Function class complexity and cluster structure with applications to transduction

Guy Lever[1]

[1]University College London
Centre for Computational Statistics and Machine Learning

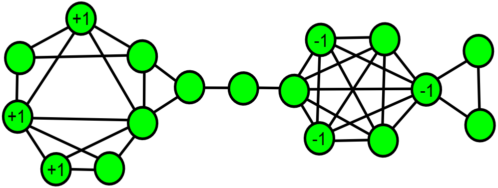
13 May, 2010

Overview

- Relate complexity to cluster structure in input space
- Cluster-structure dependent risk bounds (and algorithms)
- Investigate the complexity of learning functions defined over a graph
- Transductive and semi-supervised bounds relative to cluster structure in *resistance metric*
- Relates learning to geometry defined by data

Motivations - learning on a graph

- Predict the labelling of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$



- Understand complexity of learning over graph
- Structure poorly understood from learning theory perspective
- Existing analyses weakly dependent on graph structure
- Inspired by online bounds relative to cluster structure:

Theorem (Herbster 2008)

$M \leq \mathcal{O}\left(\mathcal{N}(\mathcal{G}, \rho, r) + \mathrm{cut}(\boldsymbol{h})\rho\right)$

- Understand the role of the structure in data generally

Preliminaries - Clustering

- $(\mathcal{X}, d)$ a metric space
- **defn.** A *clustering* of $\mathcal{S} \subset \mathcal{X}$ is any partition of $\mathcal{S}$

$$\mathcal{C} = \{\mathcal{C}_1, ...\mathcal{C}_N\}$$

- **defn.** the *center* of $\mathcal{C}_k$

$$\boldsymbol{c}_k := \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{X}} \sum_{\boldsymbol{x}' \in \mathcal{C}_k} d^2(\boldsymbol{x}', \boldsymbol{x})$$

- For each $\boldsymbol{x} \in \mathcal{S}$, $c(\boldsymbol{x}) := \boldsymbol{c}_k$ where k is such that $\boldsymbol{x} \in \mathcal{C}_k$

Preliminaries - Graph labelling

- Identify vertex $v_i \in \mathcal{V}$ with standard basis vector $\boldsymbol{e}_i$ in $\mathbb{R}^n$
- $\boldsymbol{h} \in \mathbb{R}^n$ classifies vertices $\mathcal{V} = \{v_1, ...v_n\}$ via

$$\boldsymbol{h}(v_i) := \operatorname{sgn}(\boldsymbol{h}^\top \boldsymbol{e}_i) = \operatorname{sgn}(h_i)$$

- Graph "smoothness functional" (graph cut)

$$\begin{aligned} F_{\boldsymbol{L}}(h) :&= \frac{1}{2}\boldsymbol{h}^\top \boldsymbol{L}\boldsymbol{h} \\ &= \frac{1}{2}\sum_{(i,j)\in\mathcal{E}} (h_i - h_j)^2 A_{ij} \end{aligned}$$

- $\boldsymbol{L}$ is graph *Laplacian*, $\boldsymbol{A}$ is *adjacency*
- $\mathcal{H}_\phi := \{\boldsymbol{h} \in \{-1, 1\}^n \ : \ \boldsymbol{h}^\top \boldsymbol{L}\boldsymbol{h} \le \phi\}$

Quantifying capacity

- **defn.** empirical Rademacher complexity of $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$,

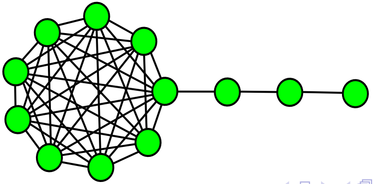
$$\widehat{\mathcal{R}}_{\mathcal{S}}(\mathcal{H}) := \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m} h(\boldsymbol{x}_i)\sigma_i\right)\right]$$

$p(\sigma_i = 1) = p(\sigma_i = -1) = \frac{1}{2}$

- **defn.** Rademacher complexity $\mathcal{R}_m(\mathcal{H}) := \mathbb{E}_{\mathcal{S}}(\widehat{\mathcal{R}}_{\mathcal{S}}(\mathcal{H}))$
- Typically sharper than VC bounds

$$\mathcal{R}_m(\mathcal{H}) \leq \mathcal{O}\left(\sqrt{\frac{\mathrm{VCdim}(\mathcal{H})}{m}}\right)$$

