Advanced Machine Learning & Perception

Instructor: Tony Jebara

Topic 9

Semi-Supervised Learning

Felix X. Yu

- Semi-supervised SVM (S^3VM^{light})
- Generative Models (EM)
- Graph-based semi-supervised learning

Semi-supervised Learning

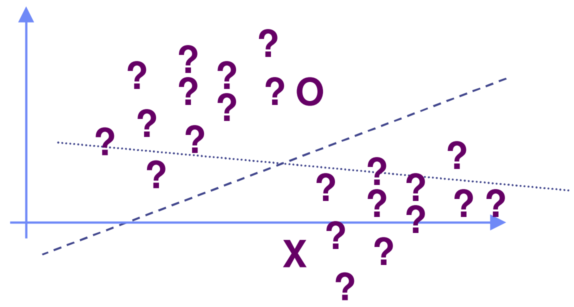
- What

Learning setting	Learning from ...
Supervised Learning	labeled data
Semi-supervised Learning	both labeled and unlabeled data
Unsupervised Learning	Unlabeled data

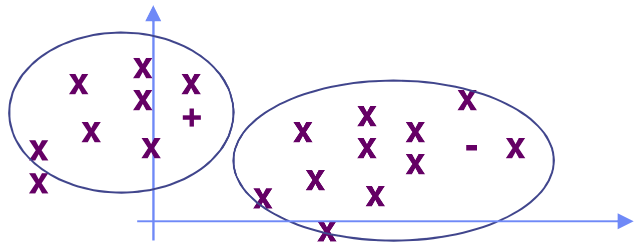
- Why
  - In many learning situations, labeling data is the most difficult and labor-intensive part so labels are limited.
  - But, getting unlabeled data is cheap.
  - Unlabeled data can help sometime.

How

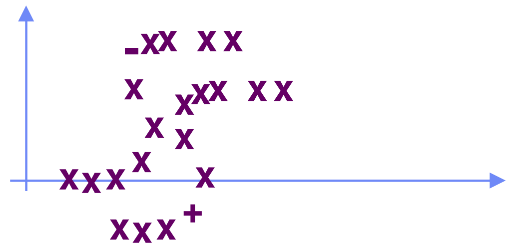
- Transduction: discriminative, find large margin region.



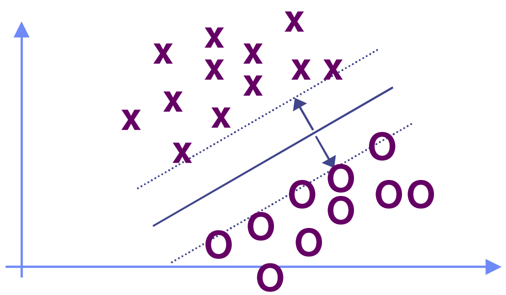
- Hidden Labels: use generative modeling to cluster data. clusters have same labels

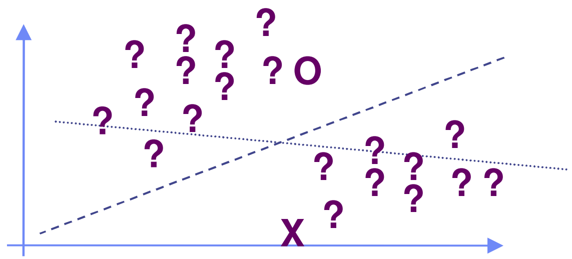


- Diffusion: spreading labels across manifold via spectral, kernel, Markov walks methods.



