In The Name of Allah

Digital Media Laboratory
Sharif University of Technology

Statistical Pattern Recognition

Semi-Supervised Learning

Hamid R. Rabiee
Jafar Muhammadi, Alireza Ghasemi

Spring 2012
http://ce.sharif.edu/courses/90-91/2/ce725-1/

Agenda

- **Introduction to Semi-Supervised Learning (SSL)**
- **Classifier based methods**
  - **Expectation-Maximization (EM)**
  - **Self-Training**
  - **Co-Training**
  - **Semi-Supervised SVM (S3VM)**
- **Data based methods**
  - **Graph-based Methods**
  - **Manifold Regularization**

Learning Problems

- **Supervised learning:**
  - Given a sample consisting of object-label pairs (x_i,y_i), find the predictive relationship between objects and labels.
- **Unsupervised learning:**
  - Given a sample consisting of only objects, look for interesting structures in the data, and group similar objects.
- **What is Semi-supervised learning?**
  - Supervised learning + Additional unlabeled data
  - Unsupervised learning + Additional labeled data

Why Semi-Supervised Learning?

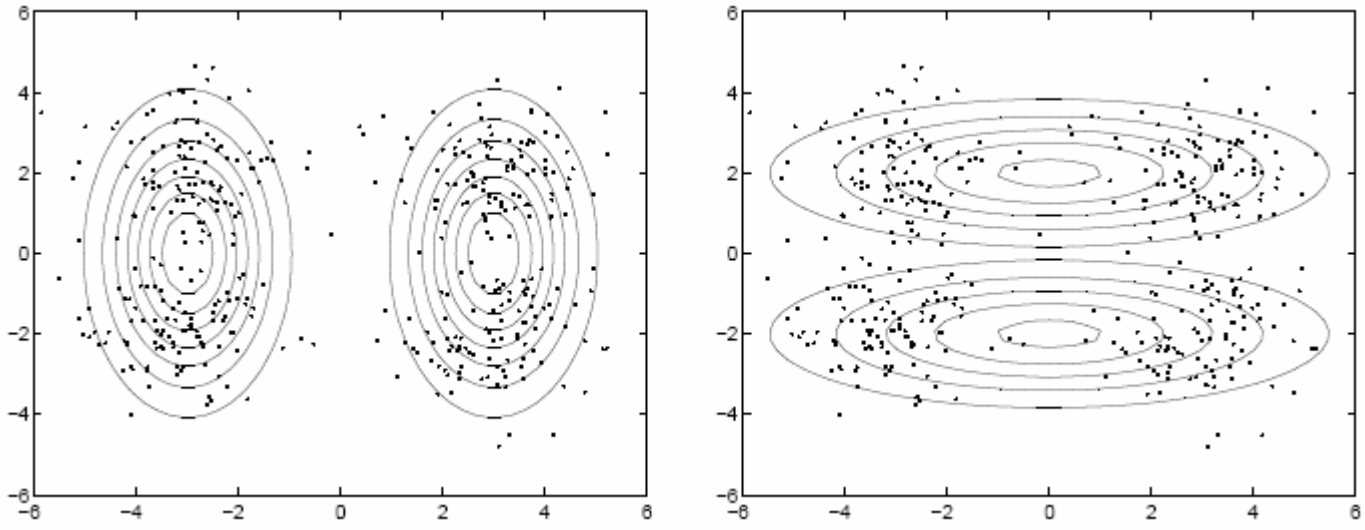
- **Data labeling is expensive and difficult**
- **Labeling is often unreliable**
- **Unlabeled examples**
  - **Easy to obtain in large numbers**
  - **e.g. webpage classification, bioinformatics, image classification**
- **Is unlabeled data useful?**
  - **In general yes, but not always**

Why Semi-Supervised Learning?

- **Unlabeled data is used to approximate the data distribution, P(x)**
  - **P(x) can incorporate in labeling function f(x), P(x) → f(x), with some assumption.**
  - **The performance will improve when the assumption is hold in given dataset!**
  - **Common assumption:**
    - Manifold Assumption: data lay on low dimensional manifold
    - Cluster Assumption: data in same cluster probably have the same label

