Semi-supervised learning II

CS534

Example: text classification

- Classify astronomy vs. travel articles
- Similarity measured by content word overlap

	d_1	d_3	d_4	d_2
asteroid	•	•		
bright	•	•		
comet		•		
year				
zodiac				
⋮				
airport				
bike				
camp			•	
yellowstone			•	•
zion				•

When labeled data alone fails

- No overlapping words

	d_1	d_3	d_4	d_2
asteroid	●			
bright	●			
comet				
year				
zodiac		●		
⋮				
airport			●	
bike			●	
camp				
yellowstone				●
zion				●

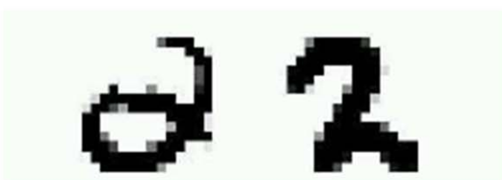
Unlabeled data as stepping stones

- Labels “propagate” via similar unlabeled articles.

	d_1	d_5	d_6	d_7	d_3	d_4	d_8	d_9	d_2
asteroid	●								
bright	●	●							
comet		●	●						
year			●	●					
zodiac				●	●				
⋮									
airport						●			
bike						●	●		
camp							●	●	
yellowstone								●	●
zion									●

Another example

- Handwritten digits recognition with pixel-wise Euclidean distance

	
not similar	‘indirectly’ similar with stepping stones

Graph-based semi-supervised learning

- Nodes: $X_l \cup X_u$
- Edges: similarity weights computed from features, e.g.,
  - K-nearest-neighbor graph, unweighted (0, 1 weights)
  - Fully connected graph, weighted ($w = \frac{\exp\left(-\left|x_i - x_j\right|^2\right)}{\sigma^2}$)
  - ϵ-radius graph
- Assumption: instances that are connected by heavy edges tend to have the same label

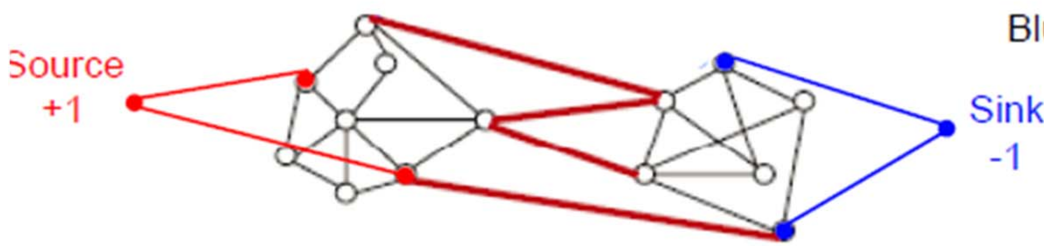
The mincut algorithm

- Fix Y_l, find $Y_u \in \{0,1\}^{n-l}$ to minimize $\sum_{ij} w_{ij} |y_i - y_j|^2$
- Equivalently, solve the following optimization problem:

$$\min_{Y \in \{0,1\}^n} \infty \underbrace{\sum_{i=1}^{l} (y_i - Y_{l i})^2}_{\substack{\text{Loss on labeled data} \\ \text{(mean square,0-1)}}} + \underbrace{\sum_{ij} w_{ij} (y_i - y_j)^2}_{\substack{\text{Graph based smoothness prior} \\ \text{on labeled and unlabeled data}}}$$

Hard constraints

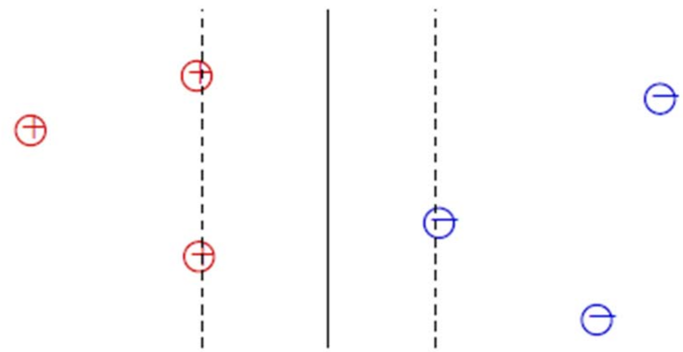
- If binary label, can be solved by min-cut on a modified graph – adding source and sink nodes with large weights to labeled examples