Semi-Supervised SVM (S^3VM)





Regular SVM for classification

Structured risk minimization

training set $(x_1, y_1), \ldots, (x_m, y_m)$

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h(log(\frac{2m}{h}+1) - log(\frac{\eta}{4})}{m}}$$

VC < D^2/M^2

SVM

$$\min_{w,b,\xi} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\xi_i$$

$$\text{subject to} \quad y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i)+b) \geq 1-\xi_i,$$
$$\xi_i \geq 0, i = 1, \ldots, l,$$

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^T Q\boldsymbol{\alpha} - \boldsymbol{e}^T\boldsymbol{\alpha}$$

$$\text{subject to} \quad \boldsymbol{y}^T\boldsymbol{\alpha} = 0,$$
$$0 \leq \alpha_i \leq C, \qquad i = 1, \ldots, l,$$

$$Q_{ij} \equiv y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

Transduction

Labeled data $(x_1, y_1) \ldots (x_l, y_l)$
Unlabeled data $x_{l+1} \ldots x_n$

$$\min_{(\mathbf{w},b),\ \mathbf{y}_u} I(\mathbf{w},b,\mathbf{y}_u) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l} V(y_i,o_i) + C^\star \sum_{i=l+1}^{n} V(y_i,o_i)$$

The above objective function is non-convex. How to "solve" it?

- S^3VM^{light} (1999)
- Convex relaxations (2004)
- CCCP (2003)
- \Delta S^3VM (2005)

…

S^3VM^{light}

- First train regular SVM on labeled data
- Use SVM to classify unlabeled points
- Use current labeling to retrain with low C^*
- Interleave regular SVM solution with unlabeled label swaps to make the objective function strictly decrease

$$y_i = 1, y_j = -1, V(1, o_i) + V(-1, o_j) > V(-1, o_i) + V(1, o_j)$$

- Slowly increase effect of unlabeled by C*

S^3VM^{light}

Algorithm 1 $\mathrm{S^3VM}^{light}$

Train an SVM with the labeled points. $o_i \leftarrow \mathbf{w} \cdot \mathbf{x}_i + b$.
Assign $y_i \leftarrow 1$ to the ur largest o_i, -1 to the others.
$\tilde{C} \leftarrow 10^{-5} C^\star$
while $\tilde{C} < C^\star$ **do**
  repeat
    Minimize (1) with $\{y_i\}$ fixed and $C^\star$ replaced by $\tilde{C}$.
    if $\exists(i,j)$ satisfying (6) **then**
      Swap the labels y_i and y_j
    end if
  until No labels have been swapped
  $\tilde{C} \leftarrow \min(1.5C, C^\star)$
end while