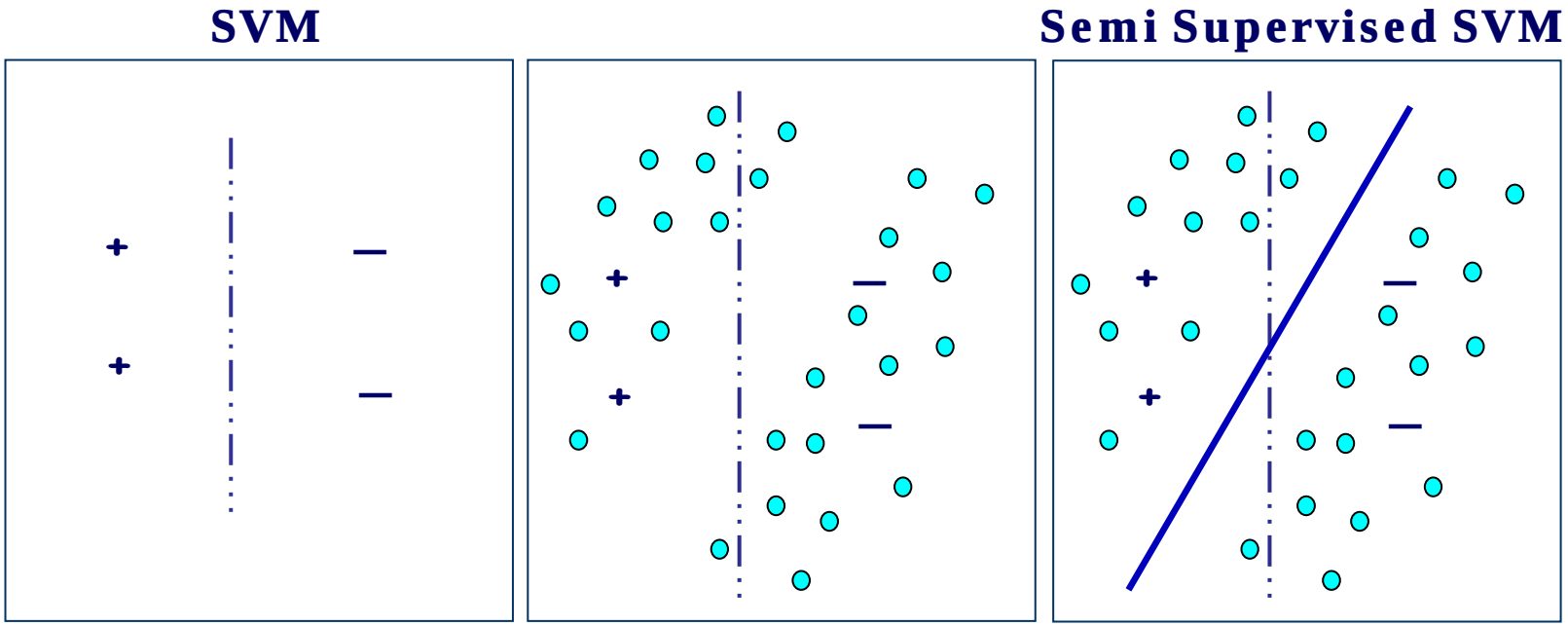
Why Semi-Supervised Learning?

- **Unlabeled Data May Hurt Learning**
  - **Right: True data distribution**
  - **Left: What is computed by maximum likelihood**



Why Semi-Supervised Learning?

- **Intuition**



Induction and Transduction

- **Transductive: Produce label only for the available unlabeled data.**
  - **The output of the method is not a classifier.**
- **Inductive: Not only produce label for unlabeled data, but also produce a classifier.**
  - **We focus on inductive semi-supervised learning..**

Glossary

- **Supervised classification**
  - $\{(x_{1:n}, y_{1:n})\}$
- **Semi-supervised classification**
  - $\{(x_{1:l}, y_{1:l}), x_{l+1:n}\}$
- **Semi-supervised clustering**
  - $\{x_{1:n}$, must-link, can't-link$\}$
- **Unsupervised learning (clustering)**
  - $\{x_{1:n}\}$

Two algorithmic approaches

- **Classifier based methods:**
  - Start from initial classifier(s), and iteratively enhance it (them)
  - EM, co-training and S3VM are classifier based methods.
- **Data based methods:**
  - Discover an inherent geometry in the data, and exploit it in finding a good classifier.
  - Graph based and manifold learning are data based methods.