- Data-dependent measure of complexity...
- e.g. consider $\mathcal{R}_m(\mathcal{H}_\phi)$ vs. $\mathrm{VCdim}(\mathcal{H}_\phi)$ on $(n, \sqrt{n})$-lollipop:



Duality of complexity on $\mathcal{H}$ and distance on $\mathcal{X}$

- $\mathcal{H}$ class of linear functions on $\mathcal{X}$
- **defn.** Norm $||\cdot||$ on $\mathcal{H}$ defines *implied metric* on $\mathcal{X}$

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) := ||\boldsymbol{x}_i - \boldsymbol{x}_j||^* = \sup_{h \in \mathcal{H}} \frac{|h(\boldsymbol{x}_i) - h(\boldsymbol{x}_j)|}{||h||}.$$

- implied metric used to measure cluster structure
- e.g. RKHS $\mathcal{H} = \overline{\mathrm{span}\{K(\boldsymbol{x}, \cdot) : \boldsymbol{x} \in \mathcal{X}\}}$, $||h||_K = \sqrt{\langle h, h \rangle_K}$ has implied metric

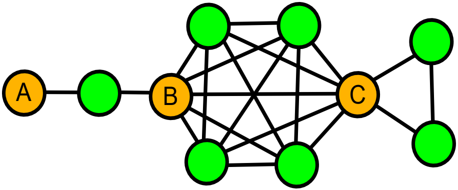
$$d_K(\boldsymbol{x}, \boldsymbol{x}') := \sqrt{K(\boldsymbol{x}, \boldsymbol{x}) + K(\boldsymbol{x}', \boldsymbol{x}') - 2K(\boldsymbol{x}, \boldsymbol{x}')}.$$

Resistance geometry on $\mathcal{G}$

- e.g. $\mathcal{H}$, functions over graph $\mathcal{G}$
- Norm $||\boldsymbol{h}||_{\boldsymbol{L}}^2 := \boldsymbol{h}^\top \boldsymbol{L}\boldsymbol{h}$ on $\mathcal{H}$
- implied metric $d_{\boldsymbol{L}} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is *resistance distance*

$$d_{\boldsymbol{L}}(v_i, v_j) := ||\boldsymbol{e}_i - \boldsymbol{e}_j||_{\boldsymbol{L}}^* = \sqrt{(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{L}^+ (\boldsymbol{e}_i - \boldsymbol{e}_j)}$$

- Edges identified as resistors



- $d_{\boldsymbol{L}}(B, C) < d_{\boldsymbol{L}}(A, B)$
- Geometry *defined by the data*
- Relate learning to *intrinsic structure* of data

Rademacher complexity and cluster structure 1

- $F : \mathcal{H} \to \mathbb{R}_{\geq 0}$ is κ-strongly convex w.r.t. $||\cdot||_F$ on $\mathcal{H}$
- $\mathcal{H}_\alpha := \{h \in \mathcal{H} \;:\; F(h) \leq \alpha\}$
- $d_F(\cdot,\cdot)$ is implied metric of $||\cdot||_F$ on $\mathcal{X}$

Theorem (refinement of Kakade et. al. 2008)

For sample $\mathcal{S} = \{\boldsymbol{x}_1, \ldots \boldsymbol{x}_m\}$, all clusterings $\mathcal{C}$ of $\mathcal{S}$, all $\alpha > 0$,

$$\widehat{\mathcal{R}}_{\mathcal{S}}(\mathcal{H}_\alpha) \leq B\sqrt{\frac{|\mathcal{C}|}{m}} + \sqrt{\frac{2\alpha\rho_{\mathcal{S}}}{m\kappa}}$$

where $\rho_{\mathcal{S}} := \frac{1}{m}\sum_{i=1}^{m} d_F^2(\boldsymbol{x}_i, c(\boldsymbol{x}_i))$ and $B := \sup_{h\in\mathcal{H}_\alpha, \boldsymbol{x}\in\mathcal{X}} |h(\boldsymbol{x})|$

- e.g. $\frac{1}{2}||\cdot||_F^2$ is 1-strongly convex w.r.t. $||\cdot||_F$
- Relates learning to cluster structure in data
- Optimized by best k-means clustering

Rademacher complexity and cluster structure 2

Theorem

For all clusterings $\mathcal{C}$ of $\mathcal{X}$ we have

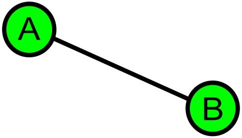
$$\mathcal{R}_m(\mathcal{H}_\alpha) \leq B\mathbb{E}_\mathcal{S}\left[\sqrt{\frac{|\mathcal{C}_\mathcal{S}|}{m}}\right] + \sqrt{\frac{2\alpha}{m\kappa}}\mathbb{E}_\mathcal{S}[\sqrt{\rho_\mathcal{S}}]$$

where $\mathcal{C}_\mathcal{S} := \{\mathcal{C}_k \in \mathcal{C} \ : \ \mathcal{S} \cap \mathcal{C}_k \neq \Phi\}$ is the clustering restricted to the sample $\mathcal{S}$.

