Advanced Machine Learning & Perception

Instructor: Tony Jebara
Topic 11

- Semi-Supervised Learning
- Exploiting Unlabeled Data
- Transduction
- Partially Labeled Data and EM
SVM Extensions

Classification
Detection
Feature/Kernel Selection

Regression
Transduction
Meta/Multi-Task Learning
Exploiting Unlabeled Data

- In many learning situations, labeling data is the most difficult and labor-intensive part so labels are limited.
- But, getting unlabeled data is cheap.
- Transduction: discriminative, find large margin region.

- Hidden Labels: use generative modeling to cluster data. Clusters have same labels.
- Diffusion: spreading labels across manifold via spectral, kernel, Markov walks methods.
Transduction

- Only min test error on test examples! Not all future test...
- As with regular SVM, minimize error on training
  but reduce generalization error term.
- Theorem: generalization error again
  depends on VC < D^2/M^2
- Again minimize by max margin (why?)
- Brute force:
  find largest
  margin over
  2^T settings of
  T test points
- C => labeled
- C* => unlabeled
- Impractical!

**OP 2** (Transductive SVM (non-sep. case))

Minimize over \((y_1^*, \ldots, y_n^*, \bar{w}, b, \xi_1, \ldots, \xi_n, \xi_1^*, \ldots, \xi_k^*)\):

\[
\frac{1}{2} \| \bar{w} \|^2 + C \sum_{i=0}^{n} \xi_i + C^* \sum_{j=0}^{k} \xi_j^*
\]

subject to:

\[
\begin{align*}
\forall_{i=1}^{n} : y_i \left[ \bar{w} \cdot \bar{x}_i + b \right] &\geq 1 - \xi_i \\
\forall_{j=1}^{k} : y_j^* \left[ \bar{w} \cdot \bar{x}_j^* + b \right] &\geq 1 - \xi_j^*
\end{align*}
\]

\[
\begin{align*}
\forall_{i=1}^{n} : \xi_i &> 0 \\
\forall_{j=1}^{k} : \xi_j^* &> 0
\end{align*}
\]
Transduction with SVMs

• First train regular SVM on \((x,y)\) labeled data
• Use SVM to classify unlabeled \((x^*,y^*)\) points
• Use current labeling to retrain with low \(C^*_+ \) & \(C^*_+\)

**OP 3 (Inductive SVM (primal))**

\[
\text{Minimize over } (\vec{w}, b, \xi, \xi^*):
\]

\[
\frac{1}{2} ||\vec{w}||^2 + C \sum_{i=1}^{n} \xi_i + C^*_+ \sum_{j:y_j^*=1} \xi_j^* + C^*_+ \sum_{j:y_j^*=-1} \xi_j^*
\]

subject to: 

\[
\forall_{i=1}^{n}: y_i [\vec{w} \cdot \vec{x}_i + b] \geq 1 - \xi_i
\]

\[
\forall_{j=1}^{k}: y_j^* [\vec{w} \cdot \vec{x}_j + b] \geq 1 - \xi_j^*
\]

• Interleave regular SVM solution with unlabeled label swaps
• Guaranteed

\[
(y_m^* y_i^* < 0) \ & \ (\xi_m^* > 0) \ & \ (\xi_i^* > 0) \ & \ (\xi_m^* + \xi_i^* > 2)
\]

• Slowly increase effect of unlabeled by \(C^*\) doubling ‘til max
Transduction with SVMs

Input:
- training examples \((\tilde{x}_1, y_1), \ldots, (\tilde{x}_n, y_n)\)
- test examples \(\tilde{x}_1^*, \ldots, \tilde{x}_k^*\)

Parameters:
- \(C_s, C^*: \) parameters from \(\text{OP}(2)\)
- \(\text{num}_+: \) number of test examples to be assigned to class +

Output:
- predicted labels of the test examples \(y_1^*, \ldots, y_k^*\)

