Advanced Machine Learning

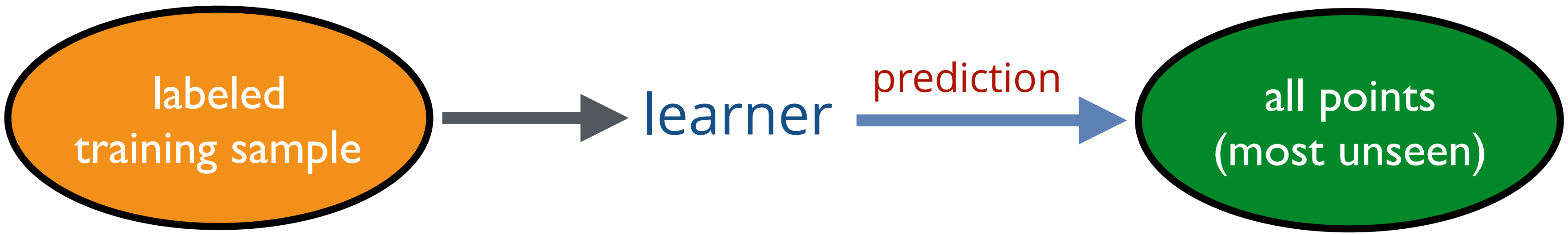
Transduction

MEHRYAR MOHRI MOHRI@

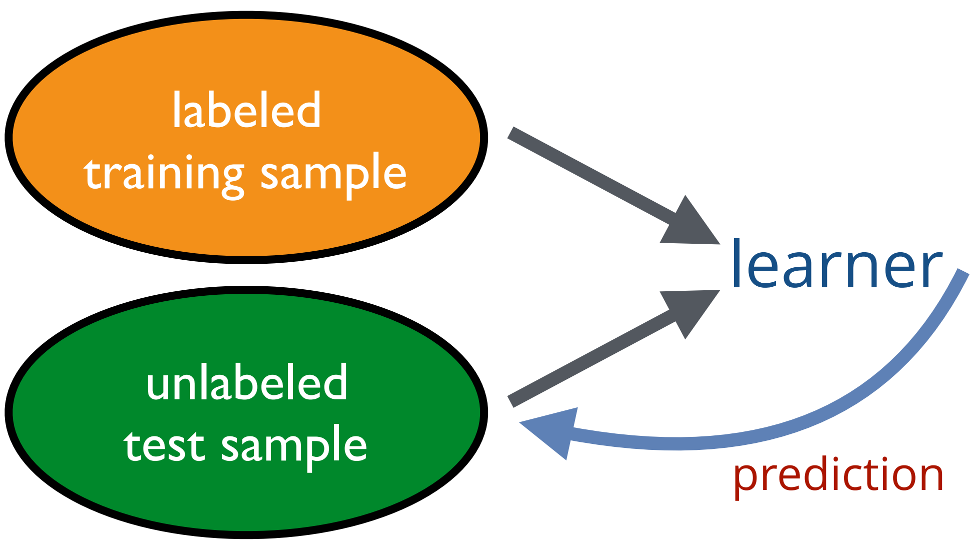
COURANT INSTITUTE & GOOGLE RESEARCH

Induction vs Transduction

- Inductive scenario:



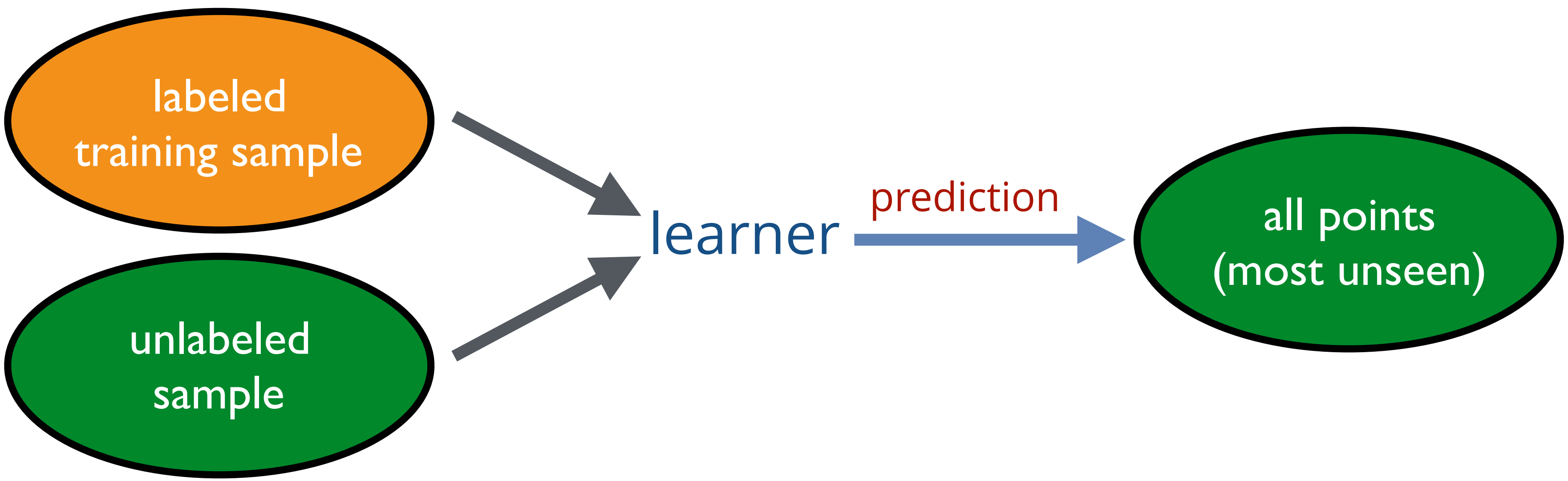
- Transductive scenario (Vapnik, 1998):



Semi-Supervised Learning

- Inductive scenario:

- Semi-supervised learning scenario:



Motivation

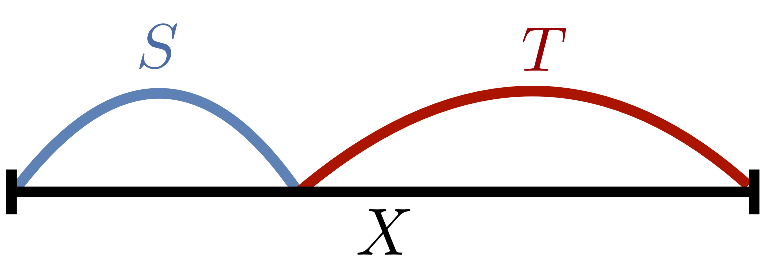
- Common scenario in many applications:
  - network predictions in computational biology.
  - web graph predictions.
  - NLP applications.
- Seemingly more favorable scenario than induction:
  - but can we (provably) benefit from that?
  - analysis of generalization in transductive setting.
  - transductive learning algorithms.

Outline

- Transduction scenario.
- Generalization bounds.
- Examples of algorithms.

Setting One

- A full sample X of size $(m+u)$ is fixed.
- The learner receives:
    - a sample $S = (x_1, \ldots, x_m)$ drawn uniformly without replacement from X as well as the labels $(y_1, \ldots, y_m)$.
    - an unlabeled test sample $T = (x_{m+1}, \ldots, x_{m+u})$ formed by the remaining points of X.



Setting One

- Loss function L taking values in $[0, 1]$.
- Hypothesis set H.
- Errors: for a hypothesis $h \in H$,
  - training error: $\widehat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} L(h(x_i), y_i)$.
  - test error: $R_T(h) = \frac{1}{u} \sum_{i=1}^{u} L(h(x_{m+i}), y_{m+i})$.
  - full sample error (not a random variable):

$$R(h) = \frac{1}{m+u} \sum_{i=1}^{m+u} L(h(x_i), y_i) = \frac{1}{m+u} \Big[m\widehat{R}_S(h) + uR_T(h) \Big].$$

Setting Two

- Distribution D over input space X.
- The learner receives:
  - a sample S of size m drawn i.i.d. from D^m as well as the corresponding labels.
  - a sample T of size u drawn i.i.d. from D^u.

Relationship btw Settings

- Any generalization bound for setting one implies a generalization bound for setting two by taking the expectation:

$$\underset{S\sim D^m, T\sim D^u}{\mathrm{E}}\left[1_{\{\sup_{h\in H} R_T(h)-\widehat{R}_S(h)>\epsilon\}}\right] = \underset{X\sim D^{m+u}}{\mathrm{E}}\left[\underset{(S,T)=X}{\mathrm{E}}\left[1_{\{\sup_{h\in H} R_T(h)-\widehat{R}_S(h)>\epsilon\}}\right]\right].$$

→ we will study generalization in setting one.

