Learning from Weakly Labeled Data

James Kwok

Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Hong Kong

(joint work with Yufeng Li, Ivor Tsang, Zhi-Hua Zhou)

ROKS 2013

Outline

- Introduction
- WellSVM
- Example applications
  1. semi-supervised learning
  2. multiple instance learning
  3. maximum margin clustering
- Conclusion

Introduction

Obtaining labeled data is expensive and difficult

- may involve hazardous experiments
- may involve expensive expertise (e.g., drug prediction)

Weakly labeled data: labels are incomplete / partially known

- semi-supervised learning (SSL)
- multiple instance learning (MIL)
- maximum margin clustering (MMC)

Introduction

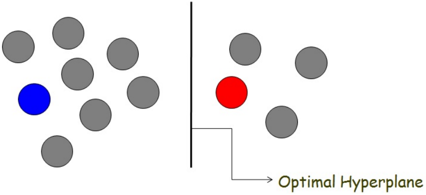
Obtaining labeled data is expensive and difficult

- may involve hazardous experiments
- may involve expensive expertise (e.g., drug prediction)

Weakly labeled data: labels are incomplete / partially known

- semi-supervised learning (SSL)
- multiple instance learning (MIL)
- maximum margin clustering (MMC)

Semi-Supervised Learning (SSL)



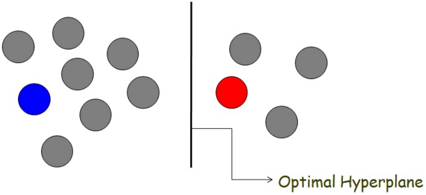
- few labeled data, lots of unlabeled data

Applications

- text categorization, medical image segmentation, word sense disambiguation, object detection

labels are partially known

Semi-Supervised Learning (SSL)



- few labeled data, lots of unlabeled data

Applications

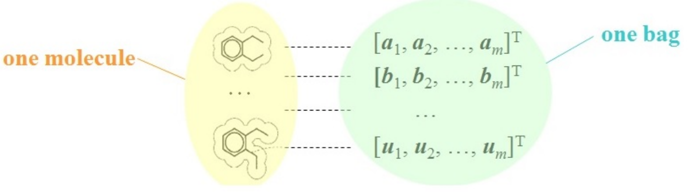
- text categorization, medical image segmentation, word sense disambiguation, object detection

labels are partially known

Multiple Instance Learning (MIL)

Example (drug activity prediction [Dietterich et al., AIJ-1997])

- given a drug molecule, predict whether it can bind to the targets (standard supervised learning?)

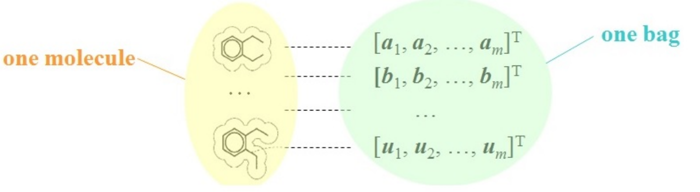


- each drug molecule can have multiple low-energy shapes or conformations
- a molecule can bind to a target if at least one of its conformations can bind
- biochemists can only tell the binding capability of a molecule, but not a particular conformation

Multiple Instance Learning (MIL)

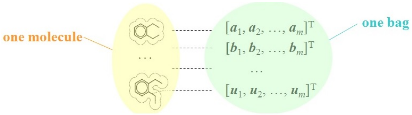
Example (drug activity prediction [Dietterich et al., AIJ-1997])

- given a drug molecule, predict whether it can bind to the targets (standard supervised learning?)

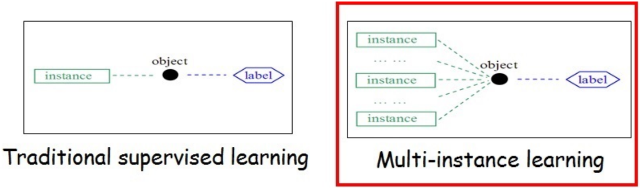


- each drug molecule can have multiple low-energy shapes or conformations
- a molecule can bind to a target if at least one of its conformations can bind
- biochemists can only tell the binding capability of a molecule, but not a particular conformation

Weak Label Information

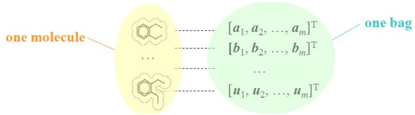


- each shape ⇒ instance; each molecule ⇒ bag
- a bag is labeled positive when it contains at least one positive instance (key instance), and is labeled negative otherwise
- only the bags (but not individual instances) have known labels

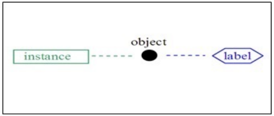


labels only implicitly known

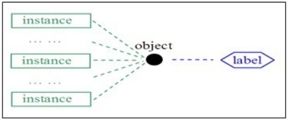
Weak Label Information



- each shape ⇒ instance; each molecule ⇒ bag
- a bag is labeled positive when it contains at least one positive instance (key instance), and is labeled negative otherwise
- only the bags (but not individual instances) have known labels



Traditional supervised learning

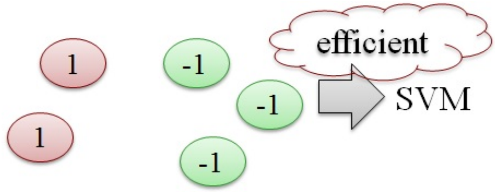


Multi-instance learning

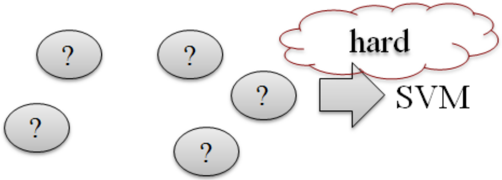
labels only implicitly known

Clustering

Supervised learning



Maximum margin clustering [Xu et al, NIPS-2005]



labels are totally unknown

Weak-Label Learning

Besides learning the parameters, needs to infer the integer-valued labels of the samples

difficult mixed-integer programming

Existing Algorithms

- global optimization
  - branch and bound [Chapelle et al., JMLR-2008], deterministic annealing [Sindhwani et al., ICML-2006]
  - not quite scalable
- semidefinite (SDP) relaxations [Xu et al., NIPS-2005]
  - convex
  - used on small data sets (thousands of examples)
- non-convex optimization
  - alternating minimization [Andrews et al., NIPS-2003], convex-concave procedure [Collobert et al., JMLR-2006]
  - often efficient, but can get stuck in local minima

Goal: A scalable yet convex optimization procedure

Existing Algorithms

- global optimization
  - branch and bound [Chapelle et al., JMLR-2008], deterministic annealing [Sindhwani et al., ICML-2006]
  - not quite scalable
- semidefinite (SDP) relaxations [Xu et al., NIPS-2005]
  - convex
  - used on small data sets (thousands of examples)
- non-convex optimization
  - alternating minimization [Andrews et al., NIPS-2003], convex-concave procedure [Collobert et al., JMLR-2006]
  - often efficient, but can get stuck in local minima

Goal: A scalable yet convex optimization procedure

Existing Algorithms

- global optimization
  - branch and bound [Chapelle et al., JMLR-2008], deterministic annealing [Sindhwani et al., ICML-2006]
  - not quite scalable
- semidefinite (SDP) relaxations [Xu et al., NIPS-2005]
  - convex
  - used on small data sets (thousands of examples)
- non-convex optimization
  - alternating minimization [Andrews et al., NIPS-2003], convex-concave procedure [Collobert et al., JMLR-2006]
  - often efficient, but can get stuck in local minima

Goal: A scalable yet convex optimization procedure

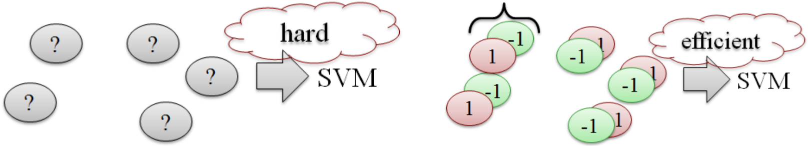
Existing Algorithms

- global optimization
  - branch and bound [Chapelle et al., JMLR-2008], deterministic annealing [Sindhwani et al., ICML-2006]
  - not quite scalable
- semidefinite (SDP) relaxations [Xu et al., NIPS-2005]
  - convex
  - used on small data sets (thousands of examples)
- non-convex optimization
  - alternating minimization [Andrews et al., NIPS-2003], convex-concave procedure [Collobert et al., JMLR-2006]
  - often efficient, but can get stuck in local minima

Goal: A scalable yet convex optimization procedure

WellSVM (WEakly LabeLed SVM)

A variant of the (convex) SVM with label generation



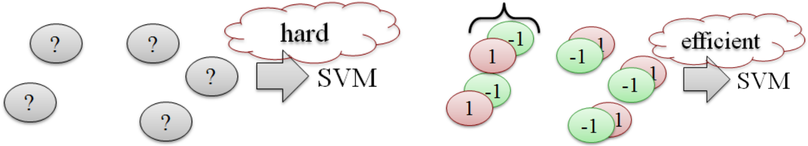
1. generate the label vectors
2. combine them via multiple kernel learning

Advantages

- a tight convex relaxation of the original mixed integer programming problem
  - at least as tight as existing convex relaxations
- can make use of state-of-the-art SVM softwares
  - scalable and efficient

WellSVM (WEakly LabeLed SVM)

A variant of the (convex) SVM with label generation



1. generate the label vectors
2. combine them via multiple kernel learning

Advantages

- a tight convex relaxation of the original mixed integer programming problem
  - at least as tight as existing convex relaxations
- can make use of state-of-the-art SVM softwares
  - scalable and efficient

Large-Margin Weak-Label Learning

- data set $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ ($\mathbf{x}_i \in \mathcal{X}$: input; $y_i \in \{\pm 1\}$: output)
- find $f : \mathcal{X} \to \{\pm 1\}$ to minimize the structural risk functional

$$\min_f \Omega(f) + C\, \ell_f(\mathcal{D})$$

  - Ω: regularizer; $\ell_f(\mathcal{D})$: empirical loss on $\mathcal{D}$
  - Ω and ℓ_f are convex

Labels $\hat{\mathbf{y}} = [\hat{y}_1, \cdots, \hat{y}_N]' \in \{\pm 1\}^N$ not available on all N examples $\Rightarrow$ need to be learned

- minimize w.r.t. f and (unknown labels in) $\hat{\mathbf{y}}$

$$\min_{\hat{\mathbf{y}} \in \mathcal{B}} \min_f \ \Omega(f) + C\, \ell_f(\{\mathbf{x}_i, \hat{y}_i\}_{i=1}^N)$$

- $\mathcal{B}$: set of candidate label assignments

Example

+ve and -ve examples are known to be approximately balanced

- $\mathcal{B} = \{\hat{\mathbf{y}} \ : \ -\beta \le \sum_{i=1}^N \hat{y}_i \le \beta\}$ for some constant β

Large-Margin Weak-Label Learning

- data set $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ ($\mathbf{x}_i \in \mathcal{X}$: input; $y_i \in \{\pm 1\}$: output)
- find $f : \mathcal{X} \rightarrow \{\pm 1\}$ to minimize the structural risk functional

$$\min_f \ \Omega(f) + C\ \ell_f(\mathcal{D})$$

  - Ω: regularizer; $\ell_f(\mathcal{D})$: empirical loss on $\mathcal{D}$
  - Ω and ℓ_f are convex

Labels $\hat{\mathbf{y}} = [\hat{y}_1, \cdots, \hat{y}_N]' \in \{\pm 1\}^N$ not available on all N examples
$\Rightarrow$ need to be learned

- minimize w.r.t. f and (unknown labels in) $\hat{\mathbf{y}}$

$$\min_{\hat{\mathbf{y}} \in \mathcal{B}} \min_f \ \ \Omega(f) + C\ \ell_f(\{\mathbf{x}_i, \hat{y}_i\}_{i=1}^N)$$

- $\mathcal{B}$: set of candidate label assignments

Example

+ve and -ve examples are known to be approximately balanced

- $\mathcal{B} = \{\hat{\mathbf{y}} \ : \ -\beta \leq \sum_{i=1}^N \hat{y}_i \leq \beta\}$ for some constant β

Large-Margin Weak-Label Learning

- data set $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ ($\mathbf{x}_i \in \mathcal{X}$: input; $y_i \in \{\pm 1\}$: output)
- find $f : \mathcal{X} \to \{\pm 1\}$ to minimize the structural risk functional

$$\min_f \Omega(f) + C\, \ell_f(\mathcal{D})$$

  - Ω: regularizer; $\ell_f(\mathcal{D})$: empirical loss on $\mathcal{D}$
  - Ω and ℓ_f are convex

Labels $\hat{\mathbf{y}} = [\hat{y}_1, \cdots, \hat{y}_N]' \in \{\pm 1\}^N$ not available on all N examples

$\Rightarrow$ need to be learned

- minimize w.r.t. f and (unknown labels in) $\hat{\mathbf{y}}$

$$\boxed{\min_{\hat{\mathbf{y}} \in \mathcal{B}} \min_f \; \Omega(f) + C\, \ell_f(\{\mathbf{x}_i, \hat{y}_i\}_{i=1}^N)}$$

- $\mathcal{B}$: set of candidate label assignments

Example

+ve and -ve examples are known to be approximately balanced

- $\mathcal{B} = \{\hat{\mathbf{y}} \; : \; -\beta \le \sum_{i=1}^N \hat{y}_i \le \beta\}$ for some constant β

Large Margin Classifiers

Primal: $\min_{\mathbf{w},\boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^N \xi_i \ : \ \hat{y}_i\mathbf{w}'\phi(\mathbf{x}_i) \geq 1-\xi_i, \ \xi_i \geq 0$

Dual: $\max_{\boldsymbol{\alpha}} \ \boldsymbol{\alpha}'\mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha} \ : \ C\mathbf{1} \geq \boldsymbol{\alpha} \geq \mathbf{0}$