- Relates learning to cluster structure in data-generating distribution

Is clustering an improvement?

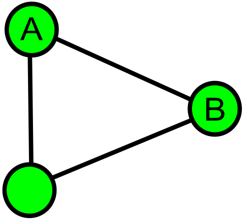
- Typical supervised setting data radius is small?
- Resistance geometry: resistance very sensitive to clustering
- $d_L^2(A, B) = 1$
- $d_L^2(A, B) = \frac{2}{3}$
- $d_L^2(A, B) = \frac{2}{4}$
- $d_L^2(A, B) = \frac{2}{5}$
- $d_L^2(A, B) = \frac{2}{6} = \mathcal{O}(\frac{1}{n})$



- non-empirical metrics not as sensitive to clustering: not *distribution dependent*

Is clustering an improvement?

- Typical supervised setting data radius is small?
- Resistance geometry: resistance very sensitive to clustering
- $d_L^2(A, B) = 1$
- $d_L^2(A, B) = \frac{2}{3}$
- $d_L^2(A, B) = \frac{2}{4}$
- $d_L^2(A, B) = \frac{2}{5}$
- $d_L^2(A, B) = \frac{2}{6} = \mathcal{O}(\frac{1}{n})$

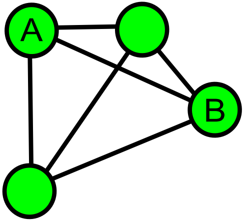


- non-empirical metrics not as sensitive to clustering: not *distribution dependent*

Is clustering an improvement?

- Typical supervised setting data radius is small?
- Resistance geometry: resistance very sensitive to clustering

- $d_L^2(A,B) = 1$
- $d_L^2(A,B) = \frac{2}{3}$
- $d_L^2(A,B) = \frac{2}{4}$
- $d_L^2(A,B) = \frac{2}{5}$
- $d_L^2(A,B) = \frac{2}{6} = \mathcal{O}(\frac{1}{n})$

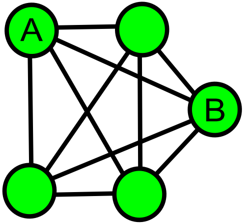


- non-empirical metrics not as sensitive to clustering: not *distribution dependent*

Is clustering an improvement?

- Typical supervised setting data radius is small?
- Resistance geometry: resistance very sensitive to clustering

- $d^2_{\mathbf{L}}(A, B) = 1$
- $d^2_{\mathbf{L}}(A, B) = \frac{2}{3}$
- $d^2_{\mathbf{L}}(A, B) = \frac{2}{4}$
- $d^2_{\mathbf{L}}(A, B) = \frac{2}{5}$
- $d^2_{\mathbf{L}}(A, B) = \frac{2}{6} = \mathcal{O}(\frac{1}{n})$

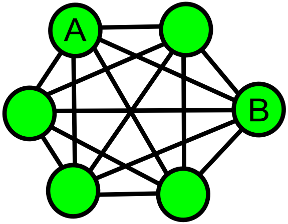


- non-empirical metrics not as sensitive to clustering: not *distribution dependent*

Is clustering an improvement?

- Typical supervised setting data radius is small?
- Resistance geometry: resistance very sensitive to clustering

- $d_L^2(A, B) = 1$
- $d_L^2(A, B) = \frac{2}{3}$
- $d_L^2(A, B) = \frac{2}{4}$
- $d_L^2(A, B) = \frac{2}{5}$
- $d_L^2(A, B) = \frac{2}{6} = \mathcal{O}(\frac{1}{n})$



- non-empirical metrics not as sensitive to clustering: not *distribution dependent*

Specialize to transduction

- Test set $\mathcal{T}$ and training set $\mathcal{S}$ presented simultaneously
- $\mathcal{S}$ drawn uniformly without replacement from $\mathcal{X} = \mathcal{S} \cup \mathcal{T}$
- $\mathcal{H}_\phi := \{\boldsymbol{h} \in \{-1, 1\}^n \ : \ \boldsymbol{h}^\top \boldsymbol{L} \boldsymbol{h} \leq \phi\}$

