Semi-supervised Learning

Introduction

Labelled data

cat

dog

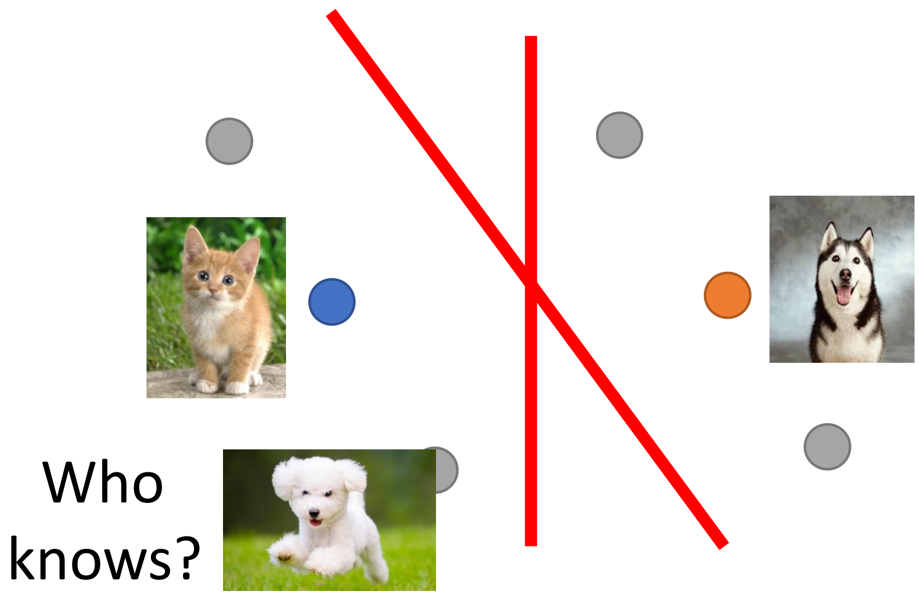
Unlabeled data

(Image of cats and dogs without labeling)

Introduction

- Supervised learning: $\{(x^r, \hat{y}^r)\}_{r=1}^{R}$
  - E.g. x^r: image, $\hat{y}^r$: class labels
- Semi-supervised learning: $\{(x^r, \hat{y}^r)\}_{r=1}^{R}$, $\{x^u\}_{u=R}^{R+U}$
  - A set of unlabeled data, usually U >> R
  - Transductive learning: unlabeled data is the testing data
  - Inductive learning: unlabeled data is not the testing data
- Why semi-supervised learning?
  - Collecting data is easy, but collecting "labelled" data is expensive
  - We do semi-supervised learning in our lives

Why semi-supervised learning helps?



The distribution of the unlabeled data tell us ***something***.

Usually with some assumptions

Outline

Semi-supervised Learning for Generative Model

Low-density Separation Assumption

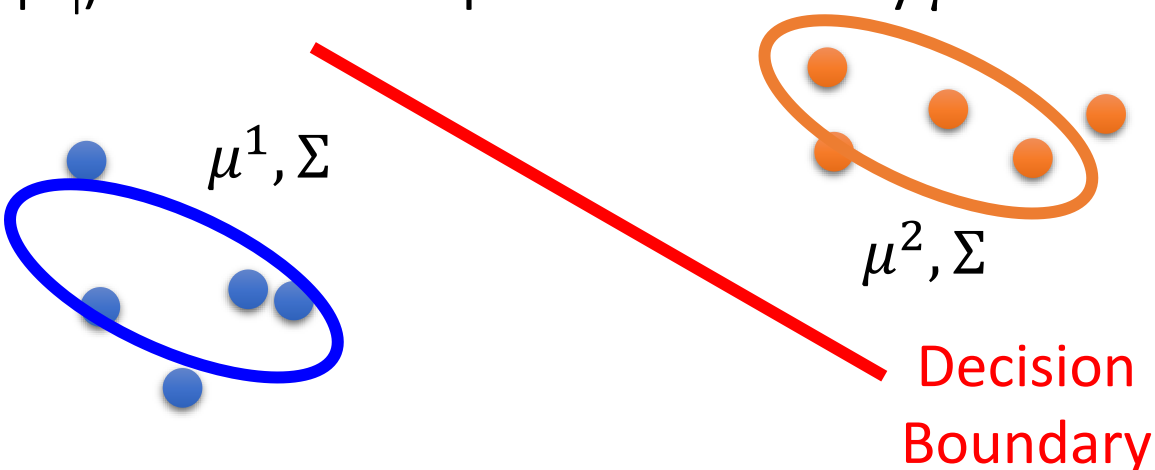
Smoothness Assumption

Better Representation

Semi-supervised Learning for Generative Model

Supervised Generative Model

- Given labelled training examples $x^r \in C_1, C_2$
  - looking for most likely prior probability $P(C_i)$ and class-dependent probability $P(x|C_i)$
  - $P(x|C_i)$ is a Gaussian parameterized by μ^i and Σ