\((\tilde{w}, b, \xi, \varsigma) := \text{solve_svm_qp}([[(\tilde{x}_1, y_1), \ldots, (\tilde{x}_n, y_n)], [], C, 0, 0]);\)

Classify the test examples using \(\langle \tilde{w}, b \rangle\). The \(\text{num}_+\) test examples with the highest value of \(\tilde{w} \cdot \tilde{x}_j^* + b\) are assigned to the class + \((y_j^* := 1)\); the remaining test examples are assigned to class - \((y_j^* := -1)\).

\[C_s^* := 10^{-5};\]
\[C_+^* := 10^{-5} + \frac{\text{num}_+}{k - \text{num}_+};\]

while \((C_s^* < C^*) \mid (C_+^* < C^*)\) \{ // Loop 1

\((\tilde{w}, b, \tilde{\xi}, \tilde{\varsigma}) := \text{solve_svm_qp}([[(\tilde{x}_1, y_1), \ldots, (\tilde{x}_n, y_n)], [(\tilde{x}_1^*, y_1^*), \ldots, (\tilde{x}_k^*, y_k^*)], C, C_s^*, C_+^*]);\)

while \((\exists m, l : (y_m^* \cdot y_l^* < 0) \&(\xi_m^* > 0) \&(\xi_l^* > 0) \&(\xi_m^* + \xi_l^* > 2))\) \{ // Loop 2

\[y_m^* := -y_m^*;\] // take a positive and a negative test
\[y_l^* := -y_l^*;\] // example, switch their labels, and retrain

\((\tilde{w}, b, \tilde{\xi}, \tilde{\varsigma}) := \text{solve_svm_qp}([[(\tilde{x}_1, y_1), \ldots, (\tilde{x}_n, y_n)], [(\tilde{x}_1^*, y_1^*), \ldots, (\tilde{x}_k^*, y_k^*)], C, C_s^*, C_+^*]);\)
\}

\[C_s^* := \min(C_s^* \cdot 2, C^*);\]
\[C_+^* := \min(C_+^* \cdot 2, C^*);\]
\}

return \((y_1^*, \ldots, y_k^*);\)
Transduction for Text

- In X vector each dim is word in language
- Stem: combine similar words physics, physician, => physic
- Remove stop words: and, the, ...
- Represent words by TF-IDF text freq times inv-doc freq

\[ X_j(w_i) = \left( \# w_i \text{ in } d_j \right) \times \log \left( \frac{\# d_j}{\# d_j \text{ where } \# w_i > 0} \right) \]

- Evaluate by P/R breakeven point (equal on ROC curve)
- Train multi-class SVM
- Map multi-class to a one versus all binary decision
Partially Labeled Data & EM

- Instead of maximizing likelihood of labeled data
  \[ l(\theta) = \sum_{i \in LAB} \log(p(x_i, y_i | \theta)) \]
- Or maximizing likelihood of unlabeled data (needs EM)
  \[ l(\theta) = \sum_{i \in UNLAB} \log\left(\sum_y p(x_i, y | \theta)\right) \]
- Maximize a combination of both weighted by \( \lambda \)
  \[ l(\theta) = \sum_{i \in LAB} \log\left(p(x_i, y_i | \theta)\right) + \lambda \sum_{i \in UNLAB} \log\left(\sum_y p(x_i, y | \theta)\right) \]
- Also, use a prior \( P(\theta) \) to help (avoids zero-counts in multinomial models)...
  \[ l(\theta) = \log p(\theta) + \sum_{i \in LAB} \log\left(p(x_i, y_i | \theta)\right) \]
  \[ + \lambda \sum_{i \in UNLAB} \log\left(\sum_y p(x_i, y | \theta)\right) \]
Partially Labeled Data & EM

- Estimate $\lambda$ by cross-validation
- Use multinomial model
- Like Naïve Bayes
- Generally improve accuracy on text problems