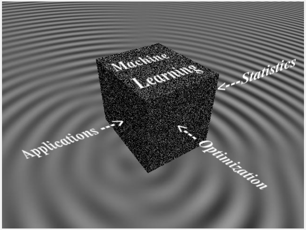
Clustering and Transductive Inference for Undirected Graphs

Kristiaan Pelckmans
Stability and Resampling Methods for Clustering

ESAT - SCD/sista
KULeuven, Leuven, Belgium

July, 2007



Overview of this talk

(0) Clustering generalities
(1) Clustering and Transductive Inference
(2) Plausible Labeling Classes
(3) Stability in Learning and Clustering
(4) Graph Algorithms, MINCUT, and Regularization

Clustering Generalities

Definition (An attempt)

For a set of n observations S and an hypothesis set $\mathcal{H}$,

$$\max_{f \in \mathcal{H}} \mathcal{J}_\gamma(f|S) = \text{Falsifiable}(f) \; + \; \gamma \, \text{Reproducible}_{S'_m \sim S}(f|S'_m)$$

- $(f|S)$ as 'f projected on the data S'
- Reproducibility $\approx$ stability, robust evidence, ...
- Falsifiability $\approx$ surprisingness, -entropy, ...
- Very general ...

- VQ and compression
- Generative model
- (Image) Segmentation
- Multiple view prediction (2 - associative clustering; many - (Krupka, 2005))
- Organization and retrieval
- ...

$\longrightarrow$ What is a good taxonomy?

- Clustering as finding Apparent Structure
- ...or 'fast learnable hypothesis class'
- Clustering for deriving plausible hypothesis space
- But interpretable: clusterwise constant (independence)
- Clustering as stage before prediction...
- ... or as developing regularization (prior)
- Different from VQ (auto-regression)
- No stochastical assumption so far...

- Clustering as finding Apparent Structure
- ...or ’fast learnable hypothesis class’
- Clustering for deriving plausible hypothesis space
- But interpretable: clusterwise constant (independence)
- Clustering as stage before prediction...
- ... or as developing regularization (prior)
- Different from VQ (auto-regression)
- No stochastical assumption so far...

- Suppose Alice and Bob have both a fair knowledge of the political structure in the world
- Alice knows a specific law in Europe/ USA/...
- $\Rightarrow$ It would be easy to communicate this piece of information

Alice  $\longrightarrow$ Bob 

- Suppose Alice and Bob have both a fair knowledge of the political structure in the world
- Alice knows a specific law in Europe/ USA/...
- $\Rightarrow$ It would be easy to communicate this piece of information

Alice  $\longrightarrow$ Bob 

So what can be learned from results in transductive inference?

- Fixed amount of $n \in \mathbb{N}_0$ nodes (objects) $V = \{v_1, \ldots, v_n\}$
- Organized in *deterministic* graph $\mathcal{G} = (V, E)$ with edges $E = \{x_{ij} \geq 0\}_{i \neq j}$ (symmetrical $x_{ij} = x_{ji}$, no loops $x_{ii} = 0$)
- Fixed label $y_i \in \{-1, 1\}$ for any node $i = 1, \ldots, n$, but only partially observed: $\mathcal{S} \subset \{1, \ldots, n\}$

