Advanced Machine Learning & Perception

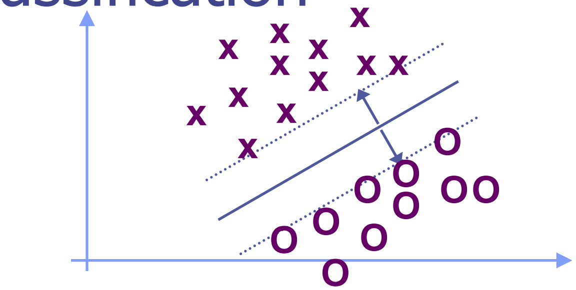
Instructor: Tony Jebara

Semi-Supervised Learning

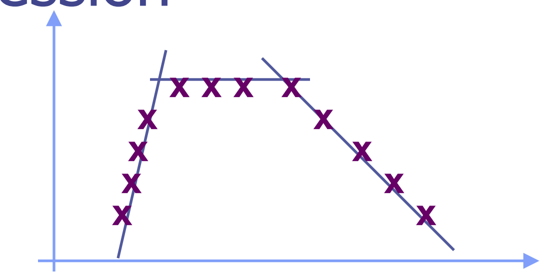
- Semi-Supervised Learning
- Exploiting Unlabeled Data
- Transduction
- Partially Labeled Data and EM

SVM Extensions

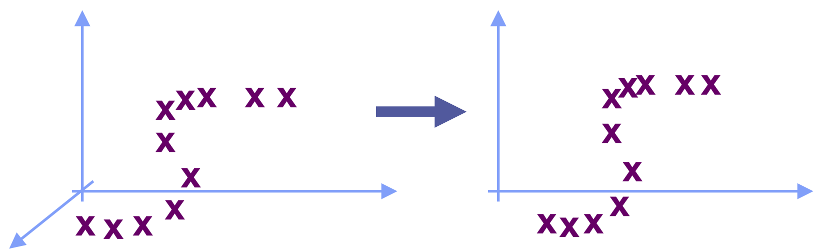
Classification



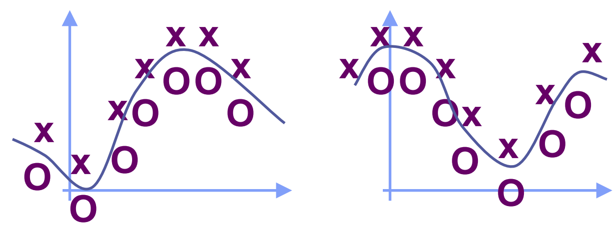
Regression



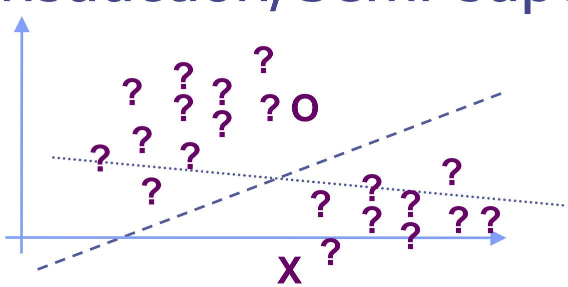
Feature/Kernel Selection



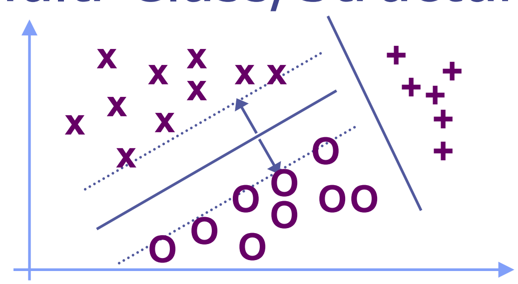
Meta/Multi-Task Learning



Transduction/Semi-supervised



Multi-Class/Structured



Semi-supervised Learning

- What

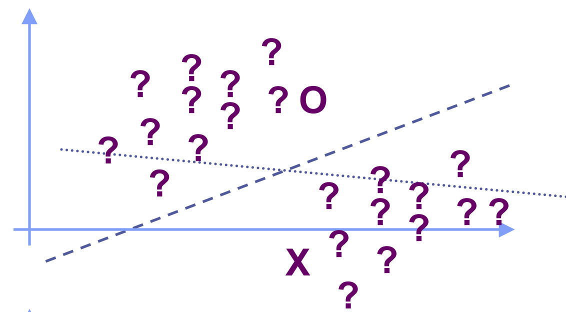
Learning setting	Learning from …
Supervised Learning	labeled data
Semi-supervised Learning	both labeled and unlabeled data
Unsupervised Learning	Unlabeled data

- 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 sometimes.