EM

- **EM is introduced in lecture 14.**
- **Information from unlabeled data can be easily inserted into the objective function of EM**
- **EM can be used to maximize the joint log-likelihood of labeled and unlabeled data:**

$$\underbrace{\sum_i \log\left(P(x_i \mid y_i,\theta)P(y_i)\right)}_{\log-\text{likelihood of labeled data}} + \underbrace{\sum_j \log\left(\sum_y P(x_j \mid y,\theta)P(y)\right)}_{\log-\text{likelihood of unlabeled data}}$$

- **EM Assumption: The data actually comes from the mixture model, where the number of components, prior p(y), and conditional p(x | y) are all correct.**

Self-Training

- **A simple semi-supervised algorithm**
  - **A committee of classifiers are trained on the labeled examples**
  - **Then classify the unlabeled examples independently**
  - **Those examples, to which most of the classifiers give the same label, are added to the training set**
  - **procedure repeats until a stop condition is met**
- **A single classifier can carry out its own self-training procedure (How?)**
- **Self-training assumption: The model's own predictions, at least the high confidence ones, tend to be correct.**

Co-Training

- **A well-known semi-supervised algorithm**
- **To function efficiently, assumes facts about data**
  - **Instances contain two sufficient sets of features**
    - i.e. an instance is $x=(x_1,x_2)$
    - Each set of features is called a View
  - **Two views of instances are independent given the label:**

$$P(x_1 \mid x_2, y) = P(x_1 \mid y)$$
$$P(x_2 \mid x_1, y) = P(x_2 \mid y)$$

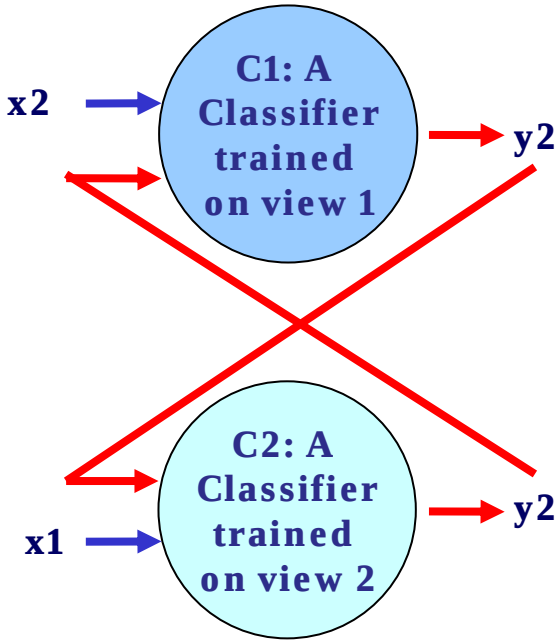
  - **Two views of instances are consistent:**

$$\exists classifiers\,\{c_1, c_2\}\,\forall x = (x_1, x_2):\ \ c_1(x_1) = c_2(x_2)$$

  - **Co-Training Assumption: conditionally independent and redundant features splits**

Co-Training

- **Co-training iteratively constructs new labeled samples from the unlabeled ones, using the following steps:**
  1. **Learns two separate classifiers for the two views using the labeled samples.**
  2. **Estimates the labels of unlabeled samples, using both classifiers.**
  3. **Uses the most confident estimated labels of each classifier (e.g. y2), with their correspondent samples (e.g. x1 and x2), as additional labeled training data.**



Co-Training

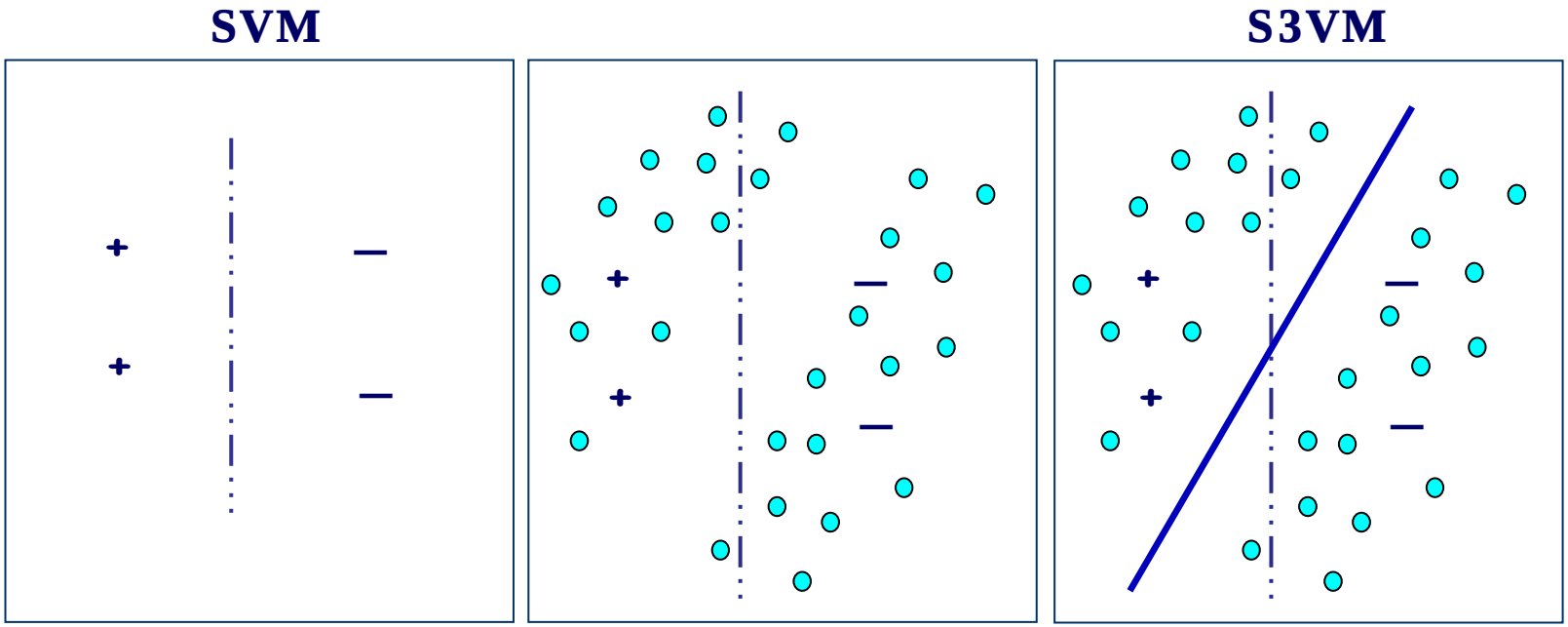
- **Advantages and Disadvantages**
  - **Advantages**
    - Simple wrapper method.
    - Applies to almost all existing classifiers
  - **Disadvantages**
    - Natural feature splits may not exist
    - Models using BOTH feature sets should do better