$$y_{\mathcal{S}} = \{y_i \in \{-1, 1\}\}_{i \in \mathcal{S}} .$$

- Predict the remaining labels

$$y_{\neg\mathcal{S}} = \{y_i \in \{-1, 1\}\}_{i \notin \mathcal{S}} .$$

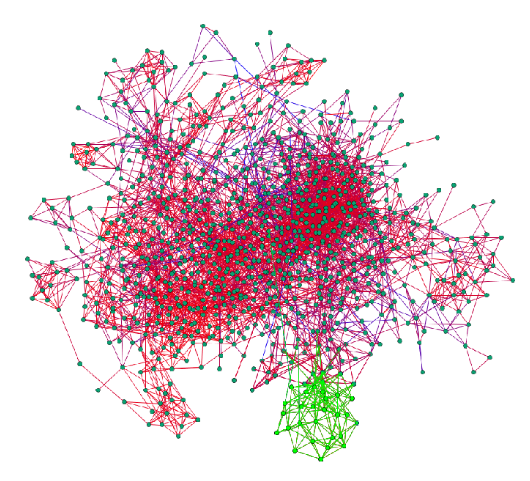
- Fixed amount of $n \in \mathbb{N}_0$ nodes (objects) $V = \{v_1, \ldots, v_n\}$
- Organized in *deterministic* graph $\mathcal{G} = (V, E)$ with edges $E = \{x_{ij} \geq 0\}_{i \neq j}$ (symmetrical $x_{ij} = x_{ji}$, no loops $x_{ii} = 0$)
- Fixed label $y_i \in \{-1, 1\}$ for any node $i = 1, \ldots, n$, but only partially observed: $\mathcal{S} \subset \{1, \ldots, n\}$

$$y_{\mathcal{S}} = \{y_i \in \{-1, 1\}\}_{i \in \mathcal{S}} .$$

- Predict the remaining labels

$$y_{\neg\mathcal{S}} = \{y_i \in \{-1, 1\}\}_{i \notin \mathcal{S}} .$$

Gene coexpression:



- Hypothesis set:

$$\mathcal{H} = \left\{ q \in \{-1, 1\}^n \right\}$$

with $|\mathcal{H}| = 2^n$

- Given a restricted hypothesis set $\mathcal{H}' \subset \mathcal{H}$ with $|\mathcal{H}'| \ll |\mathcal{H}|$, and a few observations $y_\mathcal{S}$ where $\mathcal{S}$ is uniform without replacement
- Actual risk

$$\mathcal{R}(q|\mathcal{G}) = E[I(y_J q_J < 0)] = \frac{1}{n}\sum_{i=1}^{n} I(y_i q_i < 0),$$

with E over uniform choice of j in $y_J \in \{y_i\}_i$.

- Empirical risk

$$\mathcal{R}_\mathcal{S}(q|\mathcal{G}) = \frac{1}{m}\sum_{i \in \mathcal{S}} I(y_i q_i < 0).$$

- Transductive risk

$$\mathcal{R}_{\neg\mathcal{S}}(q|\mathcal{G}) = \frac{1}{n-m}\sum_{i \notin \mathcal{S}} I(y_i q_i < 0).$$

- Hypothesis set:

$$\mathcal{H} = \{q \in \{-1, 1\}^n\}$$

  with $|\mathcal{H}| = 2^n$

- Given a restricted hypothesis set $\mathcal{H}' \subset \mathcal{H}$ with $|\mathcal{H}'| \ll |\mathcal{H}|$, and a few observations $y_\mathcal{S}$ where $\mathcal{S}$ is uniform without replacement

- Actual risk

$$\mathcal{R}(q|\mathcal{G}) = E[I(y_J q_J < 0)] = \frac{1}{n}\sum_{i=1}^{n} I(y_i q_i < 0),$$

  with E over uniform choice of j in $y_J \in \{y_i\}_i$.

- Empirical risk

$$\mathcal{R}_\mathcal{S}(q|\mathcal{G}) = \frac{1}{m}\sum_{i\in\mathcal{S}} I(y_i q_i < 0).$$

- Transductive risk

$$\mathcal{R}_{\neg\mathcal{S}}(q|\mathcal{G}) = \frac{1}{n-m}\sum_{i\notin\mathcal{S}} I(y_i q_i < 0).$$

- Hypothesis set:

$$\mathcal{H} = \{q \in \{-1, 1\}^n\}$$

with $|\mathcal{H}| = 2^n$

- Given a restricted hypothesis set $\mathcal{H}' \subset \mathcal{H}$ with $|\mathcal{H}'| \ll |\mathcal{H}|$, and a few observations $y_\mathcal{S}$ where $\mathcal{S}$ is uniform without replacement
- Actual risk

$$\mathcal{R}(q|\mathcal{G}) = E[I(y_J q_J < 0)] = \frac{1}{n}\sum_{i=1}^{n} I(y_i q_i < 0),$$

with E over uniform choice of j in $y_J \in \{y_i\}_i$.