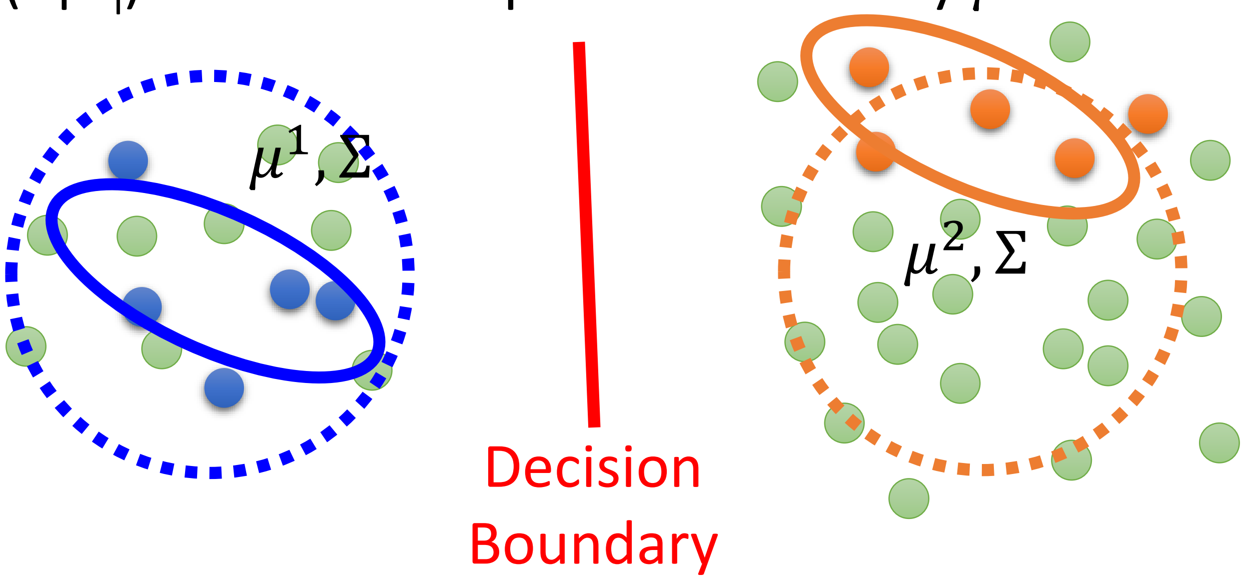


With $P(C_1)$, $P(C_2)$, μ^1, μ^2, Σ

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

Semi-supervised Generative Model

- Given labelled training examples $x^r \in C_1, C_2$
  - looking for most likely prior probability P(C_i) and class-dependent probability P(x|C_i)
  - P(x|C_i) is a Gaussian parameterized by μ^i and Σ



The unlabeled data x^u help re-estimate $P(C_1)$, $P(C_2)$, μ^1,μ^2, Σ

Semi-supervised Generative Model

The algorithm converges eventually, but the initialization influences the results.

- Initialization: $\theta = \{P(C_1), P(C_2), \mu^1, \mu^2, \Sigma\}$

E

- Step 1: compute the posterior probability of unlabeled data

$P_\theta(C_1|x^u)$ Depending on model θ

Back to step 1

M

- Step 2: update model

$$P(C_1) = \frac{N_1 + \sum_{x^u} P(C_1|x^u)}{N}$$

N: total number of examples

N_1: number of examples belonging to C_1

$$\mu^1 = \frac{\sum_{x^r \in C_1} x^r + \sum_{x^u} P(C_1|x^u) x^u}{N_1 + \sum_{x^u} P(C_1|x^u)}$$

……

EM Algorithm For GMM – Summary

- Randomly initialize parameters $\theta^{(1)}$.
- Iterate through step t=1,2,…
  - ➢Expectation Step (E-step): Compute

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{N}\sum_{k=1}^{K}\delta_{ik}^{(t)}\left\{\log\left(\frac{\pi_k}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}_k|}}\right)-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right\}$$

$$\delta_{ik}^{(t)} = \mathbb{P}\left[z_i = k\middle|\boldsymbol{x}_i;\theta^{(t)}\right]$$

Labeled data → Hard membership given by labels.