Outline

- Transduction scenario.
- Generalization bounds.
- Examples of algorithms.

Generalization Bounds

- VC-dimension bounds (Vapnik, 1998; Cortes and MM 2007).
- PAC-Baeysian bounds (Derbeko, El-Yaniv, and Meir, 2004).
- Stability bounds (El-Raniv and Pechyony 2008; Cortes, MM, Pechyony, Rastogi, 2008 and 2009).
- Rademacher complexity bounds (El-Raniv and Pechyony 2007).

McDiarmid's Inequality

(McDiarmid, 1989; corollary 6.10)

- Theorem: let $X_1, \ldots, X_m$ be random variables taking values in X and let $\Phi\colon X^m \to \mathbb{R}$ be a measurable function. Assume that there exist constants $c_1, \ldots, c_m$ such that

$$\Big| \mathrm{E}\left[\Phi(X_1^m) | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_i = x_i\right] \\ - \mathrm{E}\left[\Phi(X_1^m) | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_i = x'_i\right] \Big| \leq c_i,$$

for all $i \in [1, m]$ and $x_1^m, x_1'^m \in X^m$. Then, for any $\epsilon > 0$,

$$\Pr[|\Phi(X_1^m) - \mathrm{E}[\Phi(X_1^m)]| > \epsilon] \leq 2 \exp\left[\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right].$$

Sampling w/o Replacement

(Cortes, MM, Pechyony, Rastogi, 2008 & 2009)

- Theorem: let $X_1, \ldots, X_m$ be a sequence of r.v.'s distributed according to the uniform distribution without replacement from a set X of size $m+u$ and let $\Phi\colon X^m \to \mathbb{R}$ be a symmetric measurable function. Assume that there exists a constant c such that

$$\left|\Phi(x_1^m) - \Phi(x_1^{i-1}, x_i', x_{i+1}^m)\right| \leq c$$

for all $i \in [1, m]$ and $x_1^m, x_1'^m \in X^m$. Then, for any $\epsilon > 0$,

$$\Pr[|\Phi(X_1^m) - \mathrm{E}[\Phi(X_1^m)]| > \epsilon] \leq 2\exp\left[\frac{-2\epsilon^2}{\alpha(m,u)c^2}\right],$$

with $\alpha(m,u) = \frac{mu}{m+u-1/2}\,\frac{1}{1-1/(2\max\{m,u\})}$.

Proof

- For any $i \in [1, m]$,

$$
\begin{aligned}
&\mathrm{E}\left[\Phi(X_1^m)|X_1^i = x_1^i\right] - \mathrm{E}\left[\Phi(X_1^m)|X_1^{i-1} = x_1^{i-1}, X_i = x_i'\right] \\
&= \sum_{x_{i+1}^m} \Pr[X_{i+1}^m = x_{i+1}^m | X_1^i = x_1^i]\, \Phi(x_1^{i-1}, x_i, x_{i+1}^m) \\
&\quad - \sum_{x'^m_{i+1}} \Pr[X_{i+1}^m = x_{i+1}^m | X_1^{i-1} = x_1^{i-1}, X_i = x_i']\, \Phi(x_1^{i-1}, x_i', x'^m_{i+1}) \\
&= \left[\prod_{k=i}^{m-1} \frac{1}{m+u-k}\right] \left[\sum_{x_{i+1}^m} \Phi(x_1^{i-1}, x_i, x_{i+1}^m) - \sum_{x'^m_{i+1}} \Phi(x_1^{i-1}, x_i', x'^m_{i+1})\right] \\
&= \frac{u!}{(m+u-i)!} \left[\sum_{x_{i+1}^m} \Phi(x_1^{i-1}, x_i, x_{i+1}^m) - \sum_{x'^m_{i+1}} \Phi(x_1^{i-1}, x_i', x'^m_{i+1})\right].
\end{aligned}
$$