- Empirical risk

$$\mathcal{R}_\mathcal{S}(q|\mathcal{G}) = \frac{1}{m}\sum_{i\in\mathcal{S}} I(y_i q_i < 0).$$

- Transductive risk

$$\mathcal{R}_{\neg\mathcal{S}}(q|\mathcal{G}) = \frac{1}{n-m}\sum_{i\notin\mathcal{S}} I(y_i q_i < 0).$$

Theorem (Generalization Bound)

Let $\mathcal{S} \subset \{1, \ldots, n\}$ be uniformly sampled without replacement. Consider a set of hypothetical labelings $\mathcal{H}' \subset \mathcal{H}^n$ having a cardinality of $|\mathcal{H}'| \in \mathbb{N}$. Then the following inequality holds with probability higher than $(1 - \delta) < 1$.

$$\sup_{q \in \mathcal{H}'} \mathcal{R}(q|\mathcal{G}) - \mathcal{R}_{\mathcal{S}}(q|\mathcal{G}) \leq \sqrt{\frac{2(n-m+1)}{mn} (\ln(|\mathcal{H}'|) - \ln(\delta))}. \tag{1}$$

- Transductive risk

$$\forall q \in \mathcal{H}' : \mathcal{R}_{\neg\mathcal{S}}(q|\mathcal{G}) \leq \mathcal{R}(q|\mathcal{G}) + \left(\frac{n}{n-m}\right) \sqrt{\frac{2(n-m+1)}{mn} (\ln(|\mathcal{H}'|) + \ln(1/\delta))}.$$

Theorem (Generalization Bound)

Let $\mathcal{S} \subset \{1, \ldots, n\}$ be uniformly sampled without replacement. Consider a set of hypothetical labelings $\mathcal{H}' \subset \mathcal{H}^n$ having a cardinality of $|\mathcal{H}'| \in \mathbb{N}$. Then the following inequality holds with probability higher than $(1-\delta) < 1$.

$$\sup_{q \in \mathcal{H}'} \mathcal{R}(q|\mathcal{G}) - \mathcal{R}_{\mathcal{S}}(q|\mathcal{G}) \leq \sqrt{\frac{2(n-m+1)}{mn} \left(\ln(|\mathcal{H}'|) - \ln(\delta)\right)}. \quad (1)$$

- Transductive risk

$$\forall q \in \mathcal{H}' : \; \mathcal{R}_{\neg\mathcal{S}}(q|\mathcal{G}) \leq \mathcal{R}(q|\mathcal{G}) + \left(\frac{n}{n-m}\right) \sqrt{\frac{2(n-m+1)}{mn} \left(\ln(|\mathcal{H}'|) + \ln(1/\delta)\right)}.$$

Plausible Labeling Classes $\mathcal{H}'$

Which labelings $\mathcal{H}' = \{q\}$ are supported by a graph $\mathcal{G}$? Can be formalized as e.g.

- Which labelings $\mathcal{H}'$ can be *reconstructed* ('predicted') with a rule?
- Which labelings $\mathcal{H}'$ can be *compressed* with respect to a rule?
- Which labelings $\mathcal{H}'$ *correlate* with the graph structure?
- Which labelings $\mathcal{H}'$ are *stable* with respect to subsampling of the graph
- Which labelings $\mathcal{H}'$ have large 'between/within cluster' ratio?

Plausible Labeling Classes $\mathcal{H}'$

Which labelings $\mathcal{H}' = \{q\}$ are supported by a graph $\mathcal{G}$? Can be formalized as e.g.