Unlabeled data → Soft membership dependent on $\theta^{(t)}$.

  - ➢Maximization Step (M-step): Choose $\theta^{(t)} = \left\{\left(\pi_k^{(t)},\boldsymbol{\mu}_k^{(t)},\boldsymbol{\Sigma}_k^{(t)}\right)\right\}_{k=1}^{K}$, where

$$\pi_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N}\delta_{ik}^{(t)}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{N}\delta_{ik}^{(t)}\boldsymbol{x}_i}{\sum_{i=1}^{N}\delta_{ik}^{(t)}}$$

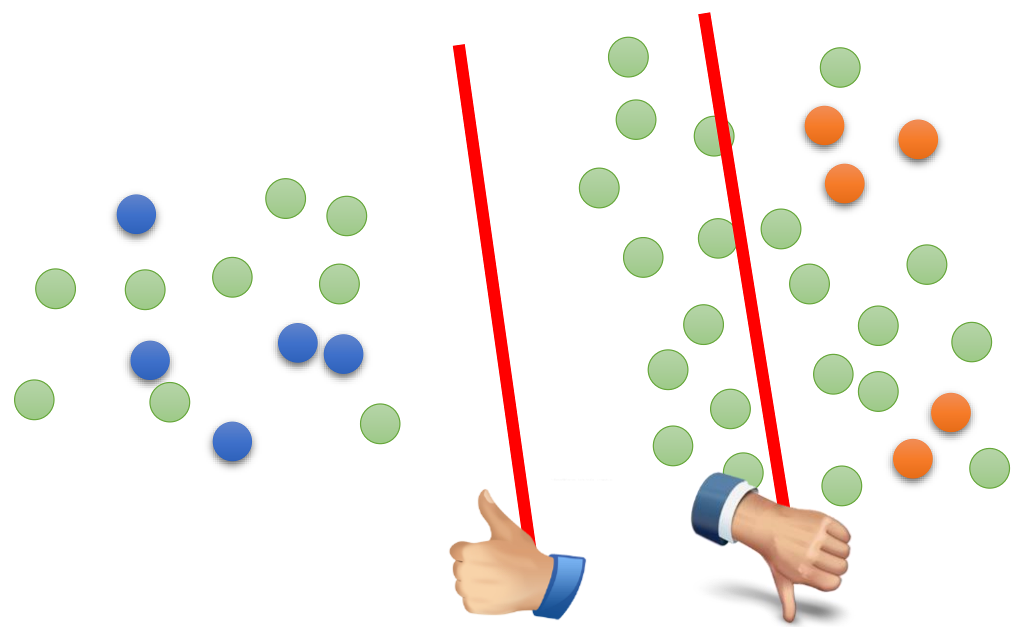
Re-estimate parameters using the currently estimated membership

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^{N}\delta_{ik}^{(t)}(\boldsymbol{x}_i-\boldsymbol{\mu}_k^{(t+1)})(\boldsymbol{x}_i-\boldsymbol{\mu}_k^{(t+1)})^T}{\sum_{i=1}^{N}\delta_{ik}^{(t)}}$$

Semi-supervised Learning
Low-density Separation

非黑即白

“Black-or-white”



Self-training

- Given: labelled data set = $\{(x^r, \hat{y}^r)\}_{r=1}^{R}$, unlabeled data set = $\{x^u\}_{u=1}^{U}$
- Repeat:
  - Train model f^* from labelled data set

    You can use any model here.

    Regression?
  - Apply f^* to the unlabeled data set
    - Obtain $\{(x^u, y^u)\}_{u=1}^{U}$

      Pseudo-label
  - Remove a set of data from unlabeled data set, and add them into the labeled data set