Corollary

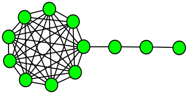
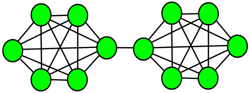
Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, for any clustering $\mathcal{C}$ of $\mathcal{V}$

$$\mathcal{R}_m^{\mathrm{trs}}(\mathcal{H}_\phi) \leq \mathbb{E}_\mathcal{S}\left[\sqrt{\frac{|\mathcal{C}_\mathcal{S}|}{m}}\right] + \sqrt{\frac{\phi\rho}{m}}$$

where $\rho := \frac{1}{n}\sum_{i=1}^n d_{\boldsymbol{L}}^2(v_i, c(v_i))$ and $\mathcal{C}_\mathcal{S} := \{\mathcal{C}_k \in \mathcal{C} \ : \ \mathcal{S} \cap \mathcal{C}_k \neq \Phi\}$ is the clustering restricted to the sample $\mathcal{S}$.

- Relates learning on graph to clustering in resistance

Comparison to VC dimension 1 - lollipops and barbells

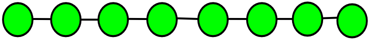
- $\mathrm{VCdim}(\mathcal{H}_\phi) \leq \mathcal{O}\left(\frac{\phi}{\phi^\star}\right)$ (Kleinberg 2004)
- $\phi^\star$ minimum # edges required to disconnect $\mathcal{G}$
- e.g. lollipop-type  : Rademacher better
- e.g. *n*-barbell graph  :

$$\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}_\phi)}{m}} \leq \mathcal{O}\left(\sqrt{\frac{\phi}{m}}\right)$$

$$\mathcal{R}^{\mathrm{trs}}_m(\mathcal{H}_\phi) \leq \sqrt{\frac{2}{m}} + \sqrt{\frac{\phi}{mn}}$$

- Advantage of clustering: resistance *between* clusters large
- Weighted graphs: even more improvement

Comparison to VC dimension 2 - paths

- e.g. path graph 

$$\sqrt{\frac{\text{VCdim}(\mathcal{H}_\phi)}{m}} \leq \mathcal{O}\left(\sqrt{\frac{\phi}{m}}\right)$$

- Rademacher bound vacuous
- Improved by passing to p resistance (Herbster and Lever 2009):
    - Family of p-norms on graph labellings $||\boldsymbol{h}||_p := \left(\sum_{(i,j)\in\mathcal{E}} |h_i - h_j|^p\right)^{\frac{1}{p}}$
    - p-resistance: $d_p(v_i, v_j) := ||\boldsymbol{e}_i - \boldsymbol{e}_j||_p^*$
- p resistance as $p \to 1$ more suitable for sparse graphs

Transductive risk analysis

- **defn.** *Transductive risk* $\mathrm{risk}_{\mathcal{T}}(\boldsymbol{h}) := \frac{1}{u}\sum_{i=1}^{u} \ell(h(x_{t_i}), y_{t_i})$
- (loss on test set $\mathcal{T} = \{(\boldsymbol{X}_{t_1}, Y_{t_1}), \ldots(\boldsymbol{X}_{t_u}, Y_{t_u})\}$)

Theorem

For any clustering $\mathcal{C}$ of $\mathcal{V}$, with probability at least $1-\delta$ over the draw of $\mathcal{S}$, simultaneously for all $\boldsymbol{h} \in \{-1, 1\}^n$

$$\mathrm{risk}_{\mathcal{T}}(\boldsymbol{h}) - \widehat{\mathrm{risk}}_{\mathcal{S}}(\boldsymbol{h}) \leq \mathcal{O}\left(\frac{n}{u}\left(\mathbb{E}_{\mathcal{S}}\left[\sqrt{\frac{|\mathcal{C}_{\mathcal{S}}|}{m}}\right] + \sqrt{\frac{F_{\boldsymbol{L}}(\boldsymbol{h})\rho}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{m}}\right)\right)$$

where $\rho = \frac{1}{n}\sum_{i=1}^{n} d_{\boldsymbol{L}}^2(v_i, c(v_i))$ *and* $\mathcal{C}_{\mathcal{S}} = \{\mathcal{C}_k \in \mathcal{C} \;:\; \mathcal{S} \cap \mathcal{C}_k \neq \Phi\}$