- Which labelings $\mathcal{H}'$ can be *reconstructed* ('predicted') with a rule?
- Which labelings $\mathcal{H}'$ can be *compressed* with respect to a rule?
- Which labelings $\mathcal{H}'$ *correlate* with the graph structure?
- Which labelings $\mathcal{H}'$ are *stable* with respect to subsampling of the graph
- Which labelings $\mathcal{H}'$ have large 'between/within cluster' ratio?

1. Cardinality: $|\mathcal{H}'|$ (finite setting!)
2. Covering balls (if many hypotheses in $q \in \mathcal{H}'$ similar)
3. Kingdom Dimension (VC-dim) of $\mathcal{H}'$:

$$\max_{\mathcal{S}} |\mathcal{S}| \text{ s.t. } \forall p \in \{-1, 1\}^{|\mathcal{S}|}, \exists q \in \mathcal{H}' : \quad p = q|_{\mathcal{S}}$$

   where $q|_{\mathcal{S}}$ denotes q restricted to $\mathcal{S}$.
4. Compression coefficient
5. Rademacher complexity

$$R(\mathcal{H}') = E\left[\sup_{q \in \mathcal{H}'} \frac{2}{n} \left| \sum_{i=1}^{n} \sigma_i q_i \right| \right]$$

1. Cardinality: $|\mathcal{H}'|$ (finite setting!)
2. Covering balls (if many hypotheses in $q \in \mathcal{H}'$ similar)
3. Kingdom Dimension (VC-dim) of $\mathcal{H}'$:

$$\max_{\mathcal{S}} |\mathcal{S}| \text{ s.t. } \forall p \in \{-1,1\}^{|\mathcal{S}|}, \exists q \in \mathcal{H}' : \quad p = q|_{\mathcal{S}}$$

   where $q|_{\mathcal{S}}$ denotes q restricted to $\mathcal{S}$.
4. Compression coefficient
5. Rademacher complexity

$$R(\mathcal{H}') = E\left[\sup_{q\in\mathcal{H}'} \frac{2}{n}\left|\sum_{i=1}^{n} \sigma_i q_i\right|\right]$$

Plausible Labeling Classes $\mathcal{H}'$ (Ct'd)

Measuring the Richness of $\mathcal{H}'$?

1. Cardinality: $|\mathcal{H}'|$ (finite setting!)
2. Covering balls (if many hypotheses in $q \in \mathcal{H}'$ similar)
3. Kingdom Dimension (VC-dim) of $\mathcal{H}'$:

$$\max_{\mathcal{S}} |\mathcal{S}| \text{ s.t. } \forall p \in \{-1, 1\}^{|\mathcal{S}|}, \exists q \in \mathcal{H}' : \quad p = q|_{\mathcal{S}}$$

   where $q|_{\mathcal{S}}$ denotes q restricted to $\mathcal{S}$.

4. Compression coefficient
5. Rademacher complexity

$$R(\mathcal{H}') = E\left[\sup_{q \in \mathcal{H}'} \frac{2}{n} \left| \sum_{i=1}^{n} \sigma_i q_i \right| \right]$$

Plausible Labeling Classes $\mathcal{H}'$ (Ct'd)

Measuring the Richness of $\mathcal{H}'$?

1. Cardinality: $|\mathcal{H}'|$ (finite setting!)
2. Covering balls (if many hypotheses in $q \in \mathcal{H}'$ similar)
3. Kingdom Dimension (VC-dim) of $\mathcal{H}'$:

$$\max_{\mathcal{S}} |\mathcal{S}| \text{ s.t. } \forall p \in \{-1, 1\}^{|\mathcal{S}|}, \exists q \in \mathcal{H}' : \quad p = q|_{\mathcal{S}}$$

   where $q|_{\mathcal{S}}$ denotes q restricted to $\mathcal{S}$.
4. Compression coefficient
5. Rademacher complexity

$$R(\mathcal{H}') = E\left[\sup_{q \in \mathcal{H}'} \frac{2}{n} \left| \sum_{i=1}^{n} \sigma_i q_i \right| \right]$$

Plausible Labeling Classes $\mathcal{H}'$ (Ct'd)

Measuring the Richness of $\mathcal{H}'$?