    How to choose the data set remains open

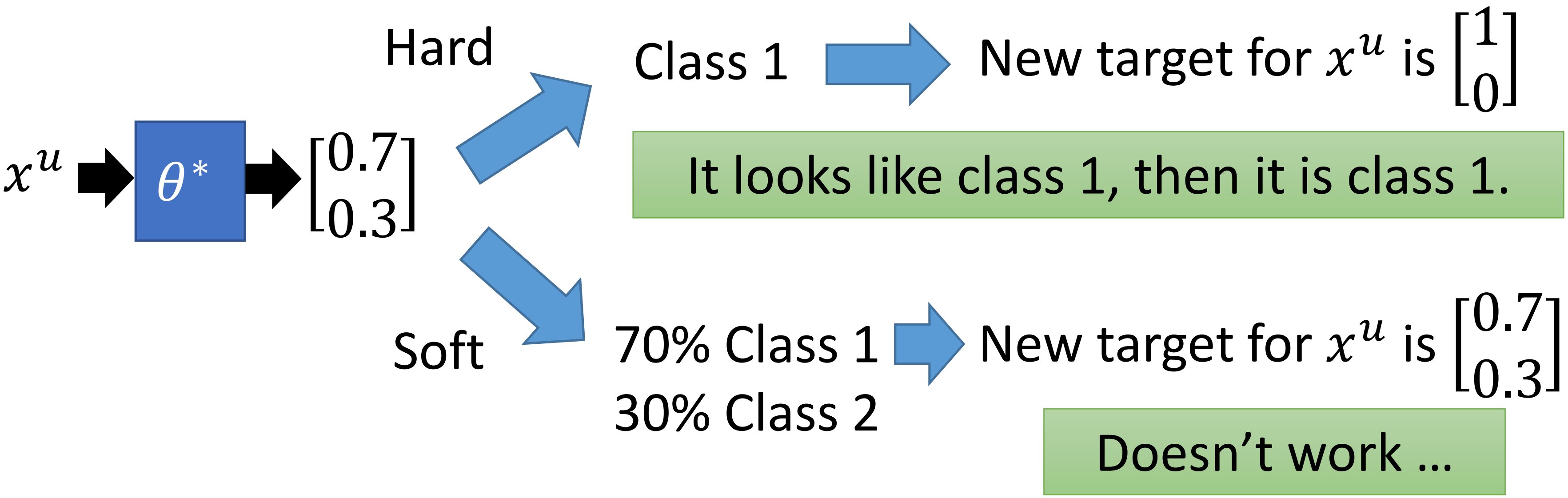
    You can also provide a weight to each data.

Self-training

- Similar to semi-supervised learning for generative model
- Hard label v.s. Soft label

  Considering using neural network

  θ^* (network parameter) from labelled data

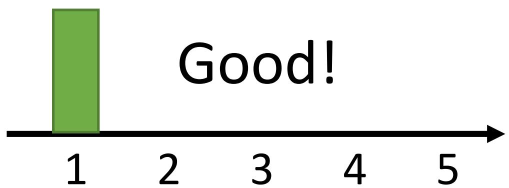


Entropy-based Regularization

Distribution

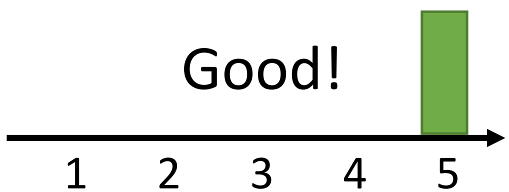
Entropy of y^u :
Evaluate how concentrate the distribution y^u is

y^u



$E(y^u) = 0$

y^u



$E(y^u) = 0$

y^u

$$E(y^u) = -ln\left(\frac{1}{5}\right) = ln5$$

$$E(y^u) = -\sum_{m=1}^{5} y_m^u ln(y_m^u)$$

As small as possible

$$L = \sum_{x^r} C(y^r, \hat{y}^r)$$ labelled data

$$+\lambda \sum_{x^u} E(y^u)$$ unlabeled data

Semi-supervised Learning
Smoothness Assumption

近朱者赤，近墨者黑

“You are known by the company you keep”

Smoothness Assumption

- Assumption: "similar" x has the same $\hat{y}$
- More precisely:
  - x is not uniform.
  - If x^1 and x^2 are close in a high density region, $\hat{y}^1$ and $\hat{y}^2$ are the same.