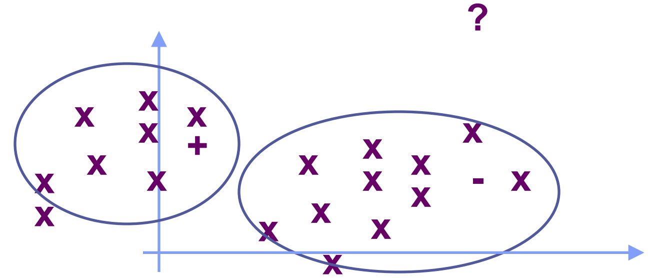
Semi-Supervised Learning

- Several approaches:

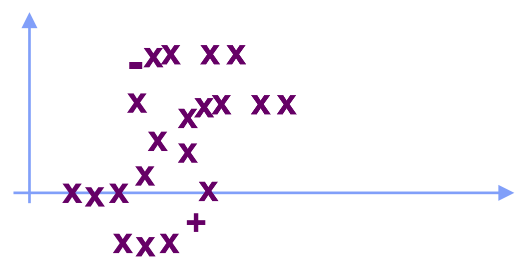
- Transduction: discriminative, find large margin region.



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

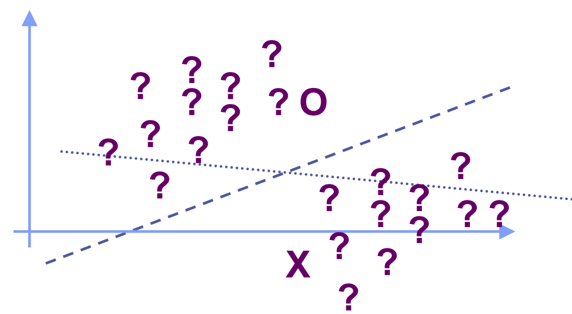


- Graphs & Diffusion: spreading labels across a graph or manifold via spectral, kernel, or Markov walks.



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 $(y_1^*, \dots, y_n^*, \vec{w}, b, \xi_1, \dots, \xi_n, \xi_1^*, \dots, \xi_k^*)$:

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

subject to:

$$\forall_{i=1}^{n} : y_i[\vec{w}\cdot\vec{x}_i + b] \geq 1 - \xi_i$$
$$\forall_{j=1}^{k} : y_j^*[\vec{w}\cdot\vec{x}_j^* + b] \geq 1 - \xi_j^*$$
$$\forall_{i=1}^{n} : \xi_i > 0$$
$$\forall_{j=1}^{k} : \xi_j^* > 0$$

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))

Minimize over $(\vec{w}, b, \vec{\xi}, \vec{\xi}^*)$:

$$\frac{1}{2}||\vec{w}||^2 + C\sum_{i=1}^{n}\xi_i + C^*_-\sum_{j:y^*_j=-1}\xi^*_j + C^*_+\sum_{j:y^*_j=1}\xi^*_j$$

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 swap if $\left(y^*_m y^*_l < 0\right)$ & $\left(\xi^*_m > 0\right)$ & $\left(\xi^*_l > 0\right)$ & $\left(\xi^*_m + \xi^*_l > 2\right)$
- Slowly increase effect of unlabeled by C* doubling ‘til max

Transduction with SVMs

Input:	– training examples $(\vec{x}_1, y_1), ..., (\vec{x}_n, y_n)$
	– test examples $\vec{x}^*_1, ..., \vec{x}^*_k$
Parameters:	– C, C^*: parameters from OP(2)
	– num_+: number of test examples to be assigned to class +
Output:	– predicted labels of the test examples $y^*_1, ..., y^*_k$

$(\vec{w}, b, \vec{\xi}, \_) := solve\_svm\_qp([(\vec{x}_1, y_1)...(\vec{x}_n, y_n)], [], C, 0, 0)$;

Classify the test examples using $< \vec{w}, b >$. The num_+ test examples with the highest value of $\vec{w} * \vec{x}^*_j + b$ are assigned to the class + $(y^*_j := 1)$; the remaining test examples are assigned to class − $(y^*_j := -1)$.