1. Cardinality: $|\mathcal{H}'|$ (finite setting!)
2. Covering balls (if many hypotheses in $q \in \mathcal{H}'$ similar)
3. Kingdom Dimension (VC-dim) of $\mathcal{H}'$:

$$\max_{\mathcal{S}} |\mathcal{S}| \text{ s.t. } \forall p \in \{-1, 1\}^{|\mathcal{S}|}, \exists q \in \mathcal{H}' : \quad p = q|_{\mathcal{S}}$$

   where $q|_{\mathcal{S}}$ denotes q restricted to $\mathcal{S}$.

4. Compression coefficient
5. Rademacher complexity

$$R(\mathcal{H}') = E\left[\sup_{q \in \mathcal{H}'} \frac{2}{n} \left| \sum_{i=1}^{n} \sigma_i q_i \right| \right]$$

Definition (Clusterwise constant hypothesis)

Assume a clustering $C^k = \{C_1, \ldots, C_k\}$ such that $C_i \cap C_j = \Phi$,

$$\mathcal{H}_{C^k} = \left\{ q \in \{-1, 1\}^n \;\middle|\; q_{C_i} = 1_{|C_i|} \text{ or } q_{C_i} = -1_{|C_i|} \quad \forall i, \text{ and } q_j = -1 \textit{elsewhere} \right\}$$

- $\text{VCdim}(\mathcal{H}_{C^k}) = k$
- Rademacher complexity $R(\mathcal{H}_{C^k})$ dependent on size clusters!
- Rademacher complexity related to normalization factor as in (Lange, 2004)

Definition (Clusterwise constant hypothesis)

Assume a clustering $C^k = \{C_1, \ldots, C_k\}$ such that $C_i \cap C_j = \Phi$,

$$\mathcal{H}_{C^k} = \left\{ q \in \{-1,1\}^n \;\middle|\; q_{C_i} = 1_{|C_i|} \text{ or } q_{C_i} = -1_{|C_i|} \quad \forall i, \quad \text{and } q_j = -1 elsewhere \right\}$$

- $\text{VCdim}(\mathcal{H}_{C^k}) = k$
- Rademacher complexity $R(\mathcal{H}_{C^k})$ dependent on size clusters!
- Rademacher complexity related to normalization factor as in (Lange, 2004)

"For any vertex v of a graph $\mathcal{G}$, the closed neighborhood $N(v)$ of v is the set of all vertices of $\mathcal{G}$ adjacent to v. We say that a set $\mathcal{S}$ of vertices is shattered if every subset $R \subset \mathcal{S}$ can be realized as $R = \mathcal{S} \cup N(v)$ for some vertex v of $\mathcal{G}$. The VCdim of $\mathcal{G}$ is defined to be the largest cardinality of a shattered set of vertices." [Haussler and Welz, 1987]

- Let the neighbor-based labeling q corresponding to $v \in V$ be defined as q_v:

$$q_{v,j} = \begin{cases} 1 & \exists e(v, v_j) \\ -1 & \text{else} \end{cases}$$

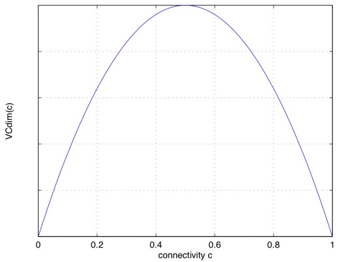
- Hypothesis set

$$\mathcal{H}' = \{q_v, \forall v \in V\}$$

- The VC-dim (Kingdom Dimension) of $\mathcal{H}'$:

$$\max_{\mathcal{S}} |\mathcal{S}| \text{ s.t. } \forall p \in \{-1, 1\}^{|\mathcal{S}|}, \exists v \in V : \quad p = q_v|_{\mathcal{S}}$$

where $q|_{\mathcal{S}}$ denotes q restricted to $\mathcal{S}$.