1. Annealing loop:
A smoothing heuristic for non-convex optimization

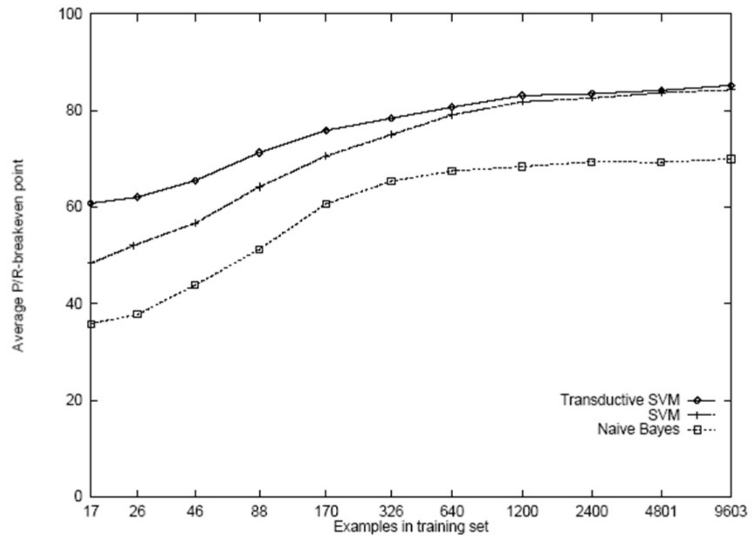
2. Convergence

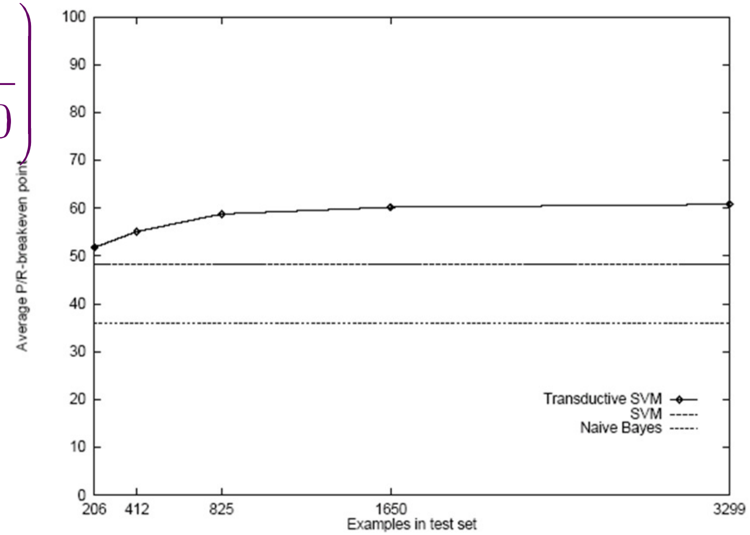
Transduction for Text

- 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

$$X_j(w_i) = (\# w_i \; in \; d_j) \times \log\left(\frac{\# d_j}{\# d_j \; where \; \# w_i > 0}\right)$$

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





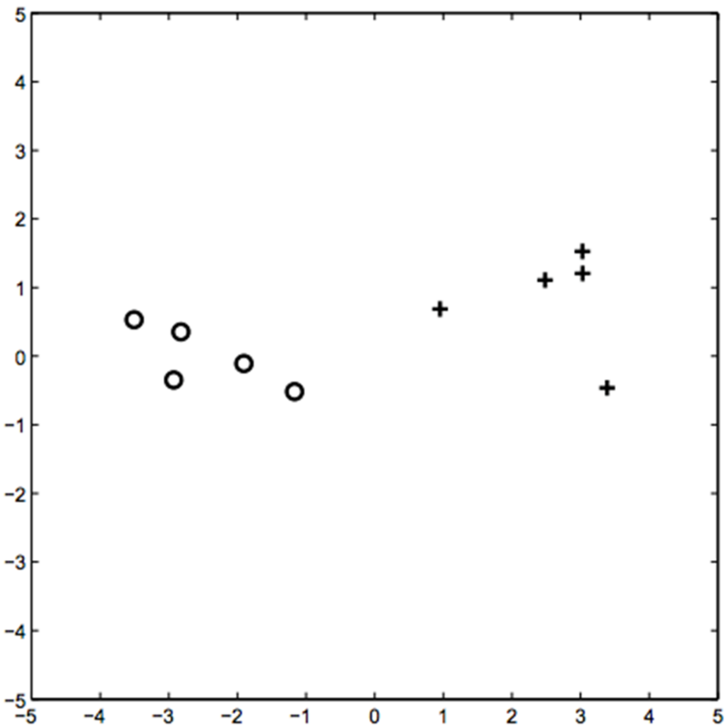
Problems

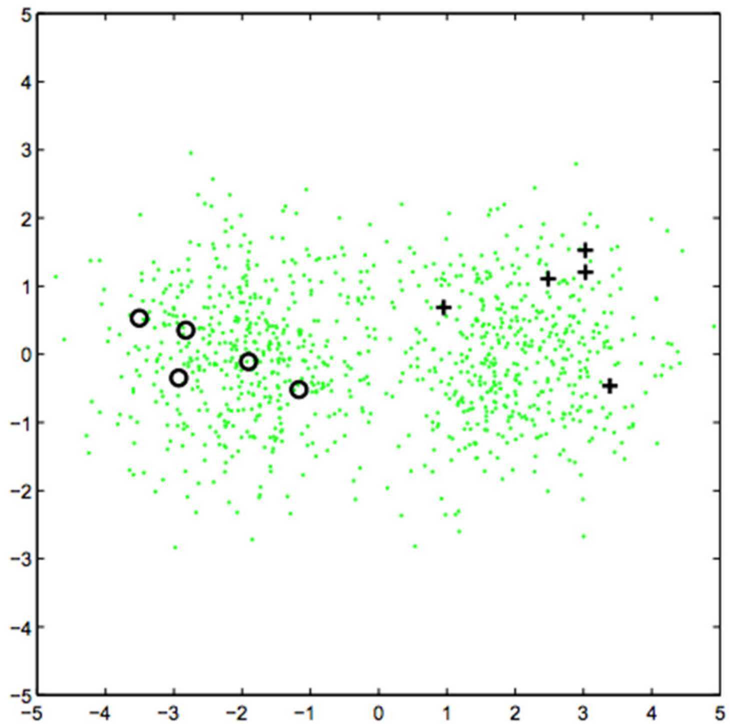
Difficulties in optimization
(S3VM light is very slow in convergence)
Unlabeled data can sometime hurt performance