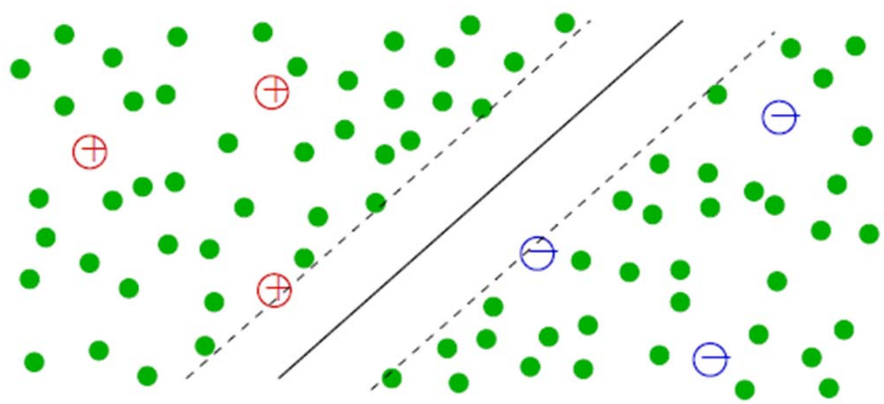
Blum & Chawla'01

Semi-supervised SVM (S^3VM)

- SVMs



- S^3VMs(Transductive SVMs)



Assumption: Unlabeled data from different classes are separated with large margin.

Standard soft margin SVMs

- keep labeled points outside the margin, while maximizing the margin:

Loss on training examples

$$\min_{h,b,\xi} \sum_{i=1}^{l} \xi_i + \overbrace{\lambda \|h\|^2_{\mathcal{H}_K}}^{\text{Magnitude of the weight in the kernel space}}$$

$$\text{subject to } \; y_i(h(x_i)+b) \geq 1 - \xi_i \quad , \forall i = 1 \ldots l$$

