Function class complexity and cluster structure with applications to transduction

Guy Lever¹

¹University 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)**

\[ M \leq O \left( \mathcal{N}(\mathcal{G}, \rho, r) + \text{cut}(h)\rho \right) \]

- Understand the role of the structure in data generally
(\mathcal{X}, d) \text{ a metric space}

**defn.** A *clustering* of $S \subset \mathcal{X}$ is any partition of $S$

$$C = \{C_1, ..., C_N\}$$

**defn.** the *center* of $C_k$

$$c_k := \arg\min_{x \in \mathcal{X}} \sum_{x' \in C_k} d^2(x', x)$$

For each $x \in S$, $c(x) := c_k$ where $k$ is such that $x \in C_k$
Identify vertex $v_i \in \mathcal{V}$ with standard basis vector $e_i$ in $\mathbb{R}^n$

$h \in \mathbb{R}^n$ classifies vertices $\mathcal{V} = \{v_1, \ldots, v_n\}$ via

$$h(v_i) := \text{sgn}(h^\top e_i) = \text{sgn}(h_i)$$

Graph “smoothness functional” (graph cut)

$$F_L(h) := \frac{1}{2} h^\top L h$$

$$= \frac{1}{2} \sum_{(i,j) \in \mathcal{E}} (h_i - h_j)^2 A_{ij}$$

$L$ is graph Laplacian, $A$ is adjacency

$H_\phi := \{h \in \{-1, 1\}^n : h^\top L h \leq \phi\}$
defn. empirical Rademacher complexity of $\mathcal{H} \subset \mathbb{R}^X$,

$$\hat{\mathcal{R}}_S(\mathcal{H}) := \mathbb{E}_\sigma \left[ \sup_{h \in \mathcal{H}} \left( \frac{1}{m} \sum_{i=1}^{m} h(x_i)\sigma_i \right) \right]$$

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

defn. Rademacher complexity $\mathcal{R}_m(\mathcal{H}) := \mathbb{E}_S(\hat{\mathcal{R}}_S(\mathcal{H}))$

Typically sharper than VC bounds

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

Data-dependent measure of complexity...

e.g. consider $\mathcal{R}_m(\mathcal{H}_\phi)$ vs. $\text{VCdim}(\mathcal{H}_\phi)$ on $(n, \sqrt{n})$-lollipop:
\( \mathcal{H} \) class of linear functions on \( \mathcal{X} \)

- **defn.** Norm \( \| \cdot \| \) on \( \mathcal{H} \) defines *implied metric* on \( \mathcal{X} \)

\[
d(x_i, x_j) : = \| x_i - x_j \|^* \\
= \sup_{h \in \mathcal{H}} \frac{|h(x_i) - h(x_j)|}{\| h \|}.
\]

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

\[
d_K(x, x') : = \sqrt{K(x, x) + K(x', x') - 2K(x, x')}.
\]
Resistance geometry on $\mathcal{G}$

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

\[
d_L(v_i, v_j) := \|e_i - e_j\|^* = \sqrt{(e_i - e_j)^\top L^+(e_i - e_j)}
\]

- Edges identified as resistors

$\quad d_L(B, C) < d_L(A, B)$

Geometry defined by the data

Relate learning to intrinsic structure of data
Rademacher complexity and cluster structure

- $F : \mathcal{H} \to \mathbb{R}_{\geq 0}$ is $\kappa$-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 $X$

**Theorem (refinement of Kakade et. al. 2008)**

For sample $S = \{x_1, \ldots, x_m\}$, all clusterings $\mathcal{C}$ of $S$, all $\alpha > 0$,

$$\hat{\mathcal{R}}_S(\mathcal{H}_\alpha) \leq B \sqrt{\frac{|\mathcal{C}|}{m}} + \sqrt{\frac{2 \alpha \rho_S}{m \kappa}}$$

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

- 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
Theorem

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

$$R_m(\mathcal{H}_\alpha) \leq \mathbb{B} \mathbb{E}_S \left[ \sqrt{\frac{|S|}{m}} \right] + \sqrt{\frac{2\alpha}{m\kappa}} \mathbb{E}_S[\sqrt{\rho_S}]$$

where $\mathcal{C}_S := \{C_k \in \mathcal{C} : S \cap C_k \neq \emptyset\}$ is the clustering restricted to the sample $S$.

- 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

• \( d_L^2(A, B) = 1 \)
• \( d_L^2(A, B) = \frac{2}{3} \)
• \( d_L^2(A, B) = \frac{2}{4} \)
• \( d_L^2(A, B) = \frac{2}{5} \)
• \( d_L^2(A, B) = \frac{2}{6} = O\left(\frac{1}{n}\right) \)

• 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

\[
\begin{align*}
  d^2_L(A, B) &= 1 \\
  d^2_L(A, B) &= \frac{2}{3} \\
  d^2_L(A, B) &= \frac{2}{4} \\
  d^2_L(A, B) &= \frac{2}{5} \\
  d^2_L(A, B) &= \frac{2}{6} = \mathcal{O}\left(\frac{1}{n}\right)
\end{align*}
\]

- non-empirical metrics not as sensitive to clustering: not 
  \textit{distribution dependent}
Is clustering an improvement?