References

Transductive inference for text classification using support vector machines, Joachims 1999
Optimization techniques for semi-supervised support vector machines, Chapelle, Olivier and Sindhwani, Vikas and Keerthi, Sathiya S, 2008
Maximum margin clustering, Xu et al. 2004

Generative Models (EM)





Partially Labeled Data & EM

- Instead of maxmizing likelihood of labeled data

$$l(\theta) = \sum_{i \in LAB} \log\left(p\left(x_i, y_i \mid \theta\right)\right)$$

- Or maximizing likelihood of unlabeled data (needs EM)

$$l(\theta) = \sum_{i \in UNLAB} \log\left(\sum_y p\left(x_i, y \mid \theta\right)\right)$$

- Maximize a combination of both weighted by λ

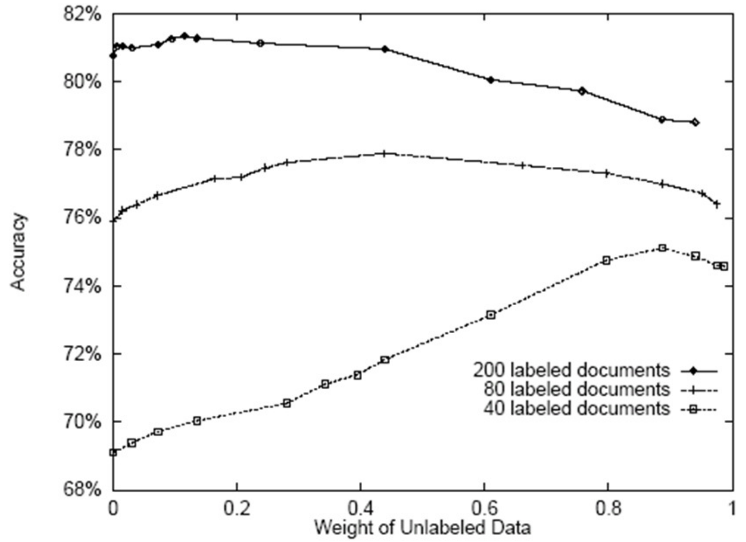
$$l(\theta) = \sum_{i \in LAB} \log\left(p\left(x_i, y_i \mid \theta\right)\right) + \lambda \sum_{i \in UNLAB} \log\left(\sum_y p\left(x_i, y \mid \theta\right)\right)$$

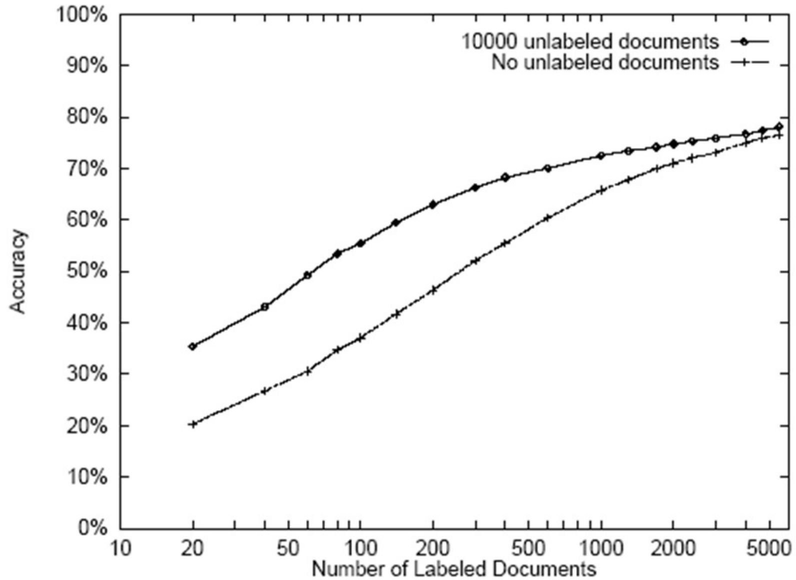
- Also, use a prior $P(\theta)$ to help (avoids zero-counts in multinomial models)…