- $\boldsymbol{\alpha}$: dual variable; $\mathbf{K}$: kernel matrix

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}} \ \boldsymbol{\alpha}'\mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha} \ : \ C\mathbf{1} \geq \boldsymbol{\alpha} \geq \mathbf{0}$$

More generally,

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$$

- convex set $\mathcal{A}$: e.g., $\{\boldsymbol{\alpha} \mid C\mathbf{1} \geq \boldsymbol{\alpha} \geq \mathbf{0}\}$
- $G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$: concave in $\boldsymbol{\alpha}$ for any fixed $\hat{\mathbf{y}}$
- $G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$ can be rewritten as $\bar{G}(\boldsymbol{\alpha}, \mathbf{M})$, where $\mathbf{M}$ is a psd matrix, and $\bar{G}$ is concave in $\boldsymbol{\alpha}$ and linear in $\mathbf{M}$
  - e.g., $\boldsymbol{\alpha}'\mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K}\odot\mathbf{M}_{\hat{\mathbf{y}}}\right)\boldsymbol{\alpha}$, where $\mathbf{M}_{\hat{\mathbf{y}}} = \hat{\mathbf{y}}\hat{\mathbf{y}}'$

Large Margin Classifiers

Primal: $\min_{\mathbf{w},\boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^N \xi_i \ : \ \hat{y}_i\mathbf{w}'\phi(\mathbf{x}_i) \geq 1-\xi_i, \ \xi_i \geq 0$

Dual: $\max_{\boldsymbol{\alpha}} \ \boldsymbol{\alpha}'\mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha} \ : \ C\mathbf{1} \geq \boldsymbol{\alpha} \geq \mathbf{0}$

- $\boldsymbol{\alpha}$: dual variable; $\mathbf{K}$: kernel matrix

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}} \ \boldsymbol{\alpha}'\mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha} \ : \ C\mathbf{1} \geq \boldsymbol{\alpha} \geq \mathbf{0}$$

More generally,

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ G(\boldsymbol{\alpha},\hat{\mathbf{y}})$$

- convex set $\mathcal{A}$: e.g., $\{\boldsymbol{\alpha} \mid C\mathbf{1} \geq \boldsymbol{\alpha} \geq \mathbf{0}\}$
- $G(\boldsymbol{\alpha},\hat{\mathbf{y}})$: concave in $\boldsymbol{\alpha}$ for any fixed $\hat{\mathbf{y}}$
- $G(\boldsymbol{\alpha},\hat{\mathbf{y}})$ can be rewritten as $\bar{G}(\boldsymbol{\alpha},\mathbf{M})$, where $\mathbf{M}$ is a psd matrix, and $\bar{G}$ is concave in $\boldsymbol{\alpha}$ and linear in $\mathbf{M}$
  - e.g., $\boldsymbol{\alpha}'\mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K}\odot\mathbf{M}_{\hat{\mathbf{y}}}\right)\boldsymbol{\alpha}$, where $\mathbf{M}_{\hat{\mathbf{y}}} = \hat{\mathbf{y}}\hat{\mathbf{y}}'$

Large Margin Classifiers

Primal: $\min_{\mathbf{w},\boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \ : \ \hat{y}_i\mathbf{w}'\phi(\mathbf{x}_i) \geq 1-\xi_i, \ \xi_i \geq 0$

Dual: $\max_{\boldsymbol{\alpha}} \ \boldsymbol{\alpha}'\mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha} \ : \ C\mathbf{1} \geq \boldsymbol{\alpha} \geq \mathbf{0}$

- $\boldsymbol{\alpha}$: dual variable; $\mathbf{K}$: kernel matrix

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}} \ \boldsymbol{\alpha}'\mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha} \ : \ C\mathbf{1} \geq \boldsymbol{\alpha} \geq \mathbf{0}$$

More generally,

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$$

- convex set $\mathcal{A}$: e.g., $\{\boldsymbol{\alpha} \mid C\mathbf{1} \geq \boldsymbol{\alpha} \geq \mathbf{0}\}$
- $G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$: concave in $\boldsymbol{\alpha}$ for any fixed $\hat{\mathbf{y}}$
- $G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$ can be rewritten as $\bar{G}(\boldsymbol{\alpha}, \mathbf{M})$, where $\mathbf{M}$ is a psd matrix, and $\bar{G}$ is concave in $\boldsymbol{\alpha}$ and linear in $\mathbf{M}$
  - e.g., $\boldsymbol{\alpha}'\mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K}\odot\mathbf{M}_{\hat{\mathbf{y}}}\right)\boldsymbol{\alpha}$, where $\mathbf{M}_{\hat{\mathbf{y}}} = \hat{\mathbf{y}}\hat{\mathbf{y}}'$

Relax

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$$

- interchange the order of $\max_{\boldsymbol{\alpha}\in\mathcal{A}}$ and $\min_{\hat{\mathbf{y}}\in\mathcal{B}}$

$$\text{(upper-bound)} \quad \max_{\boldsymbol{\alpha}\in\mathcal{A}} \min_{\hat{\mathbf{y}}\in\mathcal{B}} \; G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$$

$$= \max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max_{\theta} \theta \text{ s.t. } G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) \geq \theta, \; \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\}$$

- $\mu_t \geq 0$: dual variable for each constraint

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \min_{\boldsymbol{\mu}\in\mathcal{M}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

  - $\mathcal{M} = \{\boldsymbol{\mu} \mid \sum_t \mu_t = 1, \mu_t \geq 0\}$ (simplex)
- convex in $\boldsymbol{\mu}$ and concave in $\boldsymbol{\alpha}$ $\Rightarrow$ interchange order of max and min

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

Relax

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$$

- interchange the order of $\max_{\boldsymbol{\alpha}\in\mathcal{A}}$ and $\min_{\hat{\mathbf{y}}\in\mathcal{B}}$

$$\text{(upper-bound)} \quad \max_{\boldsymbol{\alpha}\in\mathcal{A}} \min_{\hat{\mathbf{y}}\in\mathcal{B}} \; G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$$

$$= \max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max_{\theta} \theta \text{ s.t. } G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) \geq \theta, \; \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\}$$

- $\mu_t \geq 0$: dual variable for each constraint

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \min_{\boldsymbol{\mu}\in\mathcal{M}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

  - $\mathcal{M} = \{\boldsymbol{\mu} \mid \sum_t \mu_t = 1, \mu_t \geq 0\}$ (simplex)

- convex in $\boldsymbol{\mu}$ and concave in $\boldsymbol{\alpha}$ $\Rightarrow$ interchange order of max and min

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

Relax

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$$

- interchange the order of $\max_{\boldsymbol{\alpha}\in\mathcal{A}}$ and $\min_{\hat{\mathbf{y}}\in\mathcal{B}}$

$$\text{(upper-bound)} \quad \max_{\boldsymbol{\alpha}\in\mathcal{A}} \min_{\hat{\mathbf{y}}\in\mathcal{B}} \; G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$$

$$= \max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max_{\theta} \theta \text{ s.t. } G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) \geq \theta, \; \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\}$$

- $\mu_t \geq 0$: dual variable for each constraint

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \min_{\boldsymbol{\mu}\in\mathcal{M}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

  - $\mathcal{M} = \{\boldsymbol{\mu} \mid \sum_t \mu_t = 1, \mu_t \geq 0\}$ (simplex)

- convex in $\boldsymbol{\mu}$ and concave in $\boldsymbol{\alpha}$ $\Rightarrow$ interchange order of max and min

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

Relax

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$$

- interchange the order of $\max_{\boldsymbol{\alpha}\in\mathcal{A}}$ and $\min_{\hat{\mathbf{y}}\in\mathcal{B}}$

$$\text{(upper-bound)} \quad \max_{\boldsymbol{\alpha}\in\mathcal{A}} \min_{\hat{\mathbf{y}}\in\mathcal{B}} \; G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$$

$$= \max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max_{\theta} \theta \text{ s.t. } G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) \geq \theta, \; \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\}$$

- $\mu_t \geq 0$: dual variable for each constraint

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \min_{\boldsymbol{\mu}\in\mathcal{M}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

  - $\mathcal{M} = \{\boldsymbol{\mu} \mid \sum_t \mu_t = 1, \mu_t \geq 0\}$ (simplex)

- convex in $\boldsymbol{\mu}$ and concave in $\boldsymbol{\alpha}$ $\Rightarrow$ interchange order of max and min

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

Relax

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$$

- interchange the order of $\max_{\boldsymbol{\alpha}\in\mathcal{A}}$ and $\min_{\hat{\mathbf{y}}\in\mathcal{B}}$

$$\text{(upper-bound)} \quad \max_{\boldsymbol{\alpha}\in\mathcal{A}} \min_{\hat{\mathbf{y}}\in\mathcal{B}} \ G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$$

$$= \quad \max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max_{\theta} \theta \text{ s.t. } G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) \geq \theta, \ \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\}$$

- $\mu_t \geq 0$: dual variable for each constraint

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \min_{\boldsymbol{\mu}\in\mathcal{M}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

  - $\mathcal{M} = \{\boldsymbol{\mu} \mid \sum_t \mu_t = 1, \mu_t \geq 0\}$ (simplex)

- convex in $\boldsymbol{\mu}$ and concave in $\boldsymbol{\alpha}$ $\Rightarrow$ interchange order of max and min

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

Tightest Convex Relaxation

Original problem

$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ G(\boldsymbol{\alpha},\hat{\mathbf{y}}) = \min_{\mathbf{M}\in\mathcal{Y}_0} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \bar{G}(\boldsymbol{\alpha},\mathbf{M})$

- $\mathcal{Y}_0 = \{\mathbf{M} \mid \mathbf{M} = \mathbf{M}_{\hat{\mathbf{y}}} \ (= \hat{\mathbf{y}}\hat{\mathbf{y}}'), \ \hat{\mathbf{y}} \in \mathcal{B}\}$

Our relaxation

$$\begin{aligned}\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha},\hat{\mathbf{y}}_t) &= \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \bar{G}(\boldsymbol{\alpha},\mathbf{M}_{\hat{\mathbf{y}}_t}) \\ &= \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \bar{G}\left(\boldsymbol{\alpha}, \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \mathbf{M}_{\hat{\mathbf{y}}_t}\right) \\ &= \min_{\mathbf{M}\in\mathcal{Y}_1} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \bar{G}(\boldsymbol{\alpha},\mathbf{M})\end{aligned}$$

- $\mathcal{Y}_1 = \{\mathbf{M} \mid \mathbf{M} = \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \mathbf{M}_{\hat{\mathbf{y}}_t}, \ \ \boldsymbol{\mu} \in \mathcal{M}\}$
  - convex hull of $\mathcal{Y}_0$ $\Rightarrow$ tightest convex relaxation
  - at least as tight as existing SDP relaxations

Tightest Convex Relaxation

Original problem

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; G(\boldsymbol{\alpha}, \hat{\mathbf{y}}) = \min_{\mathbf{M}\in\mathcal{Y}_0} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; \bar{G}(\boldsymbol{\alpha}, \mathbf{M})$$

- $\mathcal{Y}_0 = \{\mathbf{M} \mid \mathbf{M} = \mathbf{M}_{\hat{\mathbf{y}}} \; (= \hat{\mathbf{y}}\hat{\mathbf{y}}'), \; \hat{\mathbf{y}} \in \mathcal{B}\}$

Our relaxation

$$\begin{aligned}\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) &= \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \bar{G}(\boldsymbol{\alpha}, \mathbf{M}_{\hat{\mathbf{y}}_t}) \\ &= \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \bar{G}\left(\boldsymbol{\alpha}, \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \mathbf{M}_{\hat{\mathbf{y}}_t}\right) \\ &= \min_{\mathbf{M}\in\mathcal{Y}_1} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; \bar{G}(\boldsymbol{\alpha}, \mathbf{M})\end{aligned}$$

- $\mathcal{Y}_1 = \{\mathbf{M} \mid \mathbf{M} = \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \mathbf{M}_{\hat{\mathbf{y}}_t}, \;\; \boldsymbol{\mu} \in \mathcal{M}\}$
  - convex hull of $\mathcal{Y}_0 \Rightarrow$ tightest convex relaxation
  - at least as tight as existing SDP relaxations

Tightest Convex Relaxation

Original problem

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; G(\boldsymbol{\alpha}, \hat{\mathbf{y}}) = \min_{\mathbf{M}\in\mathcal{Y}_0} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; \bar{G}(\boldsymbol{\alpha}, \mathbf{M})$$

- $\mathcal{Y}_0 = \{\mathbf{M} \mid \mathbf{M} = \mathbf{M}_{\hat{\mathbf{y}}} \; (= \hat{\mathbf{y}}\hat{\mathbf{y}}'), \; \hat{\mathbf{y}} \in \mathcal{B}\}$

Our relaxation

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) = \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \bar{G}(\boldsymbol{\alpha}, \mathbf{M}_{\hat{\mathbf{y}}_t})$$

$$= \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \bar{G}\left(\boldsymbol{\alpha}, \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \mathbf{M}_{\hat{\mathbf{y}}_t}\right)$$

$$= \min_{\mathbf{M}\in\mathcal{Y}_1} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; \bar{G}(\boldsymbol{\alpha}, \mathbf{M})$$

- $\mathcal{Y}_1 = \{\mathbf{M} \mid \mathbf{M} = \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \mathbf{M}_{\hat{\mathbf{y}}_t}, \;\; \boldsymbol{\mu} \in \mathcal{M}\}$
  - convex hull of $\mathcal{Y}_0 \Rightarrow$ tightest convex relaxation
  - at least as tight as existing SDP relaxations

Tightest Convex Relaxation

Original problem

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; G(\boldsymbol{\alpha}, \hat{\mathbf{y}}) = \min_{\mathbf{M}\in\mathcal{Y}_0} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; \bar{G}(\boldsymbol{\alpha}, \mathbf{M})$$