"For any vertex v of a graph $\mathcal{G}$, the closed neighborhood $N(v)$ of v is the set of all vertices of $\mathcal{G}$ adjacent to v. We say that a set $\mathcal{S}$ of vertices is shattered if every subset $R \subset \mathcal{S}$ can be realized as $R = \mathcal{S} \cup N(v)$ for some vertex v of $\mathcal{G}$. The VCdim of $\mathcal{G}$ is defined to be the largest cardinality of a shattered set of vertices." [Haussler and Welz, 1987]

- Let the neighbor-based labeling q corresponding to $v \in V$ be defined as q_v:

$$q_{v,j} = \begin{cases} 1 & \exists e(v, v_j) \\ -1 & \text{else} \end{cases}$$

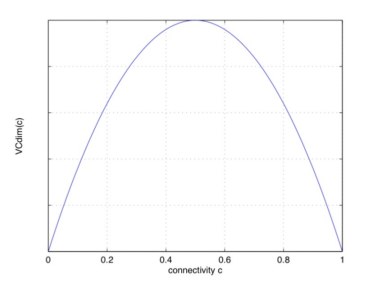
- Hypothesis set

$$\mathcal{H}' = \{q_v, \forall v \in V\}$$

- The VC-dim (Kingdom Dimension) of $\mathcal{H}'$:

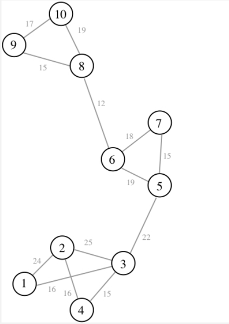
$$\max_{\mathcal{S}} |\mathcal{S}| \text{ s.t. } \forall p \in \{-1, 1\}^{|\mathcal{S}|}, \exists v \in V : \quad p = q_v|_{\mathcal{S}}$$

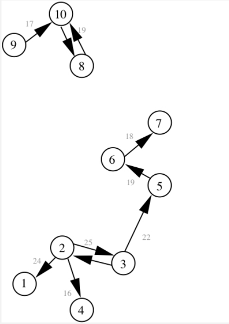
where $q|_{\mathcal{S}}$ denotes q restricted to $\mathcal{S}$.



- 1NN Predictor rule:
  $f_q(v_i) = q_{(i)}$ where $(i) = \arg\max_{j \neq i} x_{ij}$.
- Consistent predictor rule as restriction mechanism.

$$0 < q_i f_q(v_i) = q_i q_{(i)}, \quad \forall i = 1, \ldots, n$$





→ Kruskal's MSP algorithm, $VCdim(1NN) = \sharp$ disconnected components.

Plausible Labeling Classes $\mathcal{H}'$ (Ct'd)

Consider hypothesis set (Average Nearest Neighbors)

$$\mathcal{H}_k = \left\{ q \in \{-1,1\}^n : q^T L q \le k \right\}$$

Theorem (Cardinality of $\mathcal{H}_k$ (Pelckmans, Shawe-Taylor, 2007))

Let $\{\sigma_i\}_{i=1}^n$ denote the eigenvalues of the graph Laplacian $L = D - W$ where $D = \mathrm{diag}(d_1, \ldots, d_n) \in \mathbb{R}^{n\times n}$. The cardinality of the set $\mathcal{H}_k$ can then be bounded as

$$|\mathcal{H}_k| \le \sum_{d=0}^{n_\sigma(k)} \binom{n}{d} \le \left(\frac{en}{n_\sigma(k)} \right)^{n_\sigma(k)}, \tag{2}$$

where $n_\sigma(k)$ is defined as

$$n_\sigma(k) = \left| \{ \sigma_i : \ \sigma_i \le k \} \right| . \tag{3}$$

Plausible Labeling Classes $\mathcal{H}'$ (Ct'd)