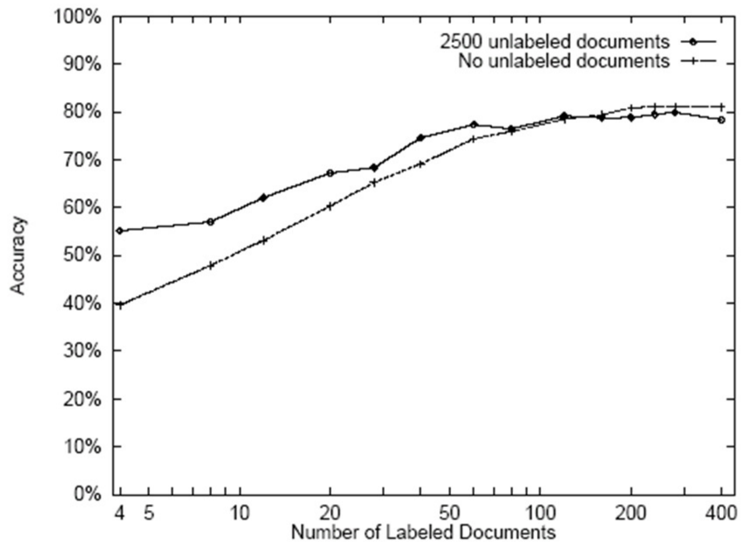
$$l(\theta) = \log p(\theta) + \sum_{i \in LAB} \log\left(p\left(x_i, y_i \mid \theta\right)\right)$$
$$+ \lambda \sum_{i \in UNLAB} \log\left(\sum_y p\left(x_i, y \mid \theta\right)\right)$$