- $\mathcal{Y}_0 = \{\mathbf{M} \mid \mathbf{M} = \mathbf{M}_{\hat{\mathbf{y}}} \; (= \hat{\mathbf{y}}\hat{\mathbf{y}}'), \; \hat{\mathbf{y}} \in \mathcal{B}\}$

Our relaxation

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) = \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \bar{G}(\boldsymbol{\alpha}, \mathbf{M}_{\hat{\mathbf{y}}_t})$$

$$= \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \bar{G}\left(\boldsymbol{\alpha}, \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \mathbf{M}_{\hat{\mathbf{y}}_t}\right)$$

$$= \min_{\mathbf{M}\in\mathcal{Y}_1} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; \bar{G}(\boldsymbol{\alpha}, \mathbf{M})$$

- $\mathcal{Y}_1 = \{\mathbf{M} \mid \mathbf{M} = \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \mathbf{M}_{\hat{\mathbf{y}}_t}, \;\; \boldsymbol{\mu} \in \mathcal{M}\}$
  - convex hull of $\mathcal{Y}_0 \Rightarrow$ tightest convex relaxation
  - at least as tight as existing SDP relaxations

Tightest Convex Relaxation

Original problem

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ G(\boldsymbol{\alpha},\hat{\mathbf{y}}) = \min_{\mathbf{M}\in\mathcal{Y}_0} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \bar{G}(\boldsymbol{\alpha},\mathbf{M})$$

- $\mathcal{Y}_0 = \{\mathbf{M} \mid \mathbf{M} = \mathbf{M}_{\hat{\mathbf{y}}} \ (= \hat{\mathbf{y}}\hat{\mathbf{y}}'), \ \hat{\mathbf{y}} \in \mathcal{B}\}$

Our relaxation

$$\begin{aligned}
\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha},\hat{\mathbf{y}}_t) &= \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \bar{G}(\boldsymbol{\alpha},\mathbf{M}_{\hat{\mathbf{y}}_t}) \\
&= \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \bar{G}\left(\boldsymbol{\alpha}, \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \mathbf{M}_{\hat{\mathbf{y}}_t}\right) \\
&= \min_{\mathbf{M}\in\mathcal{Y}_1} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \bar{G}(\boldsymbol{\alpha},\mathbf{M})
\end{aligned}$$

- $\mathcal{Y}_1 = \{\mathbf{M} \mid \mathbf{M} = \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \mathbf{M}_{\hat{\mathbf{y}}_t}, \ \ \boldsymbol{\mu} \in \mathcal{M}\}$
  - convex hull of $\mathcal{Y}_0 \Rightarrow$ tightest convex relaxation
  - at least as tight as existing SDP relaxations

Tightest Convex Relaxation

Original problem

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; G(\boldsymbol{\alpha}, \hat{\mathbf{y}}) = \min_{\mathbf{M}\in\mathcal{Y}_0} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; \bar{G}(\boldsymbol{\alpha}, \mathbf{M})$$

- $\mathcal{Y}_0 = \{\mathbf{M} \mid \mathbf{M} = \mathbf{M}_{\hat{\mathbf{y}}} \; (= \hat{\mathbf{y}}\hat{\mathbf{y}}'), \; \hat{\mathbf{y}} \in \mathcal{B}\}$

Our relaxation

$$\begin{aligned}\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) &= \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \bar{G}(\boldsymbol{\alpha}, \mathbf{M}_{\hat{\mathbf{y}}_t}) \\ &= \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \bar{G}\left(\boldsymbol{\alpha}, \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \mathbf{M}_{\hat{\mathbf{y}}_t}\right) \\ &= \min_{\mathbf{M}\in\mathcal{Y}_1} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; \bar{G}(\boldsymbol{\alpha}, \mathbf{M})\end{aligned}$$

- $\mathcal{Y}_1 = \{\mathbf{M} \mid \mathbf{M} = \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \mathbf{M}_{\hat{\mathbf{y}}_t}, \;\; \boldsymbol{\mu} \in \mathcal{M}\}$
    - convex hull of $\mathcal{Y}_0$ $\Rightarrow$ tightest convex relaxation
    - at least as tight as existing SDP relaxations

Tightest Convex Relaxation

Original problem

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; G(\boldsymbol{\alpha}, \hat{\mathbf{y}}) = \min_{\mathbf{M}\in\mathcal{Y}_0} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; \bar{G}(\boldsymbol{\alpha}, \mathbf{M})$$

- $\mathcal{Y}_0 = \{\mathbf{M} \mid \mathbf{M} = \mathbf{M}_{\hat{\mathbf{y}}} \; (= \hat{\mathbf{y}}\hat{\mathbf{y}}'), \; \hat{\mathbf{y}} \in \mathcal{B}\}$

Our relaxation

$$\begin{aligned}
\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) &= \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \bar{G}(\boldsymbol{\alpha}, \mathbf{M}_{\hat{\mathbf{y}}_t}) \\
&= \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \bar{G}\left(\boldsymbol{\alpha}, \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \mathbf{M}_{\hat{\mathbf{y}}_t}\right) \\
&= \min_{\mathbf{M}\in\mathcal{Y}_1} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; \bar{G}(\boldsymbol{\alpha}, \mathbf{M})
\end{aligned}$$

- $\mathcal{Y}_1 = \{\mathbf{M} \mid \mathbf{M} = \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \mathbf{M}_{\hat{\mathbf{y}}_t}, \;\; \boldsymbol{\mu} \in \mathcal{M}\}$
  - convex hull of $\mathcal{Y}_0 \Rightarrow$ tightest convex relaxation
  - at least as tight as existing SDP relaxations

How to Solve?

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max_{\theta} \theta \text{ s.t. } G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) \geq \theta, \ \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\}$$

- exponential number of constraints in $\mathcal{B}$
- direct optimization computationally intractable

Typically not all these constraints are active at optimality

- including only a subset of them: a very good approximation
  $\Rightarrow$ cutting plane method

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{C}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

  - $\mathcal{C}$: working set (often much smaller than $\mathcal{B}$)

How to Solve?

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max_{\theta} \theta \text{ s.t. } G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) \geq \theta, \ \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\}$$

- exponential number of constraints in $\mathcal{B}$
- direct optimization computationally intractable

Typically not all these constraints are active at optimality

- including only a subset of them: a very good approximation
  $\Rightarrow$ cutting plane method

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{C}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

  - $\mathcal{C}$: working set (often much smaller than $\mathcal{B}$)

How to Solve?

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max_{\theta} \theta \text{ s.t. } G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) \geq \theta, \ \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\}$$

- exponential number of constraints in $\mathcal{B}$
- direct optimization computationally intractable

Typically not all these constraints are active at optimality

- including only a subset of them: a very good approximation
  $\Rightarrow$ cutting plane method

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{C}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

  - $\mathcal{C}$: working set (often much smaller than $\mathcal{B}$)

How to Solve?

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max_{\theta} \theta \text{ s.t. } G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) \geq \theta, \ \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\}$$

- exponential number of constraints in $\mathcal{B}$
- direct optimization computationally intractable

Typically not all these constraints are active at optimality

- including only a subset of them: a very good approximation

$\Rightarrow$ cutting plane method

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{C}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

- $\mathcal{C}$: working set (often much smaller than $\mathcal{B}$)

How to Solve?

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max_{\theta} \theta \text{ s.t. } G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) \geq \theta, \ \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\}$$

- exponential number of constraints in $\mathcal{B}$
- direct optimization computationally intractable

Typically not all these constraints are active at optimality

- including only a subset of them: a very good approximation $\Rightarrow$ cutting plane method

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{C}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

  - $\mathcal{C}$: working set (often much smaller than $\mathcal{B}$)

Cutting Plane Algorithm by Label Generation

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{C}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

1: Initialize $\hat{\mathbf{y}}$, $\mathcal{C} = \emptyset$;
2: **repeat**
3: update $\mathcal{C} \leftarrow \{\hat{\mathbf{y}}\} \bigcup \mathcal{C}$;
4: obtain $\boldsymbol{\alpha}$ from $\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{C}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$;
5: generate a violated $\hat{\mathbf{y}}$;
6: **until** $G(\boldsymbol{\alpha}, \hat{\mathbf{y}}) > \min_{\mathbf{y}\in\mathcal{C}} G(\boldsymbol{\alpha}, \mathbf{y}) - \epsilon$ (where ϵ is a small constant) or the decrease of objective value is smaller than a threshold.

Issues

1. Given $\mathcal{C}$, how to efficiently solve the above optimization problem?
2. How to efficiently find a violated $\hat{\mathbf{y}}$ and update $\mathcal{C} \leftarrow \{\hat{\mathbf{y}}\} \bigcup \mathcal{C}$?

Cutting Plane Algorithm by Label Generation

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{C}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

1: Initialize $\hat{\mathbf{y}}$, $\mathcal{C} = \emptyset$;
2: **repeat**
3: update $\mathcal{C} \leftarrow \{\hat{\mathbf{y}}\} \bigcup \mathcal{C}$;
4: obtain $\boldsymbol{\alpha}$ from $\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{C}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$;
5: generate a violated $\hat{\mathbf{y}}$;
6: **until** $G(\boldsymbol{\alpha}, \hat{\mathbf{y}}) > \min_{\mathbf{y}\in\mathcal{C}} G(\boldsymbol{\alpha}, \mathbf{y}) - \epsilon$ (where ϵ is a small constant) or the decrease of objective value is smaller than a threshold.

Issues

1. Given $\mathcal{C}$, how to efficiently solve the above optimization problem?
2. How to efficiently find a violated $\hat{\mathbf{y}}$ and update $\mathcal{C} \leftarrow \{\hat{\mathbf{y}}\} \bigcup \mathcal{C}$?

Properties

- assume that $-G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$ is λ-strongly convex and M-Lipschitz
- $p^{(t)}$: optimal objective value at the tth iteration

$$p^{(t+1)} \leq p^{(t)} - \eta \quad (\text{where } \eta = \left(\frac{-c+\sqrt{c^2+4\epsilon}}{2}\right)^2, \ c = M\sqrt{2/\lambda})$$

The algorithm converges in no more than $\frac{p^{(1)}-p^*}{\eta}$ iterations

- magnitude of violation in the rth iteration: ϵ_r

The algorithm converges in no more than R iterations where $\sum_{r=1}^{R} \eta_r \geq p^{(1)} - p^*$, where $\eta_r = \left(\frac{-c+\sqrt{c^2+4\epsilon_r}}{2}\right)^2$

- the more effort spent on finding a violated label, the faster the convergence

Properties

- assume that $-G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$ is λ-strongly convex and M-Lipschitz
- $p^{(t)}$: optimal objective value at the tth iteration

$$p^{(t+1)} \leq p^{(t)} - \eta \quad (\text{where } \eta = \left(\frac{-c+\sqrt{c^2+4\epsilon}}{2}\right)^2, \; c = M\sqrt{2/\lambda})$$

The algorithm converges in no more than $\frac{p^{(1)} - p^*}{\eta}$ iterations

- magnitude of violation in the rth iteration: ϵ_r

The algorithm converges in no more than R iterations where $\sum_{r=1}^{R} \eta_r \geq p^{(1)} - p^*$, where $\eta_r = \left(\frac{-c+\sqrt{c^2+4\epsilon_r}}{2}\right)^2$

- the more effort spent on finding a violated label, the faster the convergence

Properties

- assume that $-G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$ is λ-strongly convex and M-Lipschitz
- $p^{(t)}$: optimal objective value at the tth iteration

$$p^{(t+1)} \leq p^{(t)} - \eta \quad (\text{where } \eta = \left(\frac{-c+\sqrt{c^2+4\epsilon}}{2}\right)^2, \; c = M\sqrt{2/\lambda})$$

The algorithm converges in no more than $\frac{p^{(1)}-p^*}{\eta}$ iterations

- magnitude of violation in the rth iteration: ϵ_r

The algorithm converges in no more than R iterations where $\sum_{r=1}^{R} \eta_r \geq p^{(1)} - p^*$, where $\eta_r = \left(\frac{-c+\sqrt{c^2+4\epsilon_r}}{2}\right)^2$

- the more effort spent on finding a violated label, the faster the convergence

Semi-Supervised Learning

- not all the training labels are known
  - $\mathcal{D}_{\mathcal{L}} = \{\mathbf{x}_i, y_i\}_{i=1}^{l}$: labeled data; $\mathcal{D}_{\mathcal{U}} = \{\mathbf{x}_j\}_{j=l+1}^{N}$: unlabeled data
  - index sets: $\mathcal{L} = \{1, \ldots, l\}$; $\mathcal{U} = \{l+1, \ldots, N\}$
- hinge loss + ℓ_2-regularizer on $\mathbf{w}$

$$\min_{\hat{\mathbf{y}} \in \mathcal{B}} \min_{\mathbf{w}, \boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{l} \xi_i + C_2 \sum_{j=l+1}^{N} \xi_i$$

$$\text{s.t.} \quad \hat{y}_i \mathbf{w}' \phi(\mathbf{x}_i) \geq 1 - \xi_i$$

Example

- $\mathcal{B} = \{\hat{\mathbf{y}} \mid \hat{\mathbf{y}}_{\mathcal{L}} = \mathbf{y}_{\mathcal{L}}, \hat{\mathbf{y}}_{\mathcal{U}} \in \{\pm 1\}^{N-l}; \frac{\mathbf{1}'\hat{\mathbf{y}}_{\mathcal{U}}}{N-l} = \frac{\mathbf{1}'\mathbf{y}_{\mathcal{L}}}{l}\}$
- $\mathbf{y}_{\mathcal{L}} = [y_1, \ldots, y_l]', \hat{\mathbf{y}}_{\mathcal{U}} = [\hat{y}_{l+1}, \ldots, \hat{y}_N]'$