Proof

- Two cases:
  - x'^{m}_{i+1} contains x_i: then there is (a unique) x^{m}_{i+1} such that $\{x'_i x'^{m}_{i+1}\} = \{x_i x^{m}_{i+1}\}$ and the corresponding terms cancel out by the symmetry of Φ.
  - x'^{m}_{i+1} does not contain x_i : then there is (a unique) x^{m}_{i+1} such that $\{x'_i x'^{m}_{i+1}\}$ differs from $\{x_i x^{m}_{i+1}\}$ by $x_i \neq x'_i$. By assumption, the corresponding terms differ in absolute value by at most c.
- Second case instances: number of x'^{m}_{i+1} permutations chosen out of the set $X - \{x_1^{i-1}, x_i, x'_i\}$:

$$\frac{(m+u-i-1)!}{(m+u-i-1-(m-i))!} = \frac{(m+u-i-1)!}{(u-1)!}.$$

Proof

- Thus,

$$\Big| \mathrm{E}\left[\Phi(X_1^m)|X_1^i = x_1^i\right] - \mathrm{E}\left[\Phi(X_1^m)|X_1^{i-1} = x_1^{i-1}, X_i = x_i'\right] \Big|$$

$$\leq \frac{u!}{(m+u-i)!} \frac{(m+u-i-1)!}{(u-1)!} c = \frac{uc}{m+u-i}.$$

- The term in McDiarmid's inequality is bounded as follows:

$$\sum_{i=1}^{m} \frac{u^2c^2}{(m+u-i)^2} = \sum_{j=u}^{m+u-1} \frac{u^2c^2}{j^2} \leq \int_{u-1/2}^{m+u-1/2} \frac{u^2c^2dx}{x^2} = \frac{muc^2}{m+u-1/2} \frac{u}{u-1/2}.$$

- The theorem follows by observing that m and u can be permuted by the symmetry of Φ.

Generalization Bound

- Theorem: for any $\delta > 0$, with probability at least $1-\delta$, for all $h \in H$,

$$R_T(h) \leq \widehat{R}_S(h) + \mathrm{E}[\Phi(S)] + \sqrt{\frac{\eta}{2}\left[\frac{1}{m} + \frac{1}{u}\right] \log \frac{1}{\delta}}\,,$$

where $\eta = \frac{m+u}{m+u-\frac{1}{2}} \frac{1}{1-\frac{1}{2\max\{m,u\}}}$.

- Proof: apply concentration bound to

$$\Phi(S) = \sup_{h \in H} R_T(h) - \widehat{R}_S(h).$$

  - observe that

$$|\Phi(S') - \Phi(S)| \leq \frac{1}{m} + \frac{1}{u} = \frac{m+u}{mu}.$$

Rademacher Complexity

- Define random variable σ_i as taking value
  - $\frac{m+u}{u}$ with probability $\frac{u}{m+u}$.
  - $-\frac{m+u}{m}$ with probability $\frac{m}{m+u}$.
- Definition: the transductive Rademacher complexity of G is

$$\mathfrak{R}_{m+u}(G) = \frac{1}{m+u} \operatorname*{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in G} \sum_{i=1}^{m+u} \sigma_i g(x_i) \right].$$

  - note: simpler definition than (El-Yaniv and Pechyony 2007).

Analysis

- For any $N \in \left[-\frac{(m+u)^2}{m}, \frac{u(m+u)^2}{u}\right]$, define

$$R(N) = \frac{1}{m+u} \operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in G} \sum_{i=1}^{m+u} \sigma_i g(x_i) \,\middle|\, \sum_{i=1}^{m+u} \sigma_i = N\right].$$

- Observe that if $\sum_{i=1}^{m+u} \sigma_i = 0$ and n σ_is take value $\frac{m+u}{u}$, then

$$n\frac{m+u}{u} - (m+u-n)\frac{m+u}{m} = 0 \Leftrightarrow n = u.$$

Thus, $\operatorname{E}_S[\Phi(S)] = R(0)$.

Analysis

- For any n

$$N = \sum_{i=1}^{m+u} \sigma_i = n\frac{m+u}{u} - (m+u-n)\frac{m+u}{m} = \frac{(m+u)^2}{mu}(n-u).$$

- Let $n_2 \geq n_1$,

$$R(N_1) = \frac{1}{m+u} \operatorname{E}\left[\sup_{g \in G} \sum_{i=1}^{n_1} \frac{m+u}{u} g(x_i) - \sum_{i=n_1+1}^{m+u} \frac{m+u}{m} g(x_i)\right].$$

$$R(N_2) = \frac{1}{m+u} \operatorname{E}\left[\sup_{g \in G} \sum_{i=1}^{n_1} \frac{m+u}{u} g(x_i) - \sum_{i=n_1+1}^{m+u} \frac{m+u}{m} g(x_i)\right.$$
$$\left. + \sum_{i=n_1+1}^{n_2} \left[\frac{m+u}{u} + \frac{m+u}{m}\right] g(x_i)\right].$$