S3VM (Semi-Supervised SVM)

- **Suppose we believe target separator goes through low density regions of the space / large margin.**
- **Aim for separator with large margin w.r.t. labeled and unlabeled data.**
- **Unfortunately, optimization problem is now NP-hard. Algorithm instead does local optimization.**
  - **Start with large margin over labeled data. Induces labels on unlabeled data.**
  - **Then try flipping labels in greedy fashion.**
- **S3VM Assumption: target separator goes through low density region between classes**
- **Quite successful on text data.**

S3VM (Semi-Supervised SVM)

✧ **SVM vs. S3VM**

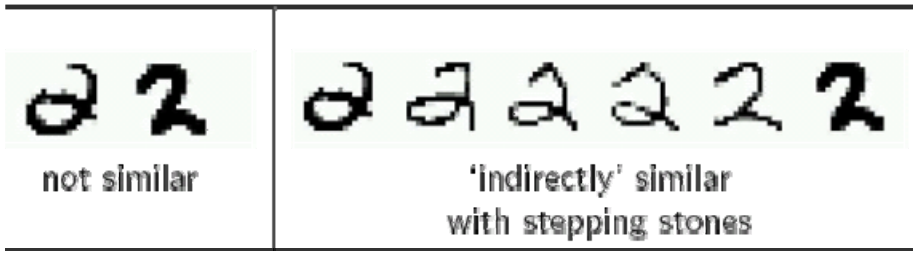


S3VM (Semi-Supervised SVM)

- **It can be viewed either as a discrete or continuous optimization problem. (How?)**
- **Many optimization approaches based on these criteria have been proposed**
  - **For discrete optimization:**
    - Branch and Bound, Simulated Annealing, Evolutionary Optimization, …
  - **For continuous optimization:**
    - Newton Method, Gradient descent, …

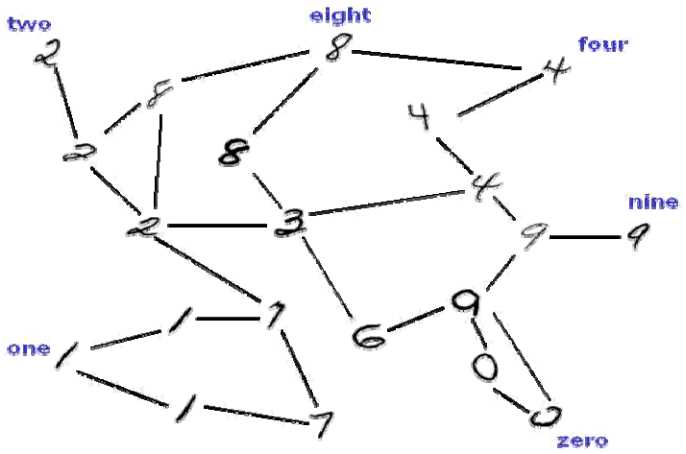
Graph-Based Methods

- **Suppose we believe that very similar examples probably have the same label.**
- **If you have a lot of labeled data, this suggests a Nearest-Neighbor type of algorithm.**
- **If you have a lot of unlabeled data, perhaps can use them as "stepping stones"**