Partially Labeled Data & EM

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







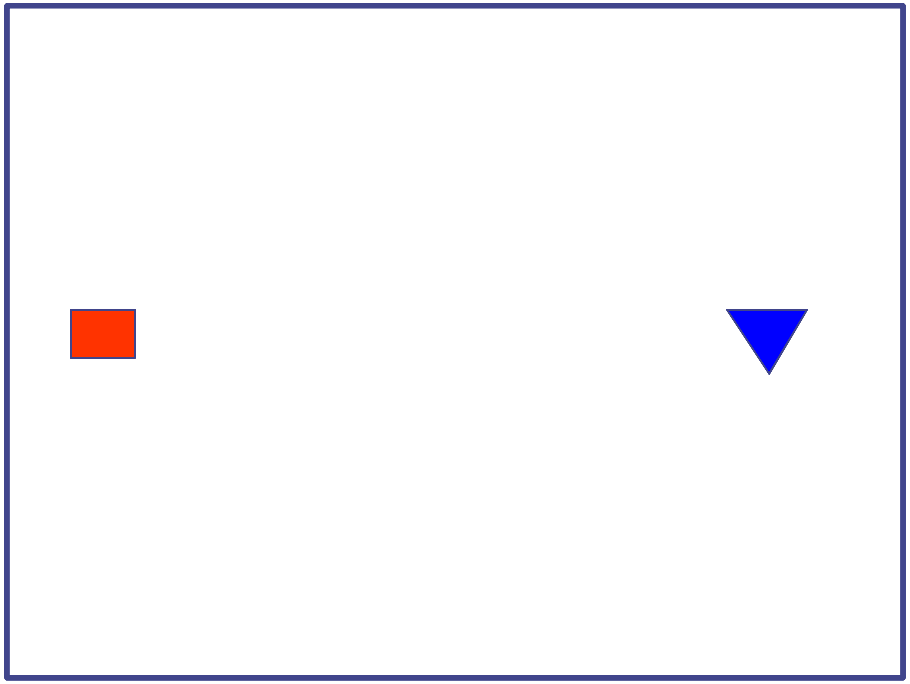
Problems

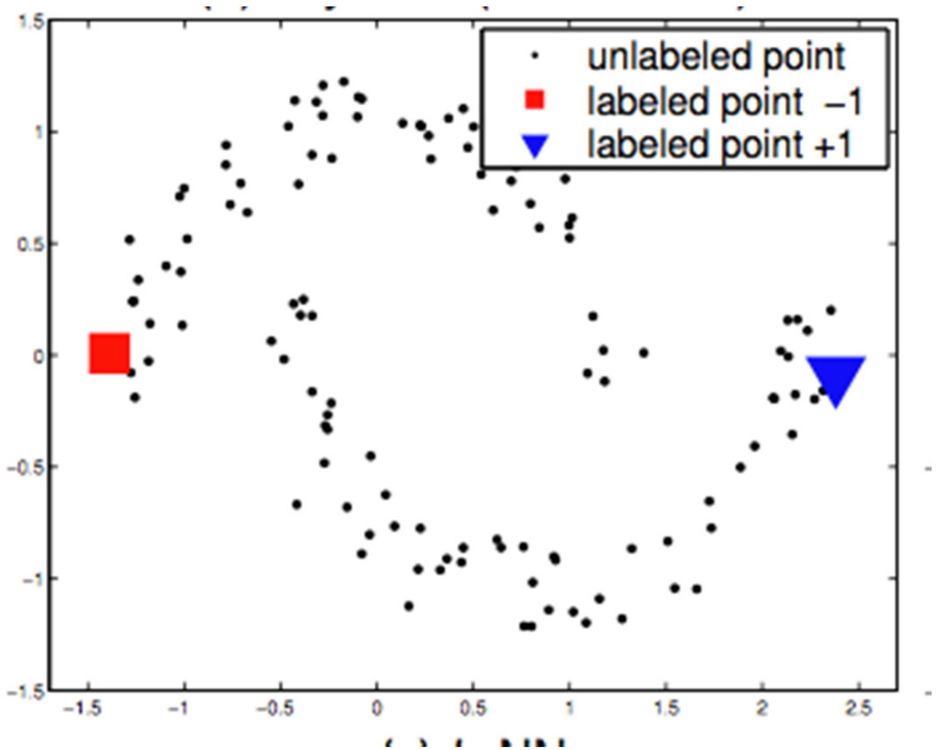
Difficulties in optimization (local minima of EM)
Unlabeled data can sometime hurt performance
How to identify the model?

References

Text Classification from Labeled and Unlabeled Documents using EM by K. Nigam, A. McCallum, S. Thrun and T. Mitchell
Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions by Zhu, Ghahramani and Lafferty

Graph based method





Basic Assumption

Similar instances should have the same label.

Local and global consistency

$$\min_f \sum_{i=1}^{l} (f(x_i) - y_i)^2 + \lambda f^\top \Delta f$$

$$\min_f \sum_{i=1}^{l} (f(x_i) - y_i)^2 + \lambda f^\top \Delta f$$

$$\sum_{i \sim j} w_{ij} (f(x_i) - f(x_j))^2 = f^\top \Delta f$$

$$D_{ii} = \textstyle\sum_{j=1}^{n} W_{ij}$$

$$\Delta = D - W$$

Problems

How to construct the graph?
(complexity and quality)

References

Local and Global Consistency by Zhou et al.
Graph Construction and b-Matching for Semi-Supervised Learning by Jebara, Wang and Chang
A tutorial on spectral clustering, Von Luxburg, Ulrike 2007

Summary

supervised	Semi-supervised	Unsupervised
SVM	S3VM	Large-margin clustering
Naïve Bayes?	EM-based SSL	Mixture of Gaussian
KNN?	Graph based SSL	Spectral methods