connected by a high density path

公館 v.s. 台北車站

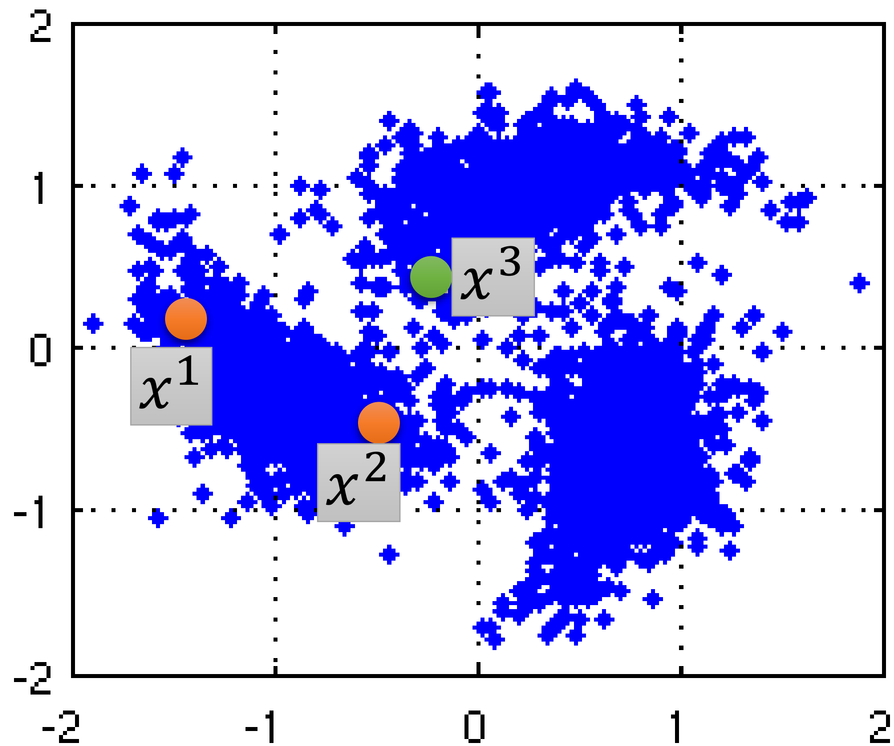
公館 v.s. 科技大樓

Source of image: http://hips.seas.harvard.edu/files/pinwheel.png

Smoothness Assumption

- Assumption: "similar" x has the same $\hat{y}$
- More precisely:
  - x is not uniform.
  - If x^1 and x^2 are close in ~~a high density region~~, $\hat{y}^1$ and $\hat{y}^2$ are the same.

connected by a high density path

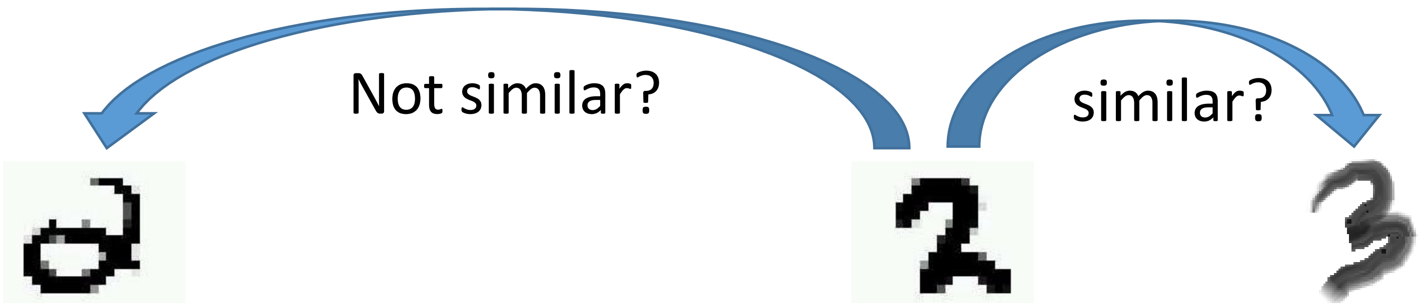


x^1 and x^2 have the same label

x^2 and x^3 have different labels

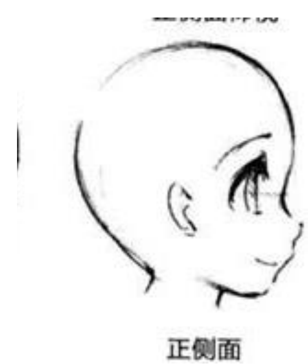
Source of image: http://hips.seas.harvard.edu/files/pinwheel.png

Smoothness Assumption



¨indirectly¨ similar
with stepping stones

(The example is from the tutorial slides of Xiaojin Zhu.)