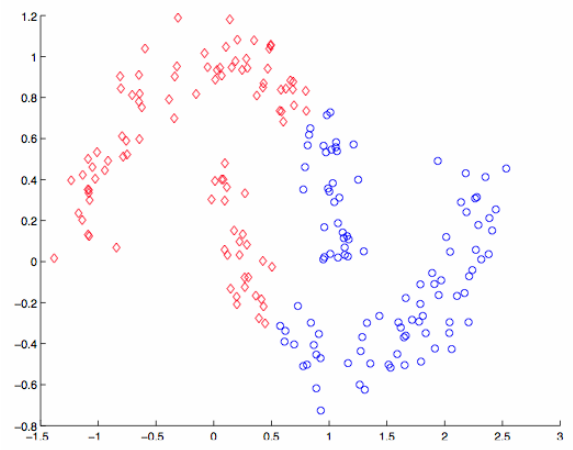


Graph-Based Methods

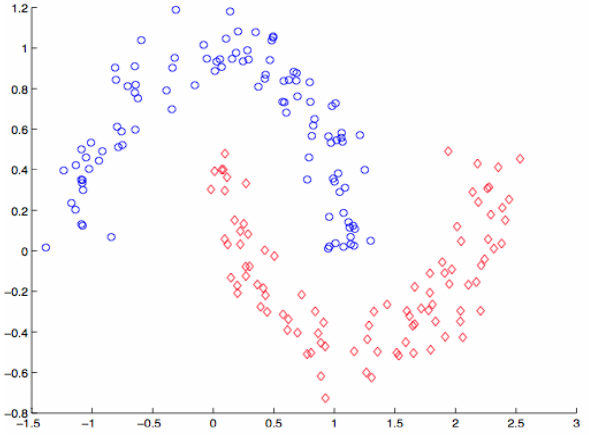
- **Idea:**
  - **Construct a graph with edges between very similar examples.**
- **Unlabeled data can help "glue" the objects of the same class together.**
- **The graph is a discrete approximation of the underlying manifold in which data lie**



Data Manifold



Supervised Learning



Manifold Learning

- **Knowing the geometry affects the estimated labels. Geometry changes the notion of similarity.**
- **Assumption: Data is distributed on some low dimensional manifold.**
- **Unlabeled data is used to estimate the geometry.**

Smoothness Assumption

- **Desired functions are smooth with respect to the underlying geometry.**

  Functions of interest do not vary much in high density regions or clusters.

  - **Example: The constant function is very smooth, however it has to respect the labeled data.**

- **The probabilistic version:**

  - **Conditional distributions P(y|x) should be smooth with respect to the marginal P(x).**
  - **Example: In a two class problem P(+|x) and P(-|x) do not vary much in clusters.**

- **Cluster Assumption**

  - **Put the decision boundary in low density area.**
    - A consequence of the smoothness assumption.

Smoothness Assumption

- **What is smoothness?**
  - **Let f : M→R. Penalty at** $x \in M$**:**

$$\frac{1}{\delta^k}\int_{\Delta}(f(x)-f(x+\delta))^2 P(x)d\delta \approx ||\nabla f||^2\ p(x)$$

  - **Total penalty:**

$$\int_M ||\nabla f||^2 P(x)dx$$

  - **P(x) is unknown, so the above quantity is estimated by the help of unlabeled data:**

$$||f||_I^2 = \sum_{i,j}(f(x_i)-f(x_j))^2 W_{ij}$$

Manifold Regularization

- **Manifold regularization method looks for the following classifier:**

$$f^{opt} = \arg\min_f \lambda_I ||f||_I^2 + \lambda_H ||f||_H^2 + \frac{1}{l}\sum_i^l (f(x_i) - y_i)^2$$

  - **First term: Smoothness of unlabeled data**
  - **Second term: Function complexity**
    - Manifold regularization method prefers low-complexity classifiers.
    - Calculating this term is complicated, and we ignore it, here.
  - **Third term: Fitness to unlabeled data**

Summary

- **We reviewed some important recent works on SSL.**
- **Different learning methods for SSL are based on different assumptions.**
- **Fulfilling these assumptions is crucial for the success of the methods.**

Any Question?

End of Lecture 15

Thank you!

Spring 2012

http://ce.sharif.edu/courses/90-91/2/ce725-1/