Analysis

- Lipschitz property:

$$|R(N_2) - R(N_1)| \leq |n_2 - n_1| \left(\frac{1}{m} + \frac{1}{u}\right) = \frac{|N_2 - N_1|}{m+u}.$$

- Thus, for $N = \sum_{i=1}^{m+u} \sigma_i$,

$$\Pr\Big[\big|R(N) - R(\mathrm{E}[N])\big| > \epsilon\Big] \leq \Pr\Big[\big|N - \mathrm{E}[N]\big| > (m+u)\epsilon\Big].$$

Transductive Rad. Comp. Bound

- Theorem: let H_L denote $\{x \mapsto L(h(x), f(x)) \colon h \in H\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $h \in H$,

$$R_T(h) \leq \widehat{R}_S(h) + \mathfrak{R}_{m+u}(H_L) + O\left(\sqrt{\min\{m,u\}}\left[\frac{1}{m} + \frac{1}{u}\right]\right) + \sqrt{\frac{\eta}{2}\left[\frac{1}{m} + \frac{1}{u}\right]\log\frac{1}{\delta}},$$

where $\eta = \frac{m+u}{m+u-\frac{1}{2}} \frac{1}{1 - \frac{1}{2\max\{m,u\}}}$.

Notes

- For large m, the bound varies only as $O(\frac{1}{\sqrt{u}})$: quite different from the induction scenario.
- H can be selected after measuring $\mathfrak{R}_{m+u}(H_L)$ since the full sample is accessible.

Transductive Stability Bound

- Theorem: let L be a loss function taking values in $[0, 1]$ and let $\mathcal{A}$ be a uniformly β-stable algorithm returning $h_S \in H$ when trained using labeled sample S. Then, for any $\delta > 0$, with probability at least $1 - \delta$,

$$R_T(h_S) \leq \widehat{R}_S(h_S) + \beta + \left(2\beta + \frac{(m+u)}{mu}\right)\sqrt{\frac{\alpha(m,u)\log\frac{1}{\delta}}{2}}.$$

Proof

- Define for any $h \in H$, $\Phi(S, h) = R_T(h) - \widehat{R}_S(h)$.
- Assume that S and S' differ by one point. Then,

$$
\begin{aligned}
&\Phi(S', h_{S'}) - \Phi(S, h_S) \\
&= \frac{1}{u} \sum_{i=1}^{u-1} L(h_{S'}(x_{m+i}), y_{m+i}) - L(h_S(x_{m+i}), y_{m+i}) \\
&\quad + \frac{1}{m} \sum_{i=1}^{m-1} L(h_{S'}(x_i), y_i) - L(h_S(x_i), y_i) \\
&\quad + \frac{1}{u} (L(h_{S'}(x'_{m+i}), y'_{m+i}) - L(h_S(x_{m+i}), y_{m+i})) \\
&\quad + \frac{1}{m} (L(h_{S'}(x'_m), y'_m) - L(h_S(x_m), y_m)).
\end{aligned}
$$

Proof

- Thus,

$$\Big|\Phi(S', h_{S'}) - \Phi(S, h_S)\Big| \leq \frac{\beta(u-1)}{u} + \frac{\beta(m-1)}{m} + \frac{1}{u} + \frac{1}{m} \leq 2\beta + \frac{1}{u} + \frac{1}{m}.$$

- Bounding the expectation:

$$\begin{aligned}
\operatorname*{E}_S[\Phi(S, h_S)] &= \frac{1}{u}\sum_{i=1}^{u} \operatorname*{E}_S[L(h_S(x_{m+i}), y_{m+i})] - \frac{1}{m}\sum_{i=1}^{m} \operatorname*{E}_S[L(h_S(x_i), y_i)] \\
&= \operatorname*{E}_{S, x' \notin S}[L(h_S(x'), y_{x'})] - \operatorname*{E}_{S, x \in S}[L(h_S(x), y_x)] \\
&= \operatorname*{E}_{S, x' \notin S}[L(h_{S-\{x\}\cup\{x'\}}(x), y_x) - L(h_S(x), y_x)] \leq \beta.
\end{aligned}$$

Outline

- Transduction scenario.
- Generalization bounds.
- Examples of algorithms.