Semi-Supervised Learning

- not all the training labels are known
  - $\mathcal{D}_{\mathcal{L}} = \{\mathbf{x}_i, y_i\}_{i=1}^{l}$: labeled data; $\mathcal{D}_{\mathcal{U}} = \{\mathbf{x}_j\}_{j=l+1}^{N}$: unlabeled data
  - index sets: $\mathcal{L} = \{1, \ldots, l\}$; $\mathcal{U} = \{l+1, \ldots, N\}$
- hinge loss + ℓ_2-regularizer on $\mathbf{w}$

$$\begin{aligned} \min_{\hat{\mathbf{y}} \in \mathcal{B}} \min_{\mathbf{w}, \boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{l} \xi_i + C_2 \sum_{j=l+1}^{N} \xi_i \\ \text{s.t.} \quad & \hat{y}_i \mathbf{w}' \phi(\mathbf{x}_i) \geq 1 - \xi_i \end{aligned}$$

Example

- $\mathcal{B} = \{\hat{\mathbf{y}} \mid \hat{\mathbf{y}}_{\mathcal{L}} = \mathbf{y}_{\mathcal{L}}, \hat{\mathbf{y}}_{\mathcal{U}} \in \{\pm 1\}^{N-l}; \frac{\mathbf{1}'\hat{\mathbf{y}}_{\mathcal{U}}}{N-l} = \frac{\mathbf{1}'\mathbf{y}_{\mathcal{L}}}{l}\}$
- $\mathbf{y}_{\mathcal{L}} = [y_1, \ldots, y_l]', \hat{\mathbf{y}}_{\mathcal{U}} = [\hat{y}_{l+1}, \ldots, \hat{y}_N]'$

Semi-Supervised Learning

- not all the training labels are known
  - $\mathcal{D}_\mathcal{L} = \{\mathbf{x}_i, y_i\}_{i=1}^{l}$: labeled data; $\mathcal{D}_\mathcal{U} = \{\mathbf{x}_j\}_{j=l+1}^{N}$: unlabeled data
  - index sets: $\mathcal{L} = \{1, \dots, l\}$; $\mathcal{U} = \{l+1, \dots, N\}$
- hinge loss + ℓ_2-regularizer on $\mathbf{w}$

$$\begin{array}{ll} \min_{\hat{\mathbf{y}}\in\mathcal{B}} \min_{\mathbf{w},\boldsymbol{\xi}} & \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{l} \xi_i + C_2 \sum_{j=l+1}^{N} \xi_i \\ \text{s.t.} & \hat{y}_i \mathbf{w}'\phi(\mathbf{x}_i) \geq 1 - \xi_i \end{array}$$

Example

- $\mathcal{B} = \{\hat{\mathbf{y}} \mid \hat{\mathbf{y}}_\mathcal{L} = \mathbf{y}_\mathcal{L}, \hat{\mathbf{y}}_\mathcal{U} \in \{\pm 1\}^{N-l}; \frac{\mathbf{1}'\hat{\mathbf{y}}_\mathcal{U}}{N-l} = \frac{\mathbf{1}'\mathbf{y}_\mathcal{L}}{l}\}$
- $\mathbf{y}_\mathcal{L} = [y_1, \dots, y_l]', \hat{\mathbf{y}}_\mathcal{U} = [\hat{y}_{l+1}, \dots, \hat{y}_N]'$

Semi-Supervised Learning...

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \min_{\mathbf{w},\boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{l} \xi_i + C_2 \sum_{j=l+1}^{N} \xi_i \ : \ \hat{y}_i \mathbf{w}'\phi(\mathbf{x}_i) \geq 1 - \xi_i$$

inner minimization $\Rightarrow$ dual

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\alpha\in\mathcal{A}} \ G(\alpha, \hat{\mathbf{y}}) \equiv \mathbf{1}'\alpha - \tfrac{1}{2}\alpha'\left(\mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\alpha$$

- $\mathcal{A} = \{\alpha \mid C_1 \geq \alpha_i \geq 0, C_2 \geq \alpha_j \geq 0, i \in \mathcal{L}, j \in \mathcal{U}\}$

$\Rightarrow$

$$\min_{\mu\in\mathcal{M}} \max_{\alpha\in\mathcal{A}} \ \mathbf{1}'\alpha - \tfrac{1}{2}\alpha'\left(\sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \mathbf{K}\odot\hat{\mathbf{y}}_t\hat{\mathbf{y}}_t'\right)\alpha$$

$\Rightarrow$

$$\min_{\mu\in\mathcal{M}} \max_{\alpha\in\mathcal{A}} \ \mathbf{1}'\alpha - \tfrac{1}{2}\alpha'\left(\sum_{t:\hat{\mathbf{y}}_t\in\mathcal{C}} \mu_t \mathbf{K}\odot\hat{\mathbf{y}}_t\hat{\mathbf{y}}_t'\right)\alpha$$

Semi-Supervised Learning...

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \min_{\mathbf{w},\boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C_1\sum_{i=1}^{l}\xi_i + C_2\sum_{j=l+1}^{N}\xi_i \;:\; \hat{y}_i\mathbf{w}'\phi(\mathbf{x}_i) \geq 1-\xi_i$$

inner minimization $\Rightarrow$ dual

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; G(\boldsymbol{\alpha},\hat{\mathbf{y}}) \equiv \mathbf{1}'\boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha}$$

- $\mathcal{A} = \{\boldsymbol{\alpha} \mid C_1 \geq \alpha_i \geq 0, C_2 \geq \alpha_j \geq 0, i\in\mathcal{L}, j\in\mathcal{U}\}$

$\Rightarrow$

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; \mathbf{1}'\boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}'\left(\sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}}\mu_t\mathbf{K}\odot\hat{\mathbf{y}}_t\hat{\mathbf{y}}_t'\right)\boldsymbol{\alpha}$$

$\Rightarrow$

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; \mathbf{1}'\boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}'\left(\sum_{t:\hat{\mathbf{y}}_t\in\mathcal{C}}\mu_t\mathbf{K}\odot\hat{\mathbf{y}}_t\hat{\mathbf{y}}_t'\right)\boldsymbol{\alpha}$$

Semi-Supervised Learning...

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \min_{\mathbf{w},\boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C_1\sum_{i=1}^{l}\xi_i + C_2\sum_{j=l+1}^{N}\xi_i \;:\; \hat{y}_i\mathbf{w}'\phi(\mathbf{x}_i) \geq 1-\xi_i$$

inner minimization $\Rightarrow$ dual

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; G(\boldsymbol{\alpha},\hat{\mathbf{y}}) \equiv \mathbf{1}'\boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha}$$

- $\mathcal{A} = \{\boldsymbol{\alpha} \mid C_1 \geq \alpha_i \geq 0, C_2 \geq \alpha_j \geq 0, i\in\mathcal{L}, j\in\mathcal{U}\}$

$\Rightarrow$

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; \mathbf{1}'\boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}'\left(\sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}}\mu_t\mathbf{K}\odot\hat{\mathbf{y}}_t\hat{\mathbf{y}}_t'\right)\boldsymbol{\alpha}$$

$\Rightarrow$

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; \mathbf{1}'\boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}'\left(\sum_{t:\hat{\mathbf{y}}_t\in\mathcal{C}}\mu_t\mathbf{K}\odot\hat{\mathbf{y}}_t\hat{\mathbf{y}}_t'\right)\boldsymbol{\alpha}$$

Semi-Supervised Learning...

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \min_{\mathbf{w},\boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C_1\sum_{i=1}^{l}\xi_i + C_2\sum_{j=l+1}^{N}\xi_i \ : \ \hat{y}_i\mathbf{w}'\phi(\mathbf{x}_i) \geq 1-\xi_i$$

inner minimization $\Rightarrow$ dual

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ G(\boldsymbol{\alpha},\hat{\mathbf{y}}) \equiv \mathbf{1}'\boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha}$$

- $\mathcal{A} = \{\boldsymbol{\alpha} \mid C_1 \geq \alpha_i \geq 0, C_2 \geq \alpha_j \geq 0, i\in\mathcal{L}, j\in\mathcal{U}\}$

$\Rightarrow$

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}'\left(\textstyle\sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}}\mu_t\mathbf{K}\odot\hat{\mathbf{y}}_t\hat{\mathbf{y}}_t'\right)\boldsymbol{\alpha}$$

$\Rightarrow$

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}'\left(\textstyle\sum_{t:\hat{\mathbf{y}}_t\in\mathcal{C}}\mu_t\mathbf{K}\odot\hat{\mathbf{y}}_t\hat{\mathbf{y}}_t'\right)\boldsymbol{\alpha}$$

Cutting Plane Algorithm

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\left(\sum_{t:\hat{\mathbf{y}}_t\in\mathcal{C}} \mu_t \mathbf{K} \odot \hat{\mathbf{y}}_t\hat{\mathbf{y}}_t'\right)\boldsymbol{\alpha}$$

Two important issues

Issue 1

Given $\mathcal{C}$, how to efficiently solve the above optimization problem?

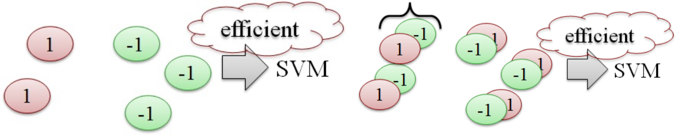
Issue 2

How to efficiently find a violated $\hat{\mathbf{y}}$?

Issue 1: How to obtain α?

$$\min_{\mu \in \mathcal{M}} \underbrace{\max_{\alpha \in \mathcal{A}} \mathbf{1}'\alpha - \frac{1}{2}\alpha' \Big(\sum_{t:\hat{\mathbf{y}}_t \in \mathcal{C}} \mu_t \mathbf{K} \odot \hat{\mathbf{y}}_t \hat{\mathbf{y}}_t' \Big) \alpha}_{\text{cf. standard SVM (with kernel matrix } \mathbf{K} \odot \hat{\mathbf{y}}\hat{\mathbf{y}}')}$$

- target kernel matrix is a convex combination of the base kernel matrices $\{\mathbf{K} \odot \hat{\mathbf{y}}_t \hat{\mathbf{y}}_t'\}$

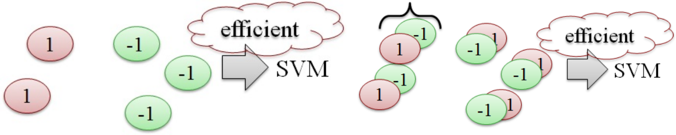


- multiple kernel learning (MKL)
  - given labels $\mathbf{y}$, find the optimal kernel $\sum_t \mu_t \mathbf{K}_t \odot \mathbf{y}\mathbf{y}'$
- multiple label-kernel learning
  - only one kernel $\mathbf{K}$, a lot of $\hat{\mathbf{y}}$'s ($\sum_t \mu_t \mathbf{K} \odot \hat{\mathbf{y}}_t \hat{\mathbf{y}}_t'$)

Issue 1: How to obtain α?

$$\min_{\mu\in\mathcal{M}} \underbrace{\max_{\boldsymbol{\alpha}\in\mathcal{A}} \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\Big(\sum_{t:\hat{\mathbf{y}}_t\in\mathcal{C}} \mu_t \mathbf{K}\odot\hat{\mathbf{y}}_t\hat{\mathbf{y}}_t'\Big)\boldsymbol{\alpha}}_{\text{cf. standard SVM (with kernel matrix } \mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\text{)}}$$

- target kernel matrix is a convex combination of the base kernel matrices $\{\mathbf{K}\odot\hat{\mathbf{y}}_t\hat{\mathbf{y}}_t'\}$



- multiple kernel learning (MKL)
  - given labels $\mathbf{y}$, find the optimal kernel $\sum_t \mu_t \mathbf{K}_t \odot \mathbf{y}\mathbf{y}'$
- multiple label-kernel learning
  - only one kernel $\mathbf{K}$, a lot of $\hat{\mathbf{y}}$'s ($\sum_t \mu_t \mathbf{K}\odot\hat{\mathbf{y}}_t\hat{\mathbf{y}}_t'$)

Multiple Label-Kernel Learning

MKL

- use the MKL-group-lasso (MKLGL) algorithm in [Xu et al., ICML-2010]
  - formulate as minimization problem $\Rightarrow$ alternating minimization

(current working set: $\mathcal{C} = \{\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_T\}$)

$$\min_{\mu \in \mathcal{M}} \min_{\mathbf{W}=[\mathbf{w}_1,\dots,\mathbf{w}_T],\xi} \frac{1}{2}\sum_{t=1}^{T} \frac{1}{\mu_t}||\mathbf{w}_t||^2 + C_1 \sum_{i=1}^{l} \xi_i + C_2 \sum_{j=l+1}^{N} \xi_j$$
$$\text{s.t. } \sum_{t=1}^{T} \hat{y}_{ti} \mathbf{w}_t' \phi(\mathbf{x}_i) \geq 1 - \xi_i$$

iterate until convergence

1. fix μ, solve for $\mathbf{w}_t$'s and ξ

   $\min \frac{1}{2}||\tilde{\mathbf{w}}||^2 + C_1 \sum_{i=1}^{l} \xi_i + C_2 \sum_{j=l+1}^{N} \xi_j \; : \; \tilde{y}_i \tilde{\mathbf{w}}' \tilde{\mathbf{x}}_i \geq 1 - \xi_i$
   - efficiently handled by standard SVM solvers
2. fix $\mathbf{w}_t$'s and ξ, update μ as $\mu_t = \frac{||\mathbf{w}_t||}{\sum_{t'=1}^{T} ||\mathbf{w}_{t'}||}$

Multiple Label-Kernel Learning

MKL

- use the MKL-group-lasso (MKLGL) algorithm in [Xu et al., ICML-2010]
  - formulate as minimization problem $\Rightarrow$ alternating minimization

(current working set: $\mathcal{C} = \{\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_T\}$)