$$\xi_i \geq 0$$

- Equivalent to

$$\min_{f} \sum_{i=1}^{l} (1 - y_i f(x_i))_+ + \lambda \|h\|^2_{\mathcal{H}_K}$$

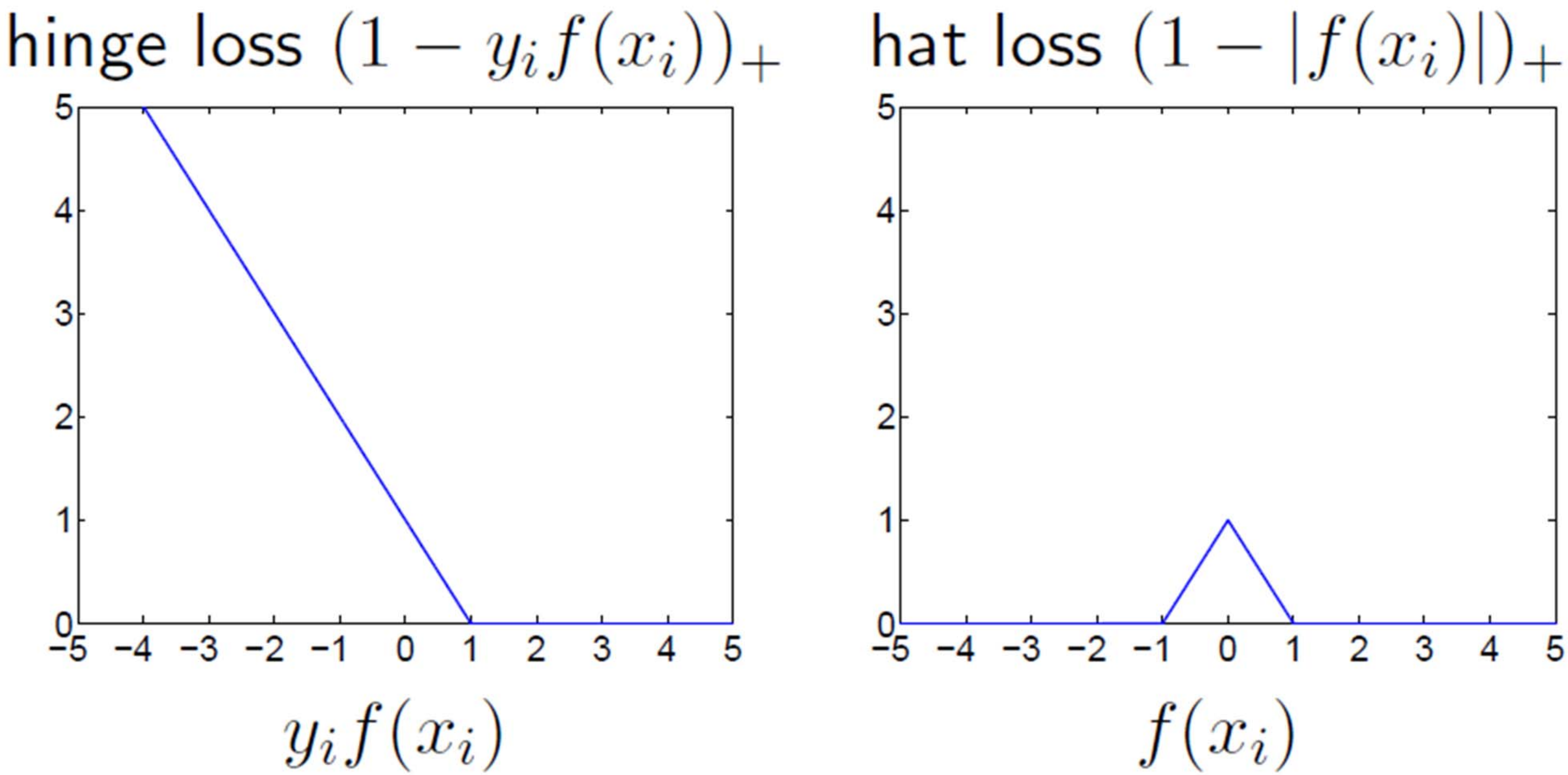
$y_i f(x_i)$ known as the margin, $\left(1 \; - \; y_i f(x_i)\right)_+$ the hinge loss

S^3VM

- To incorporate unlabeled points,
  - assign putative labels $sign(f(x))$ to $x \in X_u$
  - Hinge loss on unlabeled points becomes $(1 - |f(x)|)_+$
- New objective:

$$\min_f \sum_{i=1}^{l} (1 - y_i f(x_i))_+ + \lambda_1 \|h\|_{\mathcal{H}_K}^2 + \lambda_2 \sum_{i=l+1}^{n} (1 - |f(x_i)|)_+$$

The hat loss on unlabeled data



Prefers $f(x) \geq 1$ or $f(x) \leq -1$, i.e., unlabeled instance away from decision boundary $f(x) = 0$.

Class balance regularization

- often unbalanced – most points classified into one class
- Heuristic for encouraging class balance

$$\frac{1}{n-l} \sum_{i=l+1}^{n} f(x_i) = \frac{1}{l} \sum_{i=1}^{l} y_i.$$

Putting everything together

$$\min_f \sum_{i=1}^{l} \big(1 - y_i f(x_i)\big)_+ + \lambda_1 \|f\|_{\mathcal{H}_k}^2 + \lambda_2 \sum_{i=l+1}^{n} (1 - |f(x_i)|)_+$$

s.t. $$\frac{1}{n-l} \sum_{i=l+1}^{n} f(x_i) = \frac{1}{l} \sum_{i=1}^{l} y_i$$

- Computational difficulty
  - SVM objective is convex
  - S^3VM objective is non-convex

Summary: Semi-Supervised Learning

- Generative methods – Mixture models
- Multi-view methods – Co-training
- Graph-based methods
- Semi-Supervised SVMs – assume unlabeled data from different classes have large margin
- Many others …

SSL algorithms can use unlabeled data to help improve prediction accuracy if data satisfies appropriate assumptions