Transductive SVM (TSVM)

(Vapnik, 2008), see also (Joachims, 1999)

- Optimization problem:

$$\min_{\mathbf{w}, b, \mathbf{y}_{m+1}^{m+u}} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{m} L(\mathbf{w} \cdot \mathbf{x}_i + b, y_i) + C' \sum_{i=1}^{u} L(\mathbf{w} \cdot \mathbf{x}_{m+i} + b, y_{m+i})$$

  - classification: hinge loss.
  - regression: trivial solution, last term vanishes! (Cortes and MM, 2007).
  - theoretical guarantee: unclear.
  - computational complexity: exponential.
  - experiments: issue of uniform labeling of test points in high dimension (Joachims, 1999); poor results (Tong and Oles, 1999).

Local Transductive Regression

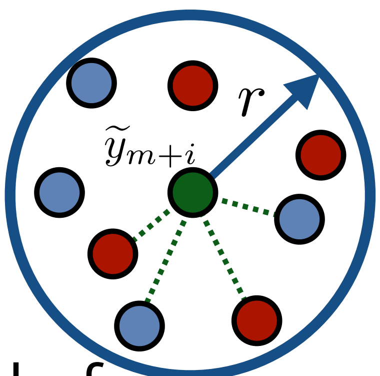
(Cortes, MM, Pechyony, Rastogi, 2008 & 2009)

- Optimization problem (LTR):

$$\min_{h \in \mathbb{H}} \|h\|_K^2 + C \sum_{i=1}^{m} L(h(x_i), y_i) + C' \sum_{i=1}^{u} L(h(x_{m+i}), \widetilde{y}_{m+i}),$$

with K a PDS kernel, and

$\widetilde{y}_{m+i}$s pseudo-labels obtained via local weighted average or any other local regression algorithm from neighborhood of radius r.



Stability Guarantee

- Theorem: assume that for all $x \in X$, $|y(x)| \leq M$ and that the local estimator has score-stability β_{loc}. Then, LTR has uniform stability

$$\beta \leq 2(C_0 M)^2 r^2 \left[\frac{C}{m} + \frac{C'}{u} + \sqrt{\left(\frac{C}{m} + \frac{C'}{u}\right)^2 + \frac{2C'\beta_{\text{loc}}}{C_0 M r^2 u}}\right],$$

with $r^2 = \sup_{x \in X} K(x, x)$ and $C_0 = 1 + r\sqrt{C + C'}$.

Graph Regularization Algo.

- Set-up:
  - weighted directed graph $G = (X, E)$.
  - hypothesis $h \colon X \to \mathbb{R}$ in H identified with vector

$$\mathbf{h} = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_{m+u}) \end{bmatrix}.$$

  - PSD matrix $\mathbf{L} \in \mathbb{R}^{(m+u)\times(m+u)}$ (similarity matrix).

Graph Regularization Algo.

- Optimization problem:

$$\min_{\mathbf{h}\in H} \mathbf{h}^\top \mathbf{L}\mathbf{h} + \frac{C}{m}(\mathbf{h}_S - \mathbf{y}_S)^\top(\mathbf{h}_S - \mathbf{y}_S)$$
$$\text{s.t.: } \mathbf{h}^\top \mathbf{u} = 0,$$

where $\mathbf{h}_S$ is the restriction of $\mathbf{h}$ to the training sample S and $\mathbf{y}_S$ the vector of training labels, and $\mathbf{u}$ a constant vector in $\mathbb{R}^{m+u}$.

Graph Regularization Algo.

- Example (Belkin et al., 2004):
  - graph assumed connected.
  - $\mathbf{L}$ is the graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where

$$\mathbf{D} = \operatorname{diag}\Big(\sum_{i=1}^{n} w_{1i}, \ldots, \sum_{i=1}^{n} w_{ni}\Big).$$

    Then, $\mathbf{h}^\top \mathbf{L} \mathbf{h} = \sum_{i \sim j} w_{ij} (h(x_i) - h(x_j))^2$.