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \min_{\mathbf{W}=[\mathbf{w}_1,\ldots,\mathbf{w}_T],\boldsymbol{\xi}} \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\mu_t}||\mathbf{w}_t||^2 + C_1\sum_{i=1}^{l}\xi_i + C_2\sum_{j=l+1}^{N}\xi_j$$
$$\text{s.t. } \sum_{t=1}^{T}\hat{y}_{ti}\mathbf{w}_t'\phi(\mathbf{x}_i) \geq 1-\xi_i$$

iterate until convergence

1. fix $\boldsymbol{\mu}$, solve for $\mathbf{w}_t$'s and $\boldsymbol{\xi}$

   $$\min \frac{1}{2}||\tilde{\mathbf{w}}||^2 + C_1\sum_{i=1}^{l}\xi_i + C_2\sum_{j=l+1}^{N}\xi_j \;:\; \tilde{y}_i\tilde{\mathbf{w}}'\tilde{\mathbf{x}}_i \geq 1-\xi_i$$

   - efficiently handled by standard SVM solvers
2. fix $\mathbf{w}_t$'s and $\boldsymbol{\xi}$, update $\boldsymbol{\mu}$ as $\mu_t = \frac{||\mathbf{w}_t||}{\sum_{t'=1}^{T}||\mathbf{w}_{t'}||}$

Issue 2: Finding a Violated Label Assignment

$$\text{Recall that } \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

$$= \max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max_{\theta} \theta \text{ s.t. } G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) \geq \theta, \ \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\}$$

To find the most violated label assignment

$$\begin{aligned} \arg\min_{\hat{\mathbf{y}}\in\mathcal{B}} G(\alpha, \hat{\mathbf{y}}) &= \arg\min_{\hat{\mathbf{y}}\in\mathcal{B}} \mathbf{1}'\alpha - \frac{1}{2}\alpha'\left(\mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\alpha \\ &= \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \frac{1}{2}\alpha'\left(\mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\alpha \\ &= \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}} \quad (\mathbf{H} \equiv \mathbf{K}\odot(\alpha\alpha')) \end{aligned}$$

- difficult

Cutting plane algorithm only requires the addition of a violated constraint at each iteration

Issue 2: Finding a Violated Label Assignment

$$\text{Recall that } \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

$$= \max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max_{\theta} \theta \text{ s.t. } G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) \geq \theta, \ \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\}$$

To find the most violated label assignment

$$\begin{aligned}
\arg\min_{\hat{\mathbf{y}}\in\mathcal{B}} G(\boldsymbol{\alpha}, \hat{\mathbf{y}}) &= \arg\min_{\hat{\mathbf{y}}\in\mathcal{B}} \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K} \odot \hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha} \\
&= \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K} \odot \hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha} \\
&= \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}} \quad (\mathbf{H} \equiv \mathbf{K} \odot (\boldsymbol{\alpha}\boldsymbol{\alpha}'))
\end{aligned}$$

- difficult

Cutting plane algorithm only requires the addition of a violated constraint at each iteration

Issue 2: Finding a Violated Label Assignment

$$\text{Recall that } \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

$$= \max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max_{\theta} \theta \text{ s.t. } G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) \geq \theta, \ \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\}$$

To find the most violated label assignment

$$\begin{aligned}
\arg\min_{\hat{\mathbf{y}}\in\mathcal{B}} G(\boldsymbol{\alpha}, \hat{\mathbf{y}}) &= \arg\min_{\hat{\mathbf{y}}\in\mathcal{B}} \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K} \odot \hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha} \\
&= \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K} \odot \hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha} \\
&= \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}} \quad (\mathbf{H} \equiv \mathbf{K} \odot (\boldsymbol{\alpha}\boldsymbol{\alpha}'))
\end{aligned}$$

- difficult

Cutting plane algorithm only requires the addition of a violated constraint at each iteration

Issue 2: Finding a Violated Label Assignment

$$\text{Recall that} \quad \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t)$$

$$= \max_{\boldsymbol{\alpha}\in\mathcal{A}} \left\{ \max_{\theta} \theta \text{ s.t. } G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) \geq \theta, \ \forall \hat{\mathbf{y}}_t \in \mathcal{B} \right\}$$

To find the most violated label assignment

$$\begin{aligned} \arg\min_{\hat{\mathbf{y}}\in\mathcal{B}} G(\boldsymbol{\alpha}, \hat{\mathbf{y}}) &= \arg\min_{\hat{\mathbf{y}}\in\mathcal{B}} \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K} \odot \hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha} \\ &= \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K} \odot \hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha} \\ &= \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}} \quad (\mathbf{H} \equiv \mathbf{K} \odot (\boldsymbol{\alpha}\boldsymbol{\alpha}')) \end{aligned}$$

- difficult

Cutting plane algorithm only requires the addition of a violated constraint at each iteration

Simple Method to Find a Violated Label Assignment

find a $\mathbf{y}$ s.t. $\mathbf{y}'\mathbf{H}\mathbf{y} > \max_{\hat{\mathbf{y}}\in\mathcal{C}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}}$

1. compute $\bar{\mathbf{y}} = \arg\max_{\hat{\mathbf{y}}\in\mathcal{C}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}}$ and $\mathbf{y}^* = \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}}$
2. $\mathbf{y}^*$ is a violated label assignment if $\bar{\mathbf{y}}'\mathbf{H}\mathbf{y}^* \neq \bar{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}}$

$$\begin{aligned}\arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}} &= \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \mathbf{r}'\hat{\mathbf{y}} \quad (\text{where } \mathbf{r} = \mathbf{H}\bar{\mathbf{y}}) \\ &= \arg\max_{\hat{\mathbf{y}}} \mathbf{r}_{\mathcal{U}}'\hat{\mathbf{y}}_{\mathcal{U}} \ : \ \hat{\mathbf{y}}_{\mathcal{U}} \in \{\pm 1\}^{N-l}, \frac{\mathbf{1}'\hat{\mathbf{y}}_{\mathcal{U}}}{N-l} = \frac{\mathbf{1}'\mathbf{y}_{\mathcal{L}}}{l}\end{aligned}$$

- at optimality, $\hat{y}_i \geq \hat{y}_j$ if $r_i > r_j$, $i, j \in \mathcal{U}$ ($\hat{y}_i$'s aligned with the sorted r_i's)

1. sort r_i's ($i \in \mathcal{U}$) in ascending order
2. to satisfy the balance constraint $\frac{\mathbf{1}'\hat{\mathbf{y}}_{\mathcal{U}}}{N-l} = \frac{\mathbf{1}'\mathbf{y}_{\mathcal{L}}}{l}$: the small $\hat{y}_i$'s are assigned -1, while the large ones are assigned 1

Simple Method to Find a Violated Label Assignment

find a $\mathbf{y}$ s.t. $\mathbf{y}'\mathbf{H}\mathbf{y} > \max_{\hat{\mathbf{y}}\in\mathcal{C}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}}$

1. compute $\bar{\mathbf{y}} = \arg\max_{\hat{\mathbf{y}}\in\mathcal{C}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}}$ and $\mathbf{y}^* = \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}}$
2. $\mathbf{y}^*$ is a violated label assignment if $\bar{\mathbf{y}}'\mathbf{H}\mathbf{y}^* \neq \bar{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}}$

$$\begin{aligned}\arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}} &= \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \mathbf{r}'\hat{\mathbf{y}} \quad (\text{where } \mathbf{r} = \mathbf{H}\bar{\mathbf{y}}) \\ &= \arg\max_{\hat{\mathbf{y}}} \mathbf{r}_{\mathcal{U}}'\hat{\mathbf{y}}_{\mathcal{U}} \ : \ \hat{\mathbf{y}}_{\mathcal{U}} \in \{\pm 1\}^{N-l}, \frac{\mathbf{1}'\hat{\mathbf{y}}_{\mathcal{U}}}{N-l} = \frac{\mathbf{1}'\mathbf{y}_{\mathcal{L}}}{l}\end{aligned}$$

- at optimality, $\hat{y}_i \geq \hat{y}_j$ if $r_i > r_j$, $i, j \in \mathcal{U}$ ($\hat{y}_i$'s aligned with the sorted r_i's)

1. sort r_i's ($i \in \mathcal{U}$) in ascending order
2. to satisfy the balance constraint $\frac{\mathbf{1}'\hat{\mathbf{y}}_{\mathcal{U}}}{N-l} = \frac{\mathbf{1}'\mathbf{y}_{\mathcal{L}}}{l}$: the small $\hat{y}_i$'s are assigned -1, while the large ones are assigned 1

Simple Method to Find a Violated Label Assignment

find a $\mathbf{y}$ s.t. $\mathbf{y}'\mathbf{H}\mathbf{y} > \max_{\hat{\mathbf{y}}\in\mathcal{C}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}}$

1. compute $\bar{\mathbf{y}} = \arg\max_{\hat{\mathbf{y}}\in\mathcal{C}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}}$ and $\mathbf{y}^* = \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}}$
2. $\mathbf{y}^*$ is a violated label assignment if $\bar{\mathbf{y}}'\mathbf{H}\mathbf{y}^* \neq \bar{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}}$

$$\begin{aligned}\arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}} &= \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \mathbf{r}'\hat{\mathbf{y}} \quad (\text{where } \mathbf{r} = \mathbf{H}\bar{\mathbf{y}}) \\ &= \arg\max_{\hat{\mathbf{y}}} \mathbf{r}_{\mathcal{U}}'\hat{\mathbf{y}}_{\mathcal{U}} \ : \ \hat{\mathbf{y}}_{\mathcal{U}} \in \{\pm 1\}^{N-l}, \frac{\mathbf{1}'\hat{\mathbf{y}}_{\mathcal{U}}}{N-l} = \frac{\mathbf{1}'\mathbf{y}_{\mathcal{L}}}{l}\end{aligned}$$

- at optimality, $\hat{y}_i \geq \hat{y}_j$ if $r_i > r_j$, $i, j \in \mathcal{U}$ ($\hat{y}_i$'s aligned with the sorted r_i's)

1. sort r_i's ($i \in \mathcal{U}$) in ascending order
2. to satisfy the balance constraint $\frac{\mathbf{1}'\hat{\mathbf{y}}_{\mathcal{U}}}{N-l} = \frac{\mathbf{1}'\mathbf{y}_{\mathcal{L}}}{l}$: the small $\hat{y}_i$'s are assigned -1, while the large ones are assigned 1

Simple Method to Find a Violated Label Assignment

find a $\mathbf{y}$ s.t. $\mathbf{y}'\mathbf{H}\mathbf{y} > \max_{\hat{\mathbf{y}} \in \mathcal{C}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}}$

1. compute $\bar{\mathbf{y}} = \arg\max_{\hat{\mathbf{y}} \in \mathcal{C}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}}$ and $\mathbf{y}^* = \arg\max_{\hat{\mathbf{y}} \in \mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}}$
2. $\mathbf{y}^*$ is a violated label assignment if $\bar{\mathbf{y}}'\mathbf{H}\mathbf{y}^* \neq \bar{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}}$

$$\begin{aligned}\arg\max_{\hat{\mathbf{y}} \in \mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}} &= \arg\max_{\hat{\mathbf{y}} \in \mathcal{B}} \mathbf{r}'\hat{\mathbf{y}} \quad (\text{where } \mathbf{r} = \mathbf{H}\bar{\mathbf{y}}) \\ &= \arg\max_{\hat{\mathbf{y}}} \mathbf{r}_{\mathcal{U}}'\hat{\mathbf{y}}_{\mathcal{U}} \ : \ \hat{\mathbf{y}}_{\mathcal{U}} \in \{\pm 1\}^{N-l}, \frac{\mathbf{1}'\hat{\mathbf{y}}_{\mathcal{U}}}{N-l} = \frac{\mathbf{1}'\mathbf{y}_{\mathcal{L}}}{l}\end{aligned}$$

- at optimality, $\hat{y}_i \geq \hat{y}_j$ if $r_i > r_j$, $i, j \in \mathcal{U}$ ($\hat{y}_i$'s aligned with the sorted r_i's)

1. sort r_i's ($i \in \mathcal{U}$) in ascending order
2. to satisfy the balance constraint $\frac{\mathbf{1}'\hat{\mathbf{y}}_{\mathcal{U}}}{N-l} = \frac{\mathbf{1}'\mathbf{y}_{\mathcal{L}}}{l}$: the small $\hat{y}_i$'s are assigned -1, while the large ones are assigned 1

Simple Method to Find a Violated Label Assignment

find a $\mathbf{y}$ s.t. $\mathbf{y}'\mathbf{H}\mathbf{y} > \max_{\hat{\mathbf{y}}\in\mathcal{C}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}}$

1. compute $\bar{\mathbf{y}} = \arg\max_{\hat{\mathbf{y}}\in\mathcal{C}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}}$ and $\mathbf{y}^* = \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}}$
2. $\mathbf{y}^*$ is a violated label assignment if $\bar{\mathbf{y}}'\mathbf{H}\mathbf{y}^* \neq \bar{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}}$

$$\begin{aligned}\arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}} &= \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \mathbf{r}'\hat{\mathbf{y}} \quad (\text{where } \mathbf{r} = \mathbf{H}\bar{\mathbf{y}}) \\ &= \arg\max_{\hat{\mathbf{y}}} \; \mathbf{r}_{\mathcal{U}}'\hat{\mathbf{y}}_{\mathcal{U}} \; : \; \hat{\mathbf{y}}_{\mathcal{U}} \in \{\pm 1\}^{N-l}, \frac{\mathbf{1}'\hat{\mathbf{y}}_{\mathcal{U}}}{N-l} = \frac{\mathbf{1}'\mathbf{y}_{\mathcal{L}}}{l}\end{aligned}$$

- at optimality, $\hat{y}_i \geq \hat{y}_j$ if $r_i > r_j$, $i, j \in \mathcal{U}$ ($\hat{y}_i$'s aligned with the sorted r_i's)