Consider hypothesis set (Average Nearest Neighbors)

$$\mathcal{H}_k = \left\{ q \in \{-1, 1\}^n : q^T L q \leq k \right\}$$

Theorem (Cardinality of $\mathcal{H}_k$ (Pelckmans, Shawe-Taylor, 2007))

Let $\{\sigma_i\}_{i=1}^n$ denote the eigenvalues of the graph Laplacian $L = D - W$ where $D = \mathrm{diag}(d_1, \ldots, d_n) \in \mathbb{R}^{n \times n}$. The cardinality of the set $\mathcal{H}_k$ can then be bounded as

$$|\mathcal{H}_k| \leq \sum_{d=0}^{n_\sigma(k)} \binom{n}{d} \leq \left(\frac{en}{n_\sigma(k)} \right)^{n_\sigma(k)}, \tag{2}$$

where $n_\sigma(k)$ is defined as

$$n_\sigma(k) = |\{\sigma_i : \ \sigma_i \leq k\}| \,. \tag{3}$$

Lemma (Permutation Stability (R. El Yaniv, D. Pechyony, 2007))

Let $Z \in \mathcal{Z}$ be a random permutation vector. Let $f : \mathcal{Z} \to \mathbb{R}$ be an (m, n) symmetric permutation function satisfying $|f(Z) - f(Z^{ij})| \leq \beta$ for all (i, j) exchanging entries in $\mathcal{S}_m = \{1, \ldots, m\}$ and $\mathcal{S}_n = \{m + 1, \ldots, n\}$. Then

$$P\left(f(Z) - E[f(Z)] \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2\beta^2 K(m, n)}\right)$$

with $K(m, n) = (n - m)^2(H(n) - H(n - m))$ and $H(k) = \sum_{i=1}^{k} \frac{1}{i^2}$ (and hence $1/K(m, n) \geq m$).

Definition (Rademacher Complexity for TI)

Given hypothesis class $\mathcal{H}'$, one has

$$R(\mathcal{H}'|\mathcal{G}) = E\left[\sup_{q\in\mathcal{H}'} \frac{2}{n}\left|\sum_{i=1}^{n} \sigma_i q_i\right|\right]$$

with $\{\sigma_i\}_{i=1}^n$ Bernouilli random variables with $P(\sigma_i = -1) = P(\sigma_i = 1) = \frac{1}{2}$.

Assume further $\mathcal{H}' = -\mathcal{H}'$ (dropping $|\cdot|$).

Theorem (Rademacher bound)

With probability exceeding $1-\delta$ *for* $\delta > 0$, *on has for all* $q \in \mathcal{H}'$ *that*

$$\mathcal{R}_{\neg\mathcal{S}}(q|\mathcal{G}) \leq \mathcal{R}_{\mathcal{S}}(q|\mathcal{G}) + \left(\frac{n^2}{4m(n-m)}\right) R(\mathcal{H}'|\mathcal{G}) + \frac{n-2m}{2m(n-m)} + 2\sqrt{\left(\frac{n}{m(n-m)}\right)\log(1/\delta)}$$

Back to the clustering story...

Definition (Stability for Clustering)

Consider an algorithm $A : \mathcal{X} \to \mathcal{C}$. If $\exists\beta$ such that

$$d(A(S_m), A(S'_m)) \leq \beta, \ \ \forall S_m, S'_m \subset \{1, \ldots, n\}$$

then for any $\epsilon > 0$, one has

$$P\Big(d(A(S_m), E[A(S_m)]) \geq \epsilon\Big) \leq \delta(\epsilon)$$

for an appropriate $\delta : \mathbb{R}_0^+ \to]0, 1]$.