- Suitable for e.g. mincut , TSVM, regularization of Belkin and Niyogi, Energy minimiization of Zhu, Pelckmans and Shawe-Taylor etc.
- Suggests algorithms obtained by minimising over clusterings and classifiers (and p)

Comparison 1

Theorem (Hanneke 2006)

With probability at least $1 - \delta$ *simultaneously for all* $\boldsymbol{h} \in \{-1, 1\}^n$,

$$\mathrm{risk}_{\mathcal{T}}(\boldsymbol{h}) \leq \widehat{\mathrm{risk}}_{\mathcal{S}}(\boldsymbol{h}) + \mathcal{O}\left(\sqrt{\frac{n(u+1)}{u^2} \frac{\frac{F_L(\boldsymbol{h})}{\phi^\star} \ln n + \ln \frac{1}{\delta}}{m}}\right)$$

where $\phi^\star$ *is the minimum number of edges that must be removed to disconnect the graph*

- New bounds preffered for highly clustered graphs

Comparison 2

Theorem (Pelckmans and Shawe-Taylor 2007)

With probability at least $1 - \delta$,

$$\sup_{\boldsymbol{h} \in \mathcal{H}_\phi} |\mathrm{risk}_{\mathcal{T}}(\boldsymbol{h}) - \widehat{\mathrm{risk}}_{\mathcal{S}}(\boldsymbol{h})| \leq \sqrt{\frac{2(n-m+1)}{nm} \log \frac{|\mathcal{H}_\phi|}{\delta}}$$

with $|\mathcal{H}_\phi| \leq \left(\frac{en}{n_\phi}\right)^{n_\phi}$ *where* $n_\phi := |\{\lambda_i : \lambda_i \leq \phi\}|$.

- Relates transductive classification risk to spectrum $\{\lambda_i\}_{i=1}^n$ of graph Laplacian

Extension to semi-supervised learning

- Relate learning to cluster structure in all labelled and unlabelled data $\mathcal{I} = \{(X_1, y_1), ...(X_m, y_m), X_{m+1}, ...X_n\}$

Theorem

ℓ a K-Lipschitz loss function. For all clusterings $\mathcal{C}, \mathcal{C}'$ of $\mathcal{I}$, with prob $1-\delta$, for all $h \in \widetilde{\mathcal{H}}_\beta \subseteq \mathcal{H}_\alpha$.

$$\mathrm{risk}^\ell(h) \leq \widehat{\mathrm{risk}}^\ell_{\mathcal{S}}(h) + \mathcal{O}\left(\mathcal{R}^{\mathrm{trs}}_m(\widetilde{\mathcal{H}}_\beta) + \widehat{\mathcal{R}}^{\mathrm{ind}}_{\mathcal{I}}(\mathcal{H}_\alpha) + \sqrt{\frac{1}{m}\log\frac{1}{\delta}}\right)$$

$$\mathcal{R}^{\mathrm{trs}}_m(\widetilde{\mathcal{H}}_\beta) \leq \mathcal{O}\left(\sqrt{\frac{|\mathcal{C}|}{m}} + \sqrt{\frac{\beta}{mn}\sum_{\boldsymbol{x}\in\mathcal{I}} d^2_{\widetilde{F}}(\boldsymbol{x}, c(\boldsymbol{x}))}\right)$$

$$\widehat{\mathcal{R}}^{\mathrm{ind}}_{\mathcal{I}}(\mathcal{H}_\alpha) \leq \mathcal{O}\left(\sqrt{\frac{|\mathcal{C}'|}{n}} + \frac{1}{n}\sqrt{\alpha\sum_{\boldsymbol{x}\in\mathcal{I}} d^2_F(\boldsymbol{x}, c'(\boldsymbol{x}))}\right)$$

$d_F(\cdot,\cdot)$ and $d_{\widetilde{F}}(\cdot,\cdot)$ are metrics on $\mathcal{X}$ implied by $||\cdot||_F$ and $||\cdot||_{\widetilde{F}}$

Conclusions

- Relate complexity to cluster structure of data
- Specialized to clustering in resistance geometry
  - Convex duality analysis of learning on a graph
- Risk analysis for transduction w.r.t. resistive geometry
- Suggests algorithms related to cluster structure
- Open problems:
  - Understand how structure of graph relates to learning
  - Spectral approach, resistance clustering, combinatorial, graph theoretic...
  - Question for data structure more generally