1. sort r_i's ($i \in \mathcal{U}$) in ascending order
2. to satisfy the balance constraint $\frac{\mathbf{1}'\hat{\mathbf{y}}_{\mathcal{U}}}{N-l} = \frac{\mathbf{1}'\mathbf{y}_{\mathcal{L}}}{l}$: the small $\hat{y}_i$'s are assigned -1, while the large ones are assigned 1

WELLSVM for Semi-Supervised Learning

1: initialize $\hat{\mathbf{y}}$, $\mathcal{C} = \emptyset$;
2: **repeat**
3: update $\mathcal{C} \leftarrow \{\mathbf{y}^*\} \bigcup \mathcal{C}$.
4: obtain the optimal $\{\boldsymbol{\mu}, \mathbf{W}\}$ or $\boldsymbol{\alpha}$ from MKL solver;
5: obtain the optimal solution $\mathbf{y}^* \equiv \arg\max_{\hat{\mathbf{y}} \in \mathcal{B}} \hat{\mathbf{y}}' \mathbf{H} \bar{\mathbf{y}}$ by sorting;
6: **until** $G(\boldsymbol{\alpha}, \mathbf{y}^*) > \min_{\mathbf{y} \in \mathcal{C}} G(\boldsymbol{\alpha}, \mathbf{y}) - \epsilon$ or the decrease of objective value is smaller than a threshold

Experiments

	#instances	#features		#instances	#features
Echocardiogram	132	8	*Clean1*	476	166
House	232	16	*Isolet*	600	51
Heart	270	9	*Australian*	690	42
Heart-stalog	270	13	*Diabetes*	768	8
Haberman	306	14	*German*	1,000	59
LiveDiscorders	345	6	*Krvskp*	3,196	36
Spectf	349	44	*Sick*	3,772	31
Ionosphere	351	34	*House-votes*	435	16

- 75% of the data for training, the rest for testing
- WellSVM (LIBSVM for nonlinear kernels, LIBLINEAR for linear kernel) vs
  1. standard SVM (using labeled data only);
  2. transductive SVM (TSVM)
  3. Laplacian SVM (LapSVM)
  4. universum SVM (USVM)
- SDP-based S^3VMs [Xu et al., NIPS-2005; De Bie et al., SSL book-2006]: cannot converge after 3 hours on the smallest data set

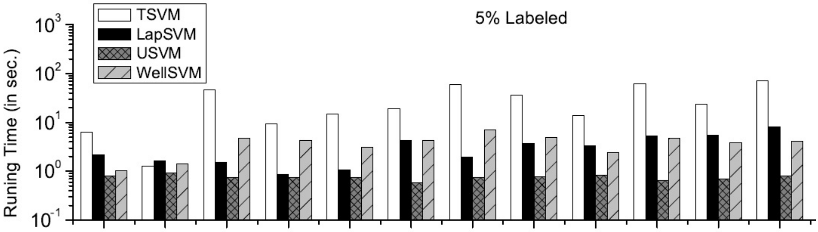
Accuracies (5% labeled examples)

	SVM	TSVM	LapSVM	USVM	WellSVM
Echocardiogram	0.80	0.74	0.64	**0.81**	0.80
House	**0.90**	**0.90**	**0.90**	**0.90**	**0.90**
Heart	0.70	0.75	0.73	0.76	**0.77**
Heart-statlog	0.73	**0.75**	0.74	**0.75**	0.73
Haberman	0.65	0.61	0.57	**0.75**	**0.75**
LiverDisorders	0.56	0.55	0.55	**0.59**	0.53
Spectf	0.73	0.68	0.61	**0.74**	0.70
Ionosphere	0.67	**0.82**	0.65	0.77	0.70
House-votes	0.88	**0.89**	0.87	0.83	**0.89**
Clean1	0.58	0.60	0.54	**0.65**	0.63
Isolet	0.97	**0.99**	0.97	0.70	0.97
Australian	0.79	**0.82**	0.78	0.80	0.81
Diabetes	0.67	0.67	0.67	**0.70**	0.69
German	**0.70**	0.69	0.62	**0.70**	**0.70**
Krvskp	0.91	**0.92**	0.80	0.91	**0.92**
Sick	**0.94**	0.89	0.90	**0.94**	**0.94**
SVM: win/tie/loss		5/7/4	8/7/1	2/9/5	**3/6/7**
avg accuracy	0.763	0.767	0.723	0.770	**0.778**

- WellSVM highly competitive

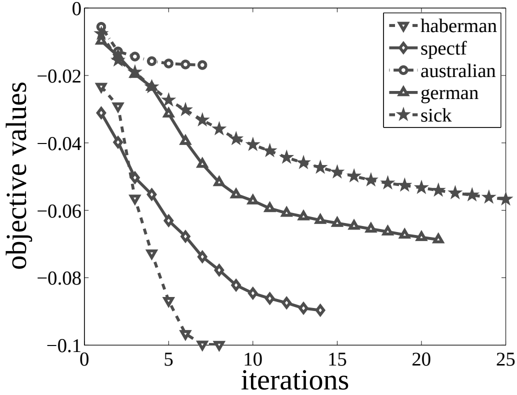
CPU Time

5% labeled samples (16 data sets)



- slowest: TSVM; fastest: USVM
- WellSVM comparable to LapSVM

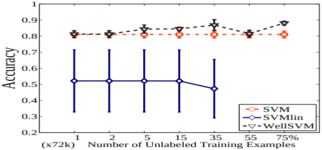
Number of WellSVM Iterations

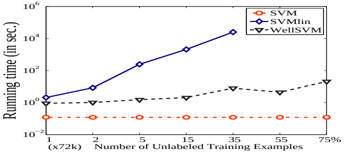


- typically, fewer than 25

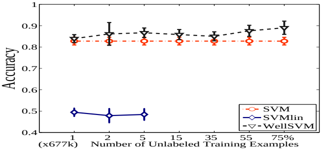
Larger Data Sets

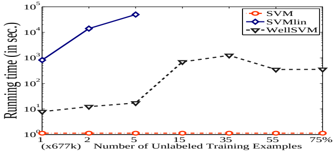
- *real-sim*: 20,958 features, 72,309 instances





- *RCV1*: 47,236 features, 677,399 instances





- linear kernel (comparison with SVMlin [Sindhwani and Keerthi, 2006])
- WellSVM is always more accurate and faster than SVMlin
- for RCV1, SVMlin cannot converge in 24 hours when $> 5\%$ examples are used for training

Comparison with Other SSL Benchmarks in the Literature

- benchmark data sets in [Chapelle, Schölkopf, Zien, SSL book-2006]
- test errors (%) (using 10 labeled examples)

	g241c	g241d	Digit1	USPS	COIL	BCI	Text
SVM	47.32	46.66	30.60	**20.03**	68.36	49.85	45.37
TSVM	**24.71**	50.08	17.77	25.20	**67.50**	49.15	40.37
WellSVM	37.37	**43.33**	**16.94**	22.74	70.73	**48.50**	**33.70**

- WellSVM is highly competitive

Comparison with SDP-based Benchmarks

- same setup in [Xu et al., 2005]
- test errors (%)

	HWD 1-7	HWD 2-3	Austr.	Flare	Vote	Diabetes
MMC	3.2	**4.7**	**32.0**	34.0	14.0	**35.6**
WELLSVM	**2.7**	5.3	40.0	**28.9**	**11.6**	41.3

- WELLSVM is again highly competitive

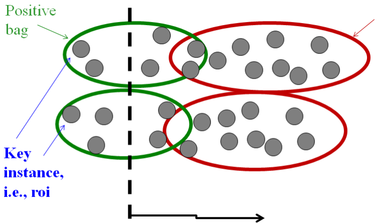
Multiple Instance Learning

- data set $\mathcal{D} = \{\mathbf{B}_i, y_i\}_{i=1}^m$
  - m: number of bags
  - bag $\mathbf{B}_i = \{\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,m_i}\}$; output $y_i \in \{\pm 1\}$
  - only bag labels available, while the instance labels are only implicitly known
- a bag is labeled positive if it contains at least one positive instance (key instance), and negative otherwise
- label of a bag is determined by its key instance, i.e., $f(\mathbf{B}_i) = \max\{f(\mathbf{x}_{i,1}), \cdots, f(\mathbf{x}_{i,m_i})\}$

Multiple Instance Learning

- data set $\mathcal{D} = \{\mathbf{B}_i, y_i\}_{i=1}^m$
  - m: number of bags
  - bag $\mathbf{B}_i = \{\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,m_i}\}$; output $y_i \in \{\pm 1\}$
  - only bag labels available, while the instance labels are only implicitly known
- a bag is labeled positive if it contains at least one positive instance (key instance), and negative otherwise
- label of a bag is determined by its key instance, i.e., $f(\mathbf{B}_i) = \max\{f(\mathbf{x}_{i,1}), \cdots, f(\mathbf{x}_{i,m_i})\}$

Multiple Instance SVM



$$\min_{\mathbf{w},\boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \textstyle\sum_{+\text{ve bag } i} \xi_i + C_2 \sum_{\text{-ve bag } i} \xi_i$$

$$\text{s.t. } \ y_i \max_{1 \le j \le m_i} \mathbf{w}'\phi(\mathbf{x}_{i,j}) \ge 1 - \xi_i$$

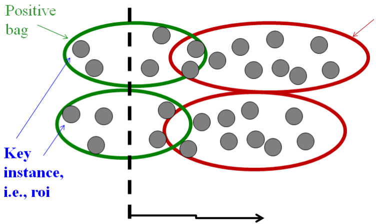
positive bag $\mathbf{B}_i$

- $\mathbf{d}_i \in \{0,1\}^{m_i}$: indicates which instance is key instance
- each +ve bag has only one key instance ($\sum_{j=1}^{m_i} d_{i,j} = 1$)

negative bag $\mathbf{B}_i$

- all its instances are negative

Multiple Instance SVM



$$\min_{\mathbf{w},\xi} \ \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum\nolimits_{+\text{ve bag } i} \xi_i + C_2 \sum\nolimits_{\text{-ve bag } i} \xi_i$$
$$\text{s.t.} \ \ y_i \max\nolimits_{1 \le j \le m_i} \mathbf{w}'\phi(\mathbf{x}_{i,j}) \ge 1 - \xi_i$$

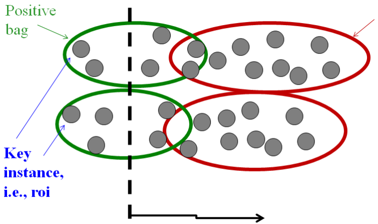
positive bag $\mathbf{B}_i$

- $\mathbf{d}_i \in \{0,1\}^{m_i}$: indicates which instance is key instance
- each +ve bag has only one key instance ($\sum_{j=1}^{m_i} d_{i,j} = 1$)

negative bag $\mathbf{B}_i$

- all its instances are negative

Multiple Instance SVM



$$\min_{\mathbf{w},\boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{\text{+ve bag } i} \xi_i + C_2 \sum_{\text{-ve bag } i} \xi_i$$
$$\text{s.t. } \ y_i \max_{1 \le j \le m_i} \mathbf{w}'\phi(\mathbf{x}_{i,j}) \ge 1 - \xi_i$$

positive bag $\mathbf{B}_i$

- $\mathbf{d}_i \in \{0,1\}^{m_i}$: indicates which instance is key instance
- each +ve bag has only one key instance ($\sum_{j=1}^{m_i} d_{i,j} = 1$)

negative bag $\mathbf{B}_i$

- all its instances are negative

Optimization Problem...

$$\min_{\mathbf{w},\boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum\nolimits_{+\text{ve bag } i} \xi_i + C_2 \sum\nolimits_{\text{-ve bag } i} \xi_i$$

$$\text{s.t.} \quad \boxed{y_i \max\nolimits_{1 \le j \le m_i} \mathbf{w}'\phi(\mathbf{x}_{i,j}) \ge 1 - \xi_i}$$

becomes

$$\min_{\mathbf{d}=[\mathbf{d}_1',\dots,\mathbf{d}_p']'} \ \min_{\mathbf{w},\boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{+\text{ve bag } i} \xi_i + C_2 \sum_{\text{-ve bag } i} \sum_{j=1}^{m_i} \xi_{i,j}$$

$$\text{s.t.} \quad \sum_{j=1}^{m_i} \mathbf{w}' d_{i,j} \phi(\mathbf{x}_{i,j}) \ge 1 - \xi_i \ \ (+\text{ve bag } i)$$

$$-\mathbf{w}'\phi(\mathbf{x}_{i,j}) \ge 1 - \xi_{i,j} \ \ (\text{each instance } j \text{ in -ve bag } i)$$

Optimization Problem...

$$\min_{\mathbf{w},\boldsymbol{\xi}} \ \tfrac{1}{2}\|\mathbf{w}\|^2 + C_1 \textstyle\sum_{\text{+ve bag } i} \xi_i + C_2 \textstyle\sum_{\text{-ve bag } i} \xi_i$$

$$\text{s.t.} \quad \boxed{y_i \max_{1\le j\le m_i} \mathbf{w}'\phi(\mathbf{x}_{i,j}) \ge 1 - \xi_i}$$

becomes

$$\min_{\mathbf{d}=[\mathbf{d}_1',\ldots,\mathbf{d}_p']'} \min_{\mathbf{w},\boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{\text{+ve bag } i} \xi_i + C_2 \sum_{\text{-ve bag } i} \sum_{j=1}^{m_i} \xi_{i,j}$$

$$\text{s.t.} \quad \sum_{j=1}^{m_i} \mathbf{w}' d_{i,j} \phi(\mathbf{x}_{i,j}) \ge 1 - \xi_i \quad (\text{+ve bag } i)$$

$$-\mathbf{w}'\phi(\mathbf{x}_{i,j}) \ge 1 - \xi_{i,j} \quad (\text{each instance } j \text{ in -ve bag } i)$$