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

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

- 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

\[
\begin{align*}
    d_L^2(A, B) &= 1 \\
    d_L^2(A, B) &= \frac{2}{3} \\
    d_L^2(A, B) &= \frac{2}{4} \\
    d_L^2(A, B) &= \frac{2}{5} \\
    d_L^2(A, B) &= \frac{2}{6} = \mathcal{O}\left(\frac{1}{n}\right)
\end{align*}
\]

- 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

\[
\begin{align*}
&d_L^2(A, B) = 1 \\
&d_L^2(A, B) = \frac{2}{3} \\
&d_L^2(A, B) = \frac{2}{4} \\
&d_L^2(A, B) = \frac{2}{5} \\
&d_L^2(A, B) = \frac{2}{6} = O\left(\frac{1}{n}\right)
\end{align*}
\]

- Non-empirical metrics not as sensitive to clustering: not distribution dependent
Test set $\mathcal{T}$ and training set $S$ presented simultaneously

- $S$ drawn uniformly without replacement from $\mathcal{X} = S \cup \mathcal{T}$

- $\mathcal{H}_\phi := \{ h \in \{-1, 1\}^n : h^\top L h \leq \phi \}$

**Corollary**

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

$$R_m^{\text{trs}}(\mathcal{H}_\phi) \leq \mathbb{E}_S \left[ \sqrt{\frac{|C_S|}{m}} \right] + \sqrt{\frac{\phi \rho}{m}}$$

where $\rho := \frac{1}{n} \sum_{i=1}^{n} d_L^2(v_i, c(v_i))$ and $C_S := \{ C_k \in \mathcal{C} : S \cap C_k \neq \emptyset \}$ is the clustering restricted to the sample $S$.

- Relates learning on graph to clustering in resistance
Comparison to VC dimension 1 - lollipops and barbells

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

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

$$R_{m}^{\text{trs}}(\mathcal{H}_\phi) \leq \sqrt{\frac{2}{m}} + \sqrt{\frac{\phi}{mn}}$$

- 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
    \[
    \|h\|_p := \left( \sum_{(i,j) \in \mathcal{E}} |h_i - h_j|^p \right)^{\frac{1}{p}}
    \]
  - $p$-resistance: \( d_p(v_i, v_j) := \|e_i - e_j\|_p^* \)
- $p$ resistance as $p \to 1$ more suitable for sparse graphs
Transductive risk analysis

- **defn.** Transductive risk \( \text{risk}_T(h) := \frac{1}{u} \sum_{i=1}^{u} \ell(h(x_{t_i}), y_{t_i}) \)

(loss on test set \( T = \{(X_{t_1}, Y_{t_1}), \ldots, (X_{t_u}, Y_{t_u})\} \))

**Theorem**

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

\[
\text{risk}_T(h) - \hat{\text{risk}}_S(h) \leq O \left( \frac{n}{u} \left( \mathbb{E}_S \left[ \sqrt{\frac{|C_S|}{m}} \right] + \sqrt{\frac{F_L(h) \rho}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{m}} \right) \right)
\]

where \( \rho = \frac{1}{n} \sum_{i=1}^{n} d_L^2(v_i, c(v_i)) \) and \( C_S = \{C_k \in C : S \cap C_k \neq \Phi\} \)

- Suitable for e.g. mincut, TSVM, regularization of Belkin and Niyogi, Energy minimization of Zhu, Pelckmans and Shawe-Taylor etc.
- Suggests algorithms obtained by minimising over clusterings and classifiers (and \( \rho \))
**Theorem (Hanneke 2006)**

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

$$\text{risk}_T(h) \leq \text{risk}_S(h) + O\left(\sqrt{\frac{n(u + 1)}{u^2}} \frac{F_L(h)}{\phi^*} \ln n + \ln \frac{1}{\delta}\right)$$

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

- New bounds preferred for highly clustered graphs
Theorem (Pelckmans and Shawe-Taylor 2007)

With probability at least $1 - \delta$,

$$\sup_{h \in \mathcal{H}_\phi} |\text{risk}_T(h) - \hat{\text{risk}}_S(h)| \leq \sqrt{\frac{2(n - m + 1)}{nm} \log \frac{\mathcal{H}_\phi}{\delta}}$$

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 $\mathcal{I} = \{(X_1, y_1), \ldots (X_m, y_m), X_{m+1}, \ldots X_n\}$

**Theorem**

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

\[
\text{risk}^\ell(h) \leq \widehat{\text{risk}}_S^\ell(h) + \mathcal{O} \left( \mathcal{R}_{m}^{\text{trs}}(\tilde{\mathcal{H}}_\beta) + \widehat{\mathcal{R}}_{\mathcal{I}}^{\text{ind}}(\mathcal{H}_\alpha) + \sqrt{\frac{1}{m \log \frac{1}{\delta}}} \right)
\]

\[
\mathcal{R}_{m}^{\text{trs}}(\tilde{\mathcal{H}}_\beta) \leq \mathcal{O} \left( \sqrt{\frac{|\mathcal{C}|}{m}} + \sqrt{\frac{\beta}{m n} \sum_{x \in \mathcal{I}} d_F^2(x, c(x))} \right)
\]

\[
\widehat{\mathcal{R}}_{\mathcal{I}}^{\text{ind}}(\mathcal{H}_\alpha) \leq \mathcal{O} \left( \sqrt{\frac{|\mathcal{C}'|}{n}} + \frac{1}{n} \sqrt{\alpha \sum_{x \in \mathcal{I}} d_{\tilde{F}}^2(x, c'(x))} \right)
\]

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