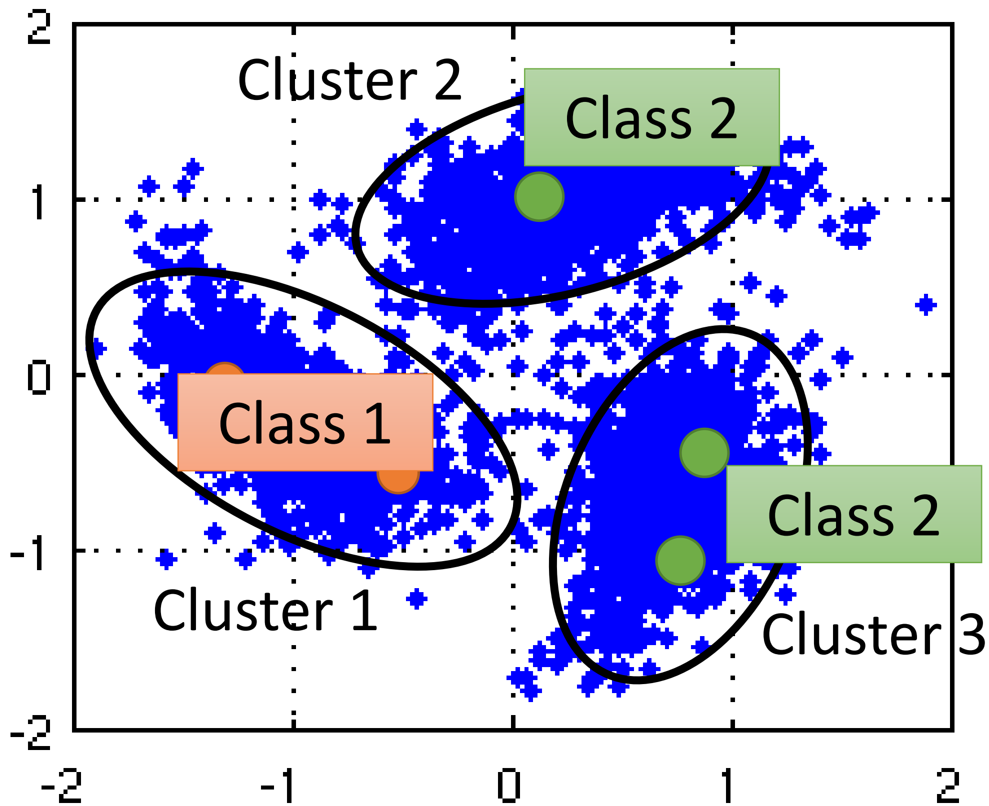


Source of image: http://www.moehui.com/5833.html/5/

Cluster and then Label

Class 1

Class 2



Using all the data to learn a classifier as usual

Graph-based Approach

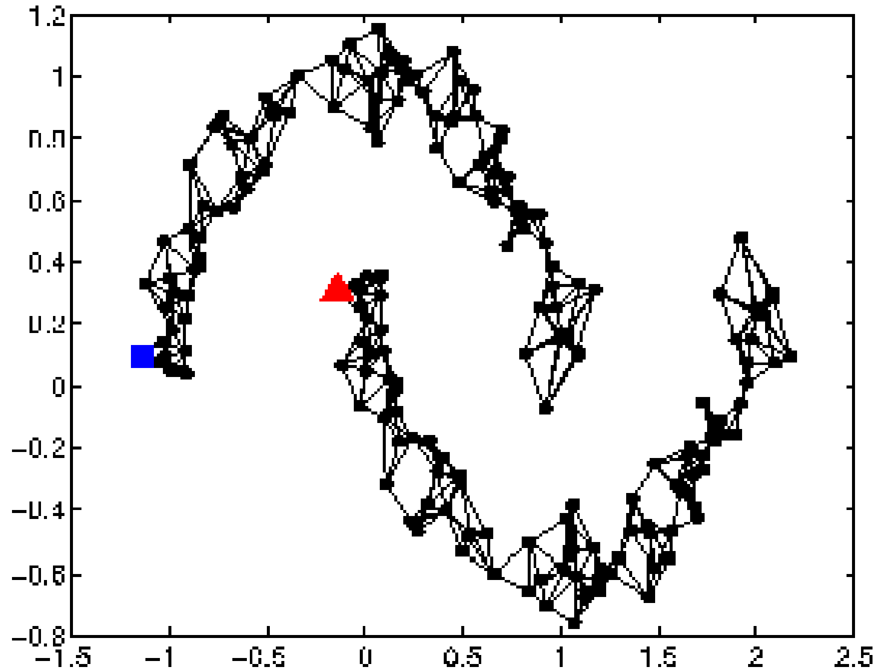
- How to know x^1 and x^2 are connected by a high density path

Represented the data points as a ***graph***

Graph representation is nature sometimes.