Convex Relaxation

dual of the inner minimization problem:

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})'\mathbf{K}^{\mathbf{d}}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})$$

- $\mathcal{A} = \left\{\boldsymbol{\alpha} \;\middle|\; \begin{array}{ll} C_1 \geq \alpha_i \geq 0 & \text{for each +ve bag } i \\ C_2 \geq \alpha_j \geq 0 & \text{for each instance } j \text{ in a -ve bag} \end{array}\right\}$
- $y_i = 1$ for +ve bag; -1 for instances in -ve bags
- $\mathbf{K}^{\mathbf{d}}_{ij} = (\psi^{\mathbf{d}}_i)'(\psi^{\mathbf{d}}_j)$, where

$$\psi^{\mathbf{d}}_i = \begin{cases} \sum_{j=1}^{m_i} d_{i,j}\phi(\mathbf{x}_{i,j}) & \text{+ve bag } i \\ \phi(\mathbf{x}_{i,j}) & \text{instance } j \text{ in -ve bag } i \end{cases}$$

$$\min_{\mathbf{d}\in\Delta}\max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})'\mathbf{K}^{\mathbf{d}}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})$$

Convex relaxation

$$\min_{\boldsymbol{\mu}\in\mathcal{M}}\max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})'\left(\sum_{t:\mathbf{d}_t\in\Delta}\mu_t\mathbf{K}^{\mathbf{d}_t}\right)(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})$$

Convex Relaxation

dual of the inner minimization problem:

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})'\mathbf{K}^{\mathbf{d}}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})$$

- $\mathcal{A} = \left\{ \boldsymbol{\alpha} \mid \begin{array}{ll} C_1 \geq \alpha_i \geq 0 & \text{for each +ve bag } i \\ C_2 \geq \alpha_j \geq 0 & \text{for each instance } j \text{ in a -ve bag} \end{array} \right\}$
- $y_i = 1$ for +ve bag; -1 for instances in -ve bags
- $\mathbf{K}^{\mathbf{d}}_{ij} = (\psi^{\mathbf{d}}_i)'(\psi^{\mathbf{d}}_j)$, where

$$\psi^{\mathbf{d}}_i = \begin{cases} \sum_{j=1}^{m_i} d_{i,j}\phi(\mathbf{x}_{i,j}) & \text{+ve bag } i \\ \phi(\mathbf{x}_{i,j}) & \text{instance } j \text{ in -ve bag } i \end{cases}$$

$$\min_{\mathbf{d}\in\Delta} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})'\mathbf{K}^{\mathbf{d}}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})$$

Convex relaxation

$$\min_{\mu\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})'\left(\sum_{t:\mathbf{d}_t\in\Delta} \mu_t \mathbf{K}^{\mathbf{d}_t}\right)(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})$$

Convex Relaxation

dual of the inner minimization problem:

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})'\mathbf{K}^{\mathbf{d}}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})$$

- $\mathcal{A} = \left\{ \boldsymbol{\alpha} \mid \begin{array}{ll} C_1 \geq \alpha_i \geq 0 & \text{for each +ve bag } i \\ C_2 \geq \alpha_j \geq 0 & \text{for each instance } j \text{ in a -ve bag} \end{array} \right\}$
- $y_i = 1$ for $+$ve bag; -1 for instances in -ve bags
- $\mathbf{K}^{\mathbf{d}}_{ij} = (\psi^{\mathbf{d}}_i)'(\psi^{\mathbf{d}}_j)$, where

$$\psi^{\mathbf{d}}_i = \begin{cases} \sum_{j=1}^{m_i} d_{i,j}\phi(\mathbf{x}_{i,j}) & \text{+ve bag } i \\ \phi(\mathbf{x}_{i,j}) & \text{instance } j \text{ in -ve bag } i \end{cases}$$

$$\min_{\mathbf{d}\in\Delta} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})'\mathbf{K}^{\mathbf{d}}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})$$

Convex relaxation

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})'\left(\sum_{t:\mathbf{d}_t\in\Delta}\mu_t\mathbf{K}^{\mathbf{d}_t}\right)(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})$$

Cutting Plane: Step 1

Multiple label-kernel learning problem

$$\min_{\boldsymbol{\mu}\in\mathcal{M},\mathbf{W}=[\mathbf{w}_1;\ldots;\mathbf{w}_T],\boldsymbol{\xi}} \quad \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\mu_t}||\mathbf{w}_t||^2 + C_1\sum_{+\text{ve bag } i}\xi_i + C_2\sum_{\text{-ve bag } i}\sum_{j=1}^{m_i}\xi_{i,j}$$

$$\text{s.t.} \quad \sum_{t=1}^{T}\left(\sum_{j=1}^{m_i}\mathbf{w}_t' d_{i,j}^t \phi(\mathbf{x}_{i,j})\right) \geq 1-\xi_i \ (+\text{ve bag } i)$$

$$-\sum_{t=1}^{T}\mathbf{w}_t'\phi(\mathbf{x}_{i,j}) \geq 1-\xi_{i,j} \ (\text{instance } j \text{ in -ve bag } i)$$

- apply MKL algorithm

Cutting Plane: Step 2

find the most violated label assignment

$$\arg\min_{\mathbf{d}\in\Delta} \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})'\mathbf{K}^{\mathbf{d}}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}}) \quad = \quad \arg\max_{\mathbf{d}\in\Delta} \mathbf{d}'\mathbf{H}\mathbf{d} + \boldsymbol{\tau}'\mathbf{d}$$

- for some $\mathbf{H}$ and $\boldsymbol{\tau}$

find a violated label assignment

1. compute $\bar{\mathbf{d}} = \arg\max_{\mathbf{d}\in\mathcal{C}} \mathbf{d}'\mathbf{H}\mathbf{d} + \boldsymbol{\tau}'\mathbf{d}$ and $\mathbf{d}^* = \arg\max_{\mathbf{d}\in\Delta} \mathbf{d}'\mathbf{H}\bar{\mathbf{d}} + \frac{\boldsymbol{\tau}'\mathbf{d}}{2}$
2. $\mathbf{d}^*$ is a violated label assignment if $\mathbf{d}^{*\prime}\mathbf{H}\bar{\mathbf{d}} + \frac{\boldsymbol{\tau}'\mathbf{d}^*}{2} > \bar{\mathbf{d}}'\mathbf{H}\bar{\mathbf{d}} + \frac{\boldsymbol{\tau}'\bar{\mathbf{d}}}{2}$

find $\mathbf{d}^*$ via sorting (let $\mathbf{r} = \mathbf{H}\bar{\mathbf{d}} + \frac{\boldsymbol{\tau}}{2}$)

- $\max_{\mathbf{d}} \mathbf{r}'\mathbf{d} \; : \; \mathbf{1}'\mathbf{d}_i = 1, \mathbf{d}_i \in \{0,1\}^{m_i}$
- (recall that $\mathbf{d} = [\mathbf{d}_1', \ldots, \mathbf{d}_p']'$) solve the subproblems for each +ve bag individually
- set the the largest element in each $\mathbf{d}_i$ to 1, others to zero

Cutting Plane: Step 2

find the most violated label assignment

$$\arg\min_{\mathbf{d}\in\Delta} \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})'\mathbf{K}^{\mathbf{d}}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}}) = \arg\max_{\mathbf{d}\in\Delta} \mathbf{d}'\mathbf{H}\mathbf{d} + \boldsymbol{\tau}'\mathbf{d}$$

- for some $\mathbf{H}$ and $\boldsymbol{\tau}$

find a violated label assignment

1. compute $\bar{\mathbf{d}} = \arg\max_{\mathbf{d}\in\mathcal{C}} \mathbf{d}'\mathbf{H}\mathbf{d} + \boldsymbol{\tau}'\mathbf{d}$ and $\mathbf{d}^* = \arg\max_{\mathbf{d}\in\Delta} \mathbf{d}'\mathbf{H}\bar{\mathbf{d}} + \frac{\boldsymbol{\tau}'\mathbf{d}}{2}$
2. $\mathbf{d}^*$ is a violated label assignment if $\mathbf{d}^{*\prime}\mathbf{H}\bar{\mathbf{d}} + \frac{\boldsymbol{\tau}'\mathbf{d}^*}{2} > \bar{\mathbf{d}}'\mathbf{H}\bar{\mathbf{d}} + \frac{\boldsymbol{\tau}'\bar{\mathbf{d}}}{2}$

find $\mathbf{d}^*$ via sorting (let $\mathbf{r} = \mathbf{H}\bar{\mathbf{d}} + \frac{\boldsymbol{\tau}}{2}$)

- $\max_{\mathbf{d}} \mathbf{r}'\mathbf{d} \;:\; \mathbf{1}'\mathbf{d}_i = 1, \mathbf{d}_i \in \{0,1\}^{m_i}$
- (recall that $\mathbf{d} = [\mathbf{d}_1', \dots, \mathbf{d}_p']'$) solve the subproblems for each +ve bag individually
- set the the largest element in each $\mathbf{d}_i$ to 1, others to zero

Cutting Plane: Step 2

find the most violated label assignment

$$\arg\min_{\mathbf{d}\in\Delta} \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}})'\mathbf{K}^{\mathbf{d}}(\boldsymbol{\alpha}\odot\hat{\mathbf{y}}) = \arg\max_{\mathbf{d}\in\Delta} \mathbf{d}'\mathbf{H}\mathbf{d} + \boldsymbol{\tau}'\mathbf{d}$$

- for some $\mathbf{H}$ and $\boldsymbol{\tau}$

find a violated label assignment

1. compute $\bar{\mathbf{d}} = \arg\max_{\mathbf{d}\in\mathcal{C}} \mathbf{d}'\mathbf{H}\mathbf{d} + \boldsymbol{\tau}'\mathbf{d}$ and
   $\mathbf{d}^* = \arg\max_{\mathbf{d}\in\Delta} \mathbf{d}'\mathbf{H}\bar{\mathbf{d}} + \frac{\boldsymbol{\tau}'\mathbf{d}}{2}$
2. $\mathbf{d}^*$ is a violated label assignment if $\mathbf{d}^{*\prime}\mathbf{H}\bar{\mathbf{d}} + \frac{\boldsymbol{\tau}'\mathbf{d}^*}{2} > \bar{\mathbf{d}}'\mathbf{H}\bar{\mathbf{d}} + \frac{\boldsymbol{\tau}'\bar{\mathbf{d}}}{2}$

find $\mathbf{d}^*$ via sorting (let $\mathbf{r} = \mathbf{H}\bar{\mathbf{d}} + \frac{\boldsymbol{\tau}}{2}$)

- $\max_{\mathbf{d}} \mathbf{r}'\mathbf{d} \ : \ \mathbf{1}'\mathbf{d}_i = 1, \mathbf{d}_i \in \{0,1\}^{m_i}$
- (recall that $\mathbf{d} = [\mathbf{d}_1', \ldots, \mathbf{d}_p']'$) solve the subproblems for each +ve bag individually
- set the the largest element in each $\mathbf{d}_i$ to 1, others to zero

Experiment: CBIR

Content-based image retrieval (CBIR)

- task: classify/retrieve images based on content

- each image (bag) is composed of several segments (instances)
- an image is labeled positive when at least one of its segments is positive
- 500 COREL images from five image categories

Multiple Instance Learning for Locating ROIs

compare with

1. MI-SVM, mi-SVM [Andrews et al., NIPS-2003];
2. SVM with MI-Kernel [Gärtner et al., ICML-2002]
3. non-SVM methods: Diverse Density; EM-DD; CkNN-ROI

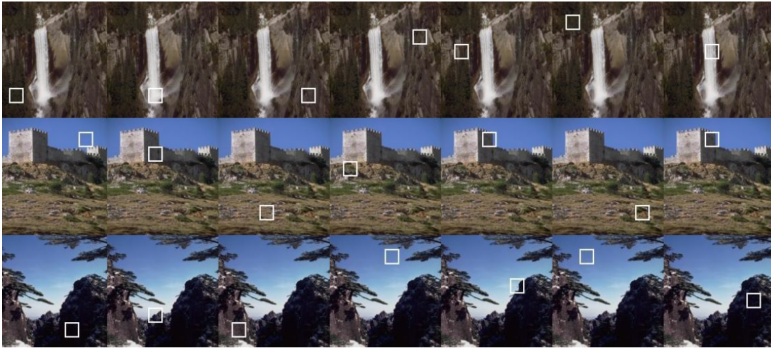
(the higher the better)

	method	*castle*	*firework*	*mountain*	*sunset*	*waterfall*
SVM methods	WELLSVM	0.57	**0.68**	**0.59**	0.32	**0.39**
	mi-SVM	0.51	0.56	0.18	0.32	0.37
	MI-SVM	0.52	0.63	0.18	0.29	0.06
	MI-Kernel	0.56	0.57	0.23	0.24	0.20
non-SVM methods	DD	0.24	0.15	0.56	0.30	0.26
	EM-DD	**0.69**	0.65	0.54	**0.36**	0.30
	C*k*NN-ROI	0.48	0.65	0.47	0.31	0.20

- WELLSVM achieves the best performance among all the SVM-based methods
- WELLSVM is still always better than DD and C*k*NN-ROI, and is highly comparable to EM-DD

Location the Region of Interest (ROI)

- usually user is only interested in some image regions (regions of interest)
- determining whether a region is a ROI $\equiv$ finding the key instance
- left to right: DD, EM-DD, CkNN-ROI, MI-SVM, mi-SVM, MI-Kernel, and WELLSVM



Maximum Margin Clustering

- all the class labels are unknown

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \min_{\mathbf{w},\boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^N \xi_i \ : \ \hat{y}_i\mathbf{w}'\phi(\mathbf{x}_i) \geq 1-\xi_i$$

  - balance constraint: $\mathcal{B} = \{\hat{\mathbf{y}} \mid \hat{y}_i \in \{+1,-1\}; -\beta \leq \mathbf{1}'\hat{\mathbf{y}} \leq \beta\}$ for some $\beta \geq 0$