Stability in Clustering

Idea: encode $d(A(S_m), A(S'_m))$ as

$$\left|c(A(S_m)) - c(A(S'_m))\right|_1 ,$$

for an appropriate encoding function $c : \mathcal{C} \rightarrow \{0,1\}^{n_c}$, and $S_m = D(Z|m)$, or shortly as $Z|m$, then

Corollary (Stability for Clustering)

Consider an algorithm $A : \mathcal{X} \rightarrow \mathcal{C}$. If $\exists \beta(A, m)$ such that

$$\frac{1}{n_c}\left|c(A(Z|m)) - c(A(Z'|m))\right|_1 \leq \beta(A,m), \quad \forall S_m, S'_m \sim \{1,\ldots,n\}$$

then for any $\epsilon > 0$, one has

$$P\left(\frac{1}{n_c}\left|c(A(Z|m) - E_Z[c(A(Z|m))]\right|_1 \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2\beta^2(A,m)K(m,n)}\right)$$

with $K(m,n) = (n-m)^2(H(n) - H(n-m))$ and $H(k) = \sum_{i=1}^{k}\frac{1}{i^2}$ (and hence $1/K(m,n) \geq m$).

Stability in Clustering

Idea: encode $d(A(S_m), A(S'_m))$ as

$$\left|c(A(S_m)) - c(A(S'_m))\right|_1,$$

for an appropriate encoding function $c : \mathcal{C} \rightarrow \{0,1\}^{n_c}$, and $S_m = D(Z|m)$, or shortly as $Z|m$, then

Corollary (Stability for Clustering)

Consider an algorithm $A : \mathcal{X} \rightarrow \mathcal{C}$. If $\exists \beta(A, m)$ such that

$$\frac{1}{n_c}\left|c(A(Z|m)) - c(A(Z'|m))\right|_1 \leq \beta(A,m), \quad \forall S_m, S'_m \sim \{1,\ldots,n\}$$

then for any $\epsilon > 0$, one has

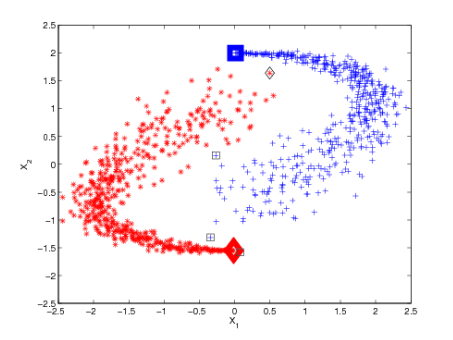
$$P\left(\frac{1}{n_c}\left|c(A(Z|m) - E_Z[c(A(Z|m))]\right|_1 \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2\beta^2(A,m)K(m,n)}\right)$$

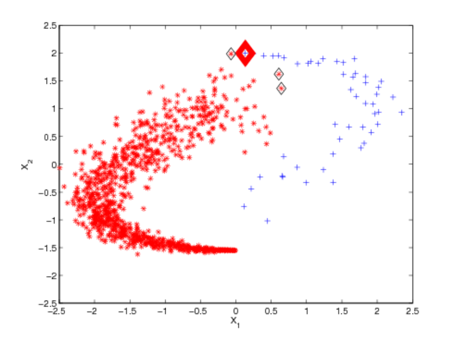
with $K(m,n) = (n-m)^2(H(n) - H(n-m))$ and $H(k) = \sum_{i=1}^{k}\frac{1}{i^2}$ (and hence $1/K(m,n) \geq m$).

Remarks

- Natural encoding by mapping points on canonical 'cluster identity'.
- McDiarmid inequality...
- Notions of weak stability
- Norm $\|\cdot\|_1$? Better choices with $\sup_{n_c}$
- Does k comes in only via $\beta(A, m)$?

Can be recast as a convex graph flow algorithm!





Incorporating prior knowledge by relaxing as an LP/QP - e.g. $\sum_{i=1}^{n}(1 - q_i) \geq B$ for $B \approx n$.

Conclusion

- ! Learning in finite universes!
- ! Clustering as setting the stage for prediction (particular sense)
- ! Clusterwise constant hypothesis class
- ! Stability results in Learning
- ? Cluster-ability and Learnability
- ? Falsifiable vs. Reproducible
- ? Need for a taxonomy?