  - $\mathbf{u} = (1, \ldots, 1)^\top$.
  - data assumed centered: $\mathbf{u}^\top \mathbf{y} = 0$, and graph connected.
  - $\longrightarrow$ zero eigenvalue of Laplacian has multiplicity one and the solutions $\mathbf{h}$ in $\operatorname{range}(\mathbf{L})$.

Graph Regularization Algo.

- Lagrangian:

$$\mathcal{L} = \mathbf{h}^\top \mathbf{L} \mathbf{h} + \frac{C}{m} (\mathbf{h}_S - \mathbf{y}_S)^\top (\mathbf{h}_S - \mathbf{y}_S) + \beta \mathbf{h}^\top \mathbf{u}.$$

- Differentiating and applying orthogonal projection to $\mathbf{u}$:

$$\mathbf{P}\left(\mathbf{L} + \frac{C}{m}\mathbf{I}_S\right)\mathbf{h} = \frac{C}{m}\mathbf{P}\mathbf{y}_S - \beta \mathbf{P}\mathbf{u} = \frac{C}{m}\mathbf{P}\mathbf{y}_S$$

$$\Rightarrow \mathbf{h} = \left[\mathbf{P}\left(\frac{m}{C}\mathbf{L} + \mathbf{I}_S\right)\right]^{-1} \mathbf{P}\mathbf{y}_S. \qquad \left(\mathbf{P}\left(\frac{m}{C}\mathbf{L} + \mathbf{I}_S\right) \text{ invertible}\right)$$

Stability Guarantee

(Cortes, MM, Pechyony, Rastogi, 2009)

- Theorem: assume that for all $h \in H$ and $x \in X$, $|h(x) - y_x| \leq M$. Then, the graph Laplacian regularization algorithm has uniform stability

$$\beta \leq \frac{4CM^2}{m} \min\left\{\frac{1}{\lambda_2}, \rho_G\right\},$$

where λ_2 is the smallest non-trivial eigenvalue of $\mathbf{L}$ and ρ_G the diameter of the graph (longest shortest path).

Proof

- The graph Laplacian algorithm can be shown to coincide with LTR with the kernel matrix $\mathbf{K} = \mathbf{L}^+$: for all $\mathbf{h} \in \mathrm{range}(\mathbf{L})$,

$$\mathbf{K}\mathbf{L}\mathbf{h} = \mathbf{L}^+\mathbf{L}\mathbf{h} = \mathbf{h}$$
$$\mathbf{h}'^\top \mathbf{L}\mathbf{K}\mathbf{L}\mathbf{h} = \mathbf{h}'^\top \mathbf{L}\mathbf{h}.$$

- The result follows by applying the stability bound for LTR with the bound on the $K(x, x)$ in terms of λ_2 and ρ_G.

Notes

- For a hypercube, $\lambda_2 = 2$.
- Does not perform well in experiments in comparison with LTR.

References

- Mikhail Belkin, Irina Matveeva, and Partha Niyogi. Regularization and semi-supervised learning on large graphs. In COLT, pages 624–638. Springer, 2004.
- Olivier Chapelle, Vladimir Vapnik, and Jason Weston. Transductive inference for estimating values of functions. In NIPS, pages 421–427. 1999.
- Corinna Cortes and Mehryar Mohri. On transductive regression. In NIPS, pages 305–312. 2007.
- Corinna Cortes, Mehryar Mohri, Dmitry Pechyony, and Ashish Rastogi. Stability of transductive regression algorithms. In ICML. 2008.
- Corinna Cortes, Mehryar Mohri, Dmitry Pechyony, and Ashish Rastogi. Stability analysis and learning bounds for transductive regression algorithms. ArXiv 0904.0814. 2009.
- Philip Derbeko, Ran El-Yaniv, and Ron Meir. Explicit learning curves for transduction and application to clustering and compression algorithms. J. Artif. Intell. Res. (JAIR), 22:117–142, 2004.

References

- Thorsten Joachims. Transductive Inference for Text Classification using Support Vector Machines. ICML 1999: 200-209.
- Colin McDiarmid. On the method of bounded differences. In Surveys in Combinatorics, pages 148–188. Cambridge University Press, Cambridge, 1989.
- Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. In COLT, 2007.
- Vladimir N. Vapnik. Statistical Learning Theory. Wiley-Interscience, New York, 1998.
- Zhang, Tong and Frank Oles. A probability analysis on the value of unlabeled data for classification problems. In ICML, pp. 1191–1198, 2000.