Use dual in inner minimization problem

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha}$$

- $\mathcal{A} = \{\boldsymbol{\alpha} \mid C\mathbf{1} \geq \boldsymbol{\alpha} \geq \mathbf{0}\}$

Convex relaxation

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\left(\sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t\mathbf{K}\odot\hat{\mathbf{y}}_t\hat{\mathbf{y}}_t'\right)\boldsymbol{\alpha}$$

Maximum Margin Clustering

- all the class labels are unknown

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \min_{\mathbf{w},\boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^N \xi_i \ : \ \hat{y}_i\mathbf{w}'\phi(\mathbf{x}_i) \geq 1-\xi_i$$

  - balance constraint: $\mathcal{B} = \{\hat{\mathbf{y}} \mid \hat{y}_i \in \{+1,-1\}; -\beta \leq \mathbf{1}'\hat{\mathbf{y}} \leq \beta\}$ for some $\beta \geq 0$

Use dual in inner minimization problem

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K}\odot\hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\alpha}$$

- $\mathcal{A} = \{\boldsymbol{\alpha} \mid C\mathbf{1} \geq \boldsymbol{\alpha} \geq \mathbf{0}\}$

Convex relaxation

$$\min_{\mu\in\mathcal{M}} \max_{\alpha\in\mathcal{A}} \ \mathbf{1}'\alpha - \frac{1}{2}\alpha'\left(\sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t \mathbf{K}\odot\hat{\mathbf{y}}_t\hat{\mathbf{y}}_t'\right)\alpha$$

Maximum Margin Clustering

- all the class labels are unknown

$$\min_{\hat{\mathbf{y}} \in \mathcal{B}} \min_{\mathbf{w}, \boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \xi_i \ : \ \hat{y}_i \mathbf{w}' \phi(\mathbf{x}_i) \geq 1 - \xi_i$$

  - balance constraint: $\mathcal{B} = \{\hat{\mathbf{y}} \mid \hat{y}_i \in \{+1, -1\}; -\beta \leq \mathbf{1}'\hat{\mathbf{y}} \leq \beta\}$ for some $\beta \geq 0$

Use dual in inner minimization problem

$$\min_{\hat{\mathbf{y}} \in \mathcal{B}} \max_{\boldsymbol{\alpha} \in \mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}' \left(\mathbf{K} \odot \hat{\mathbf{y}}\hat{\mathbf{y}}' \right) \boldsymbol{\alpha}$$

- $\mathcal{A} = \{\boldsymbol{\alpha} \mid C\mathbf{1} \geq \boldsymbol{\alpha} \geq \mathbf{0}\}$

Convex relaxation

$$\min_{\boldsymbol{\mu} \in \mathcal{M}} \max_{\boldsymbol{\alpha} \in \mathcal{A}} \ \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}' \left(\sum_{t: \hat{\mathbf{y}}_t \in \mathcal{B}} \mu_t \mathbf{K} \odot \hat{\mathbf{y}}_t \hat{\mathbf{y}}_t' \right) \boldsymbol{\alpha}$$

Cutting Plane

Step 1 (MKL problem)

$$\min \ \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\mu_t}||\mathbf{w}_t||^2 + C\sum_{i=1}^{N}\xi_i \ : \ \sum_{t=1}^{T}\hat{y}_{ti}\mathbf{w}_t'\phi(\mathbf{x}_i) \geq 1-\xi_i$$

Step 2 (let $\mathbf{H} = \mathbf{K}\odot(\alpha\alpha')$)

1. compute $\bar{\mathbf{y}} = \arg\max_{\hat{\mathbf{y}}\in\mathcal{C}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}}$ and $\mathbf{y}^* = \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}}$
2. $\mathbf{y}^*$ is a violated label assignment if $\bar{\mathbf{y}}'\mathbf{H}\mathbf{y}^* \geq \bar{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}}$

find $\mathbf{y}^*$ via sorting (let $\mathbf{r} = \mathbf{H}\bar{\mathbf{y}}$)

- $\max_{\hat{\mathbf{y}}} \ \mathbf{r}'\hat{\mathbf{y}} \ : \ -\beta \leq \hat{\mathbf{y}}'\mathbf{1} \leq \beta, \hat{\mathbf{y}} \in \{-1,+1\}^N$
- $\hat{y}_i$'s align with the sorted values of r_i's

Cutting Plane

Step 1 (MKL problem)

$$\min \ \frac{1}{2}\sum_{t=1}^{T} \frac{1}{\mu_t}||\mathbf{w}_t||^2 + C\sum_{i=1}^{N} \xi_i \ : \ \sum_{t=1}^{T} \hat{y}_{ti}\mathbf{w}_t'\phi(\mathbf{x}_i) \geq 1 - \xi_i$$

Step 2 (let $\mathbf{H} = \mathbf{K} \odot (\boldsymbol{\alpha}\boldsymbol{\alpha}')$)

1. compute $\bar{\mathbf{y}} = \arg\max_{\hat{\mathbf{y}}\in\mathcal{C}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}}$ and $\mathbf{y}^* = \arg\max\limits_{\hat{\mathbf{y}}\in\mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}}$
2. $\mathbf{y}^*$ is a violated label assignment if $\bar{\mathbf{y}}'\mathbf{H}\mathbf{y}^* \geq \bar{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}}$

find $\mathbf{y}^*$ via sorting (let $\mathbf{r} = \mathbf{H}\bar{\mathbf{y}}$)

- $\max\limits_{\hat{\mathbf{y}}} \ \mathbf{r}'\hat{\mathbf{y}} \ : \ -\beta \leq \hat{\mathbf{y}}'\mathbf{1} \leq \beta, \hat{\mathbf{y}} \in \{-1, +1\}^N$
- $\hat{y}_i$'s align with the sorted values of r_i's

Cutting Plane

Step 1 (MKL problem)

$$\min \; \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\mu_t}||\mathbf{w}_t||^2 + C\sum_{i=1}^{N}\xi_i \;\; : \;\; \sum_{t=1}^{T}\hat{y}_{ti}\mathbf{w}_t'\phi(\mathbf{x}_i) \geq 1-\xi_i$$

Step 2 (let $\mathbf{H} = \mathbf{K}\odot(\boldsymbol{\alpha}\boldsymbol{\alpha}')$)

1. compute $\bar{\mathbf{y}} = \arg\max_{\hat{\mathbf{y}}\in\mathcal{C}} \hat{\mathbf{y}}'\mathbf{H}\hat{\mathbf{y}}$ and $\mathbf{y}^* = \arg\max_{\hat{\mathbf{y}}\in\mathcal{B}} \hat{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}}$
2. $\mathbf{y}^*$ is a violated label assignment if $\bar{\mathbf{y}}'\mathbf{H}\mathbf{y}^* \geq \bar{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}}$

find $\mathbf{y}^*$ via sorting (let $\mathbf{r} = \mathbf{H}\bar{\mathbf{y}}$)

- $\max_{\hat{\mathbf{y}}} \mathbf{r}'\hat{\mathbf{y}} \;:\; -\beta \leq \hat{\mathbf{y}}'\mathbf{1} \leq \beta, \hat{\mathbf{y}} \in \{-1,+1\}^N$
- $\hat{y}_i$'s align with the sorted values of r_i's

Experiments

1. k-means clustering (KM)
2. kernel k-means clustering (KKM)
3. normalized cut (NC)
4. GMMC [Valizadegan and Jin, NIPS-2007]
5. IterSVR [Zhang et al., ICML-2007]
6. CPMMC [Zhao et al., ICDM-2008]

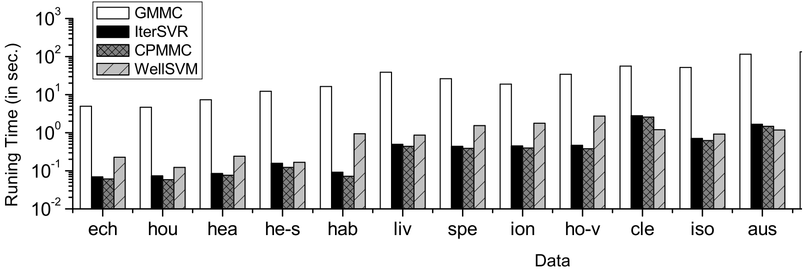
- Gaussian kernel
- initialization:
  - 20 random label assignments are generated
  - the one with the maximum kernel alignment is selected

Clustering Accuracies

	KM	KKM	NC	MMC	IterSVR	GMMC	WellSVM
Echocardiogram	0.76	0.77	0.76	0.7	0.78	0.82	**0.83**
House	0.89	0.88	0.89	0.78	0.87	0.53	**0.93**
Heart	0.59	0.59	0.57	0.7	0.59	0.56	**0.74**
Heart-statlog	0.79	0.79	0.79	0.77	0.76	0.56	**0.81**
Haberman	0.59	0.64	0.7	0.6	0.57	**0.74**	**0.74**
LiverDisorders	0.54	0.56	0.57	0.55	0.51	**0.58**	**0.58**
Spectf	0.57	**0.77**	0.63	0.64	0.53	0.73	0.73
Ionosphere	0.71	0.74	0.7	0.73	0.65	0.64	**0.77**
House-votes	0.87	0.87	0.86	0.6	0.82	0.61	**0.88**
Clean1	0.54	0.62	0.52	**0.66**	0.53	0.56	0.56
Isolet	0.96	0.95	0.98	0.56	**1.00**	0.5	**1.00**
Australian	0.55	0.57	0.56	0.6	0.51	0.56	**0.82**
Diabetes	0.67	**0.69**	0.66	**0.69**	0.66	0.65	0.68
German	0.56	0.62	0.66	0.56	0.64	**0.7**	**0.7**
Krvskp	0.51	0.55	0.56	-	0.51	0.52	**0.57**
Sick	0.63	0.77	0.84	-	0.59	**0.94**	**0.94**

- WellSVM outperforms existing clustering approaches on most data sets

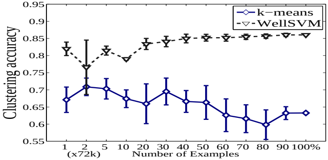
CPU Time

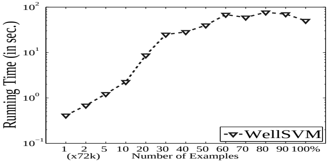


- local optimization methods (IterSVR and CPMMC): often efficient
- global optimization method: WellSVM scales much better than GMMC
  - on average, WellSVM is about 10 times faster (scales much better than GMMC)

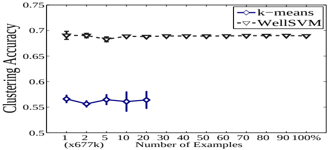
Large-Scale Experiments (Linear Kernel)

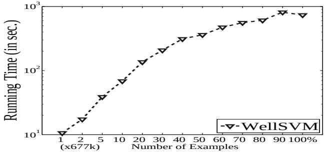
- *real-sim*: 20,958 features, 72,309 instances





- *RCV1*: 47,236 features, 677,399 instances





- WELLSVM outperforms k-means
- can be used on large data sets (takes fewer than 1,000 seconds on *RCV1*)

Conclusion

- Learning from weakly labeled data, where the training labels are incomplete
- WELLSVM : convex; based on "label generation"
  - tight relaxation
  - reduces to a sequence of standard SVM training $\Rightarrow$ much more scalable
- promising experimental results on
  1. semi-supervised learning (labels are partially known)
  2. multiple instance learning (labels are implicitly known)
  3. clustering (labels are totally unknown)

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\alpha\in\mathcal{A}} \; G(\alpha, \hat{\mathbf{y}})$$

try

$$\min_{\mu\in\mathcal{M}} \max_{\alpha\in\mathcal{A}} \; \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\alpha, \hat{\mathbf{y}}_t) + \text{cutting plane}$$

Conclusion

- Learning from weakly labeled data, where the training labels are incomplete
- WELLSVM : convex; based on "label generation"
  - tight relaxation
  - reduces to a sequence of standard SVM training $\Rightarrow$ much more scalable
- promising experimental results on
  1. semi-supervised learning (labels are partially known)
  2. multiple instance learning (labels are implicitly known)
  3. clustering (labels are totally unknown)

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\alpha\in\mathcal{A}} \; G(\alpha, \hat{\mathbf{y}})$$

try

$$\min_{\mu\in\mathcal{M}} \max_{\alpha\in\mathcal{A}} \; \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\alpha, \hat{\mathbf{y}}_t) + \text{cutting plane}$$

Conclusion

- Learning from weakly labeled data, where the training labels are incomplete
- WELLSVM : convex; based on "label generation"
  - tight relaxation
  - reduces to a sequence of standard SVM training $\Rightarrow$ much more scalable
- promising experimental results on
  1. semi-supervised learning (labels are partially known)
  2. multiple instance learning (labels are implicitly known)
  3. clustering (labels are totally unknown)

$$\min_{\hat{\mathbf{y}}\in\mathcal{B}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; G(\boldsymbol{\alpha}, \hat{\mathbf{y}})$$

try

$$\min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; \sum_{t:\hat{\mathbf{y}}_t\in\mathcal{B}} \mu_t G(\boldsymbol{\alpha}, \hat{\mathbf{y}}_t) + \text{cutting plane}$$

References

- Y.-F. Li, I.W. Tsang, J.T. Kwok, Z.-H. Zhou. Convex and Scalable Weakly Labeled SVMs. To appear in **Journal of Machine Learning Research**, 2013 (also as arXiv:1303.1271).
- Y.-F. Li, J.T. Kwok, I.W. Tsang, Z.-H. Zhou. A convex method for locating regions of interest with multi-instance learning. **ECML-2009**.
- Y.-F. Li, I.W. Tsang, J.T. Kwok, Z.-H. Zhou. Tighter and convex maximum margin clustering. **AISTATS-2009**.