$C^*_- := 10^{-5}$; // some small number
$C^*_+ := 10^{-5} * \frac{num_+}{k - num_+}$;

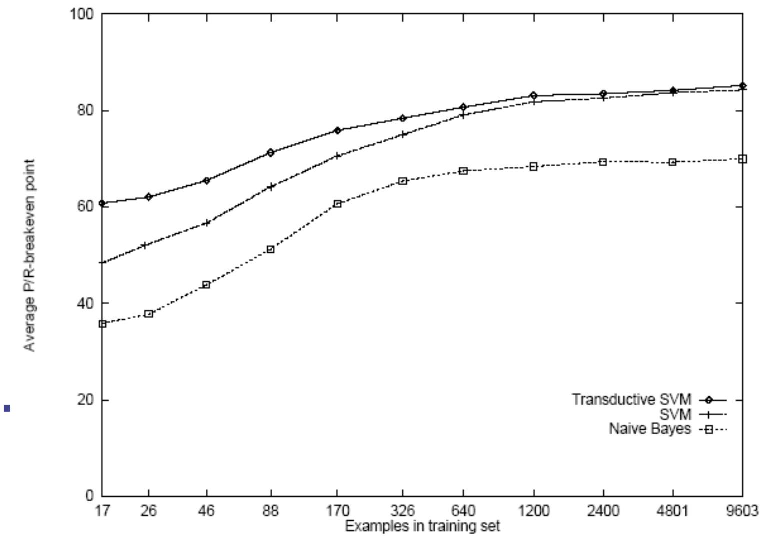
while$((C^*_- < C^*) \parallel (C^*_+ < C^*))\{$ // Loop 1
    $(\vec{w}, b, \vec{\xi}, \vec{\xi}^*) := solve\_svm\_qp([(\vec{x}_1, y_1)...(\vec{x}_n, y_n)], [(\vec{x}^*_1, y^*_1)...(\vec{x}^*_k, y^*_k)], C, C^*_-, C^*_+)$;
    while$(\exists m, l : (y^*_m * y^*_l < 0) \& (\xi^*_m > 0) \& (\xi^*_l > 0) \& (\xi^*_m + \xi^*_l > 2))$ $\{$ // Loop 2
        $y^*_m := -y^*_m$; // take a positive and a negative test
        $y^*_l := -y^*_l$; // example, switch their labels, and retrain
        $(\vec{w}, b, \vec{\xi}, \vec{\xi}^*) := solve\_svm\_qp([(\vec{x}_1, y_1)...(\vec{x}_n, y_n)], [(\vec{x}^*_1, y^*_1)...(\vec{x}^*_k, y^*_k)], C, C^*_-, C^*_+)$;
    }
    $C^*_- := min(C^*_- * 2, C^*)$;
    $C^*_+ := min(C^*_+ * 2, C^*)$;
}
return$(y^*_1, ..., y^*_k)$;