E.g. Hyperlink of webpages, citation of papers

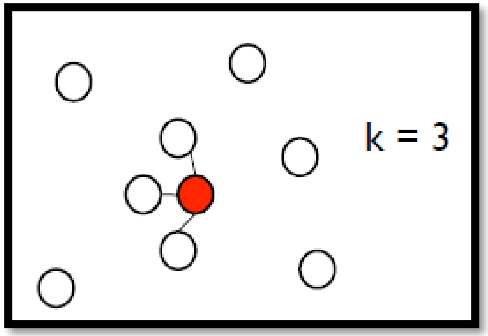
Sometimes you have to construct the graph yourself.

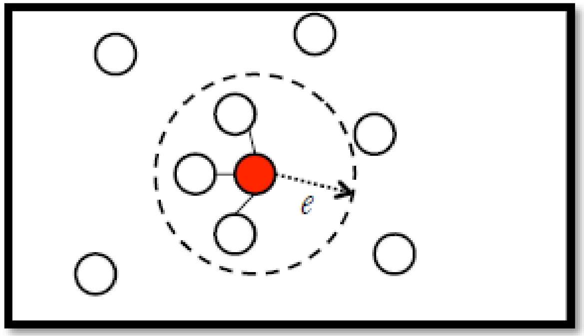


Graph-based Approach - Graph Construction

The images are from the tutorial slides of **Amarnag Subramanya and Partha Pratim Talukdar**

- Define the similarity $s(x^i, x^j)$ between x^i and x^j
- Add edge:
  - K Nearest Neighbor
  - *e*-Neighborhood

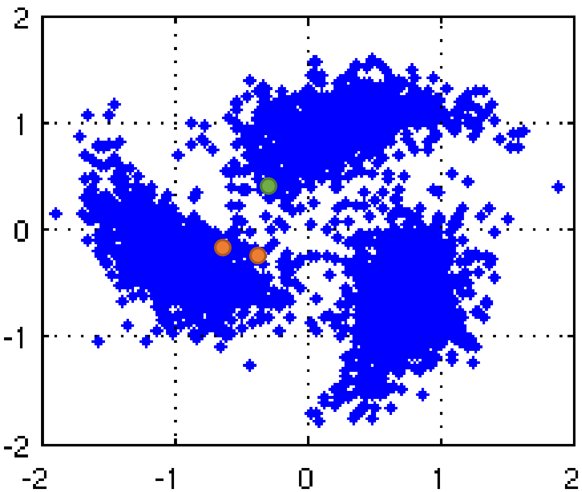




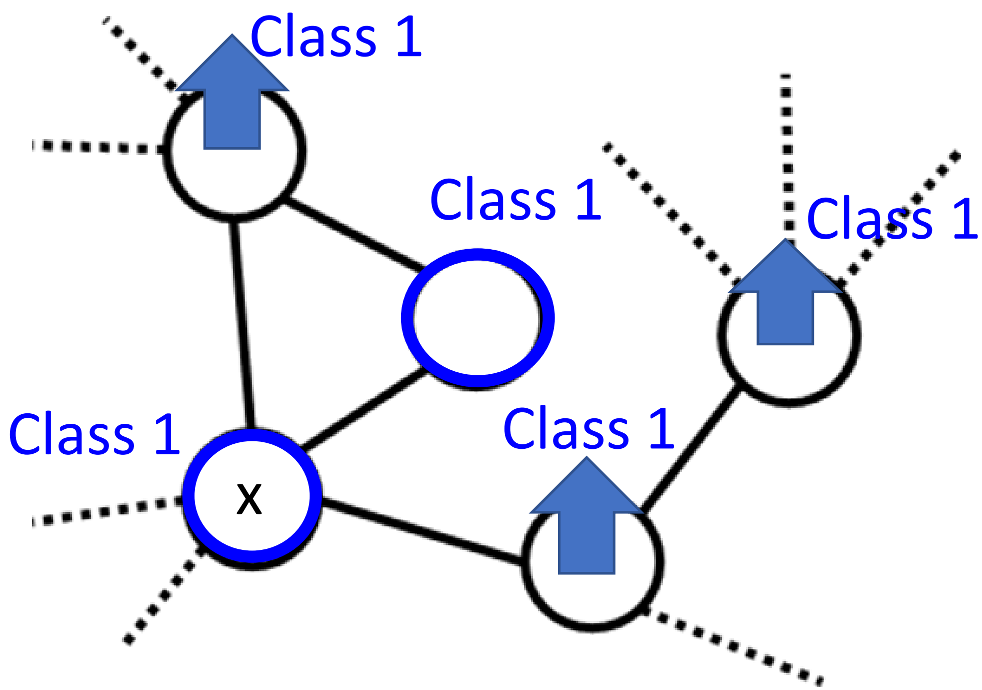
- Edge weight is proportional to $s(x^i, x^j)$