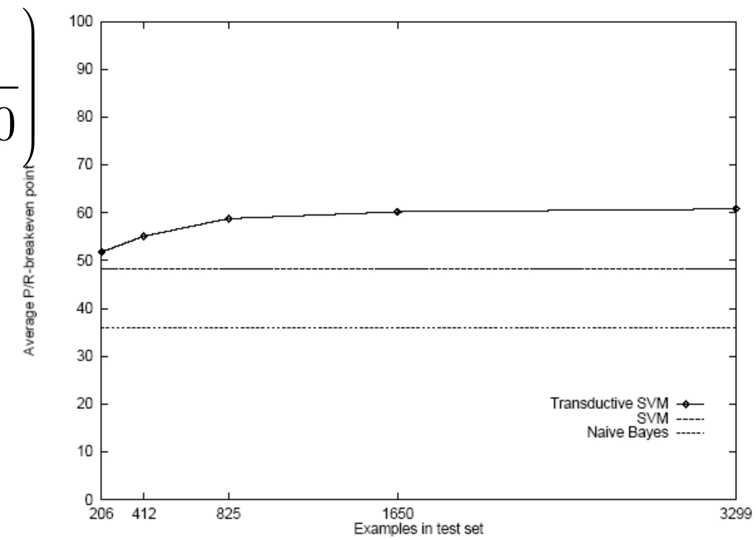
Transduction for Text

- 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

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

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





Generative Models and EM

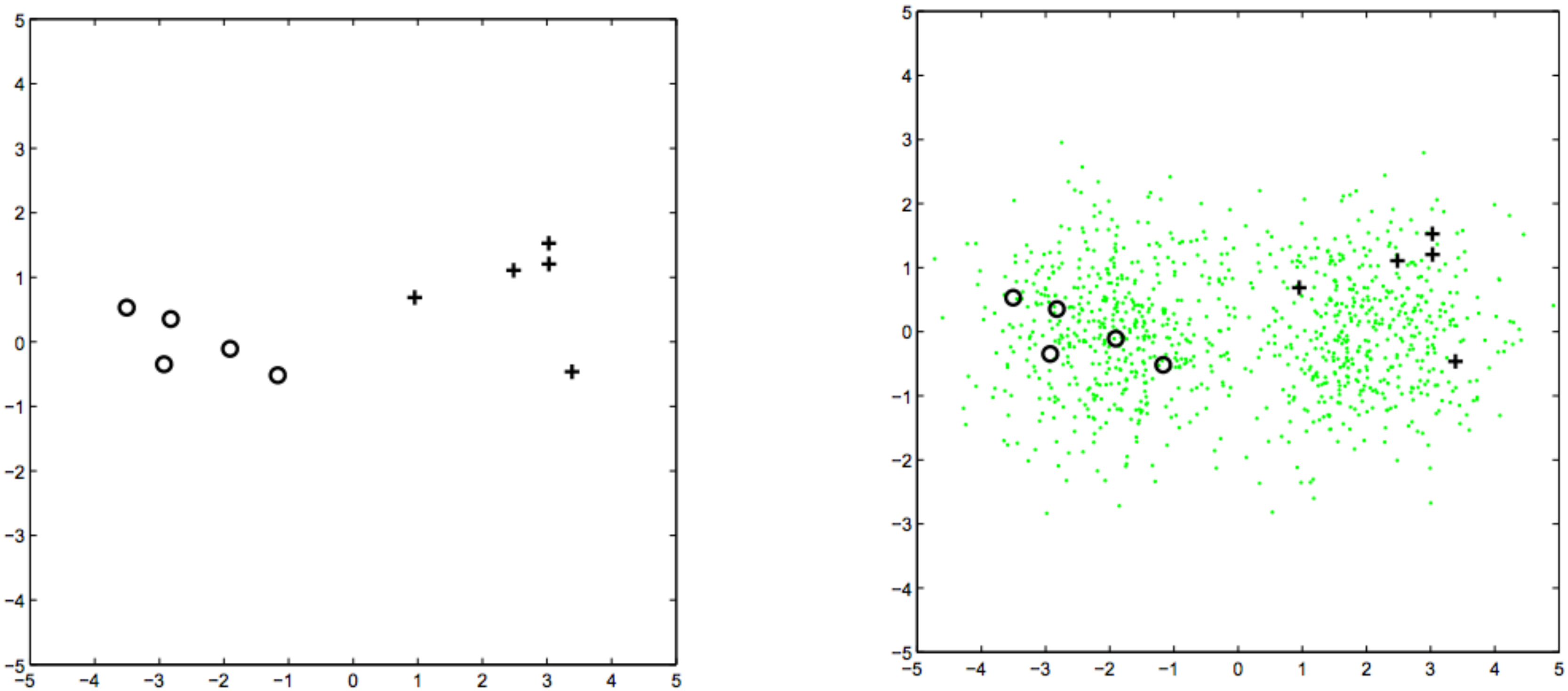


Figure credit: tutorial on semi-supervised learning Xiaojin Zhu

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