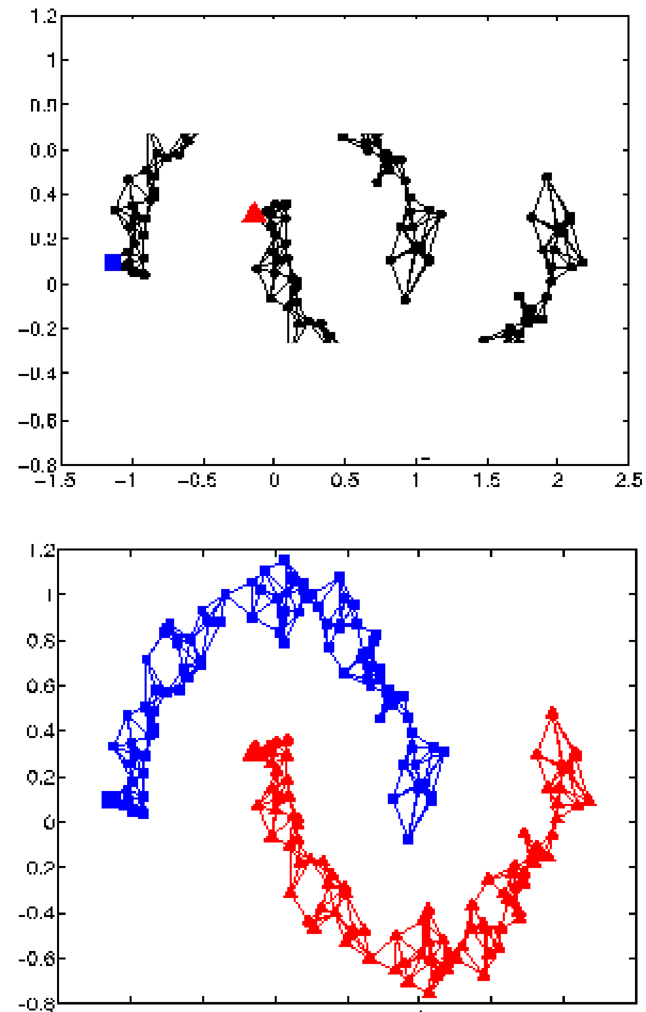
Gaussian Radial Basis Function:

$$s(x^i, x^j) = exp\left(-\gamma \left\|x^i - x^j\right\|^2\right)$$



Graph-based Approach





The labelled data influence their neighbors.

Propagate through the graph

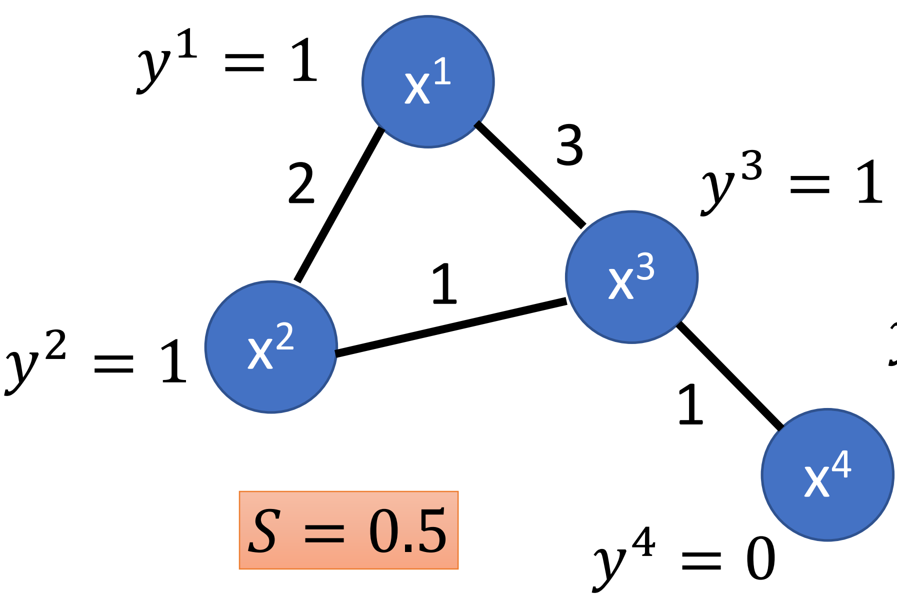
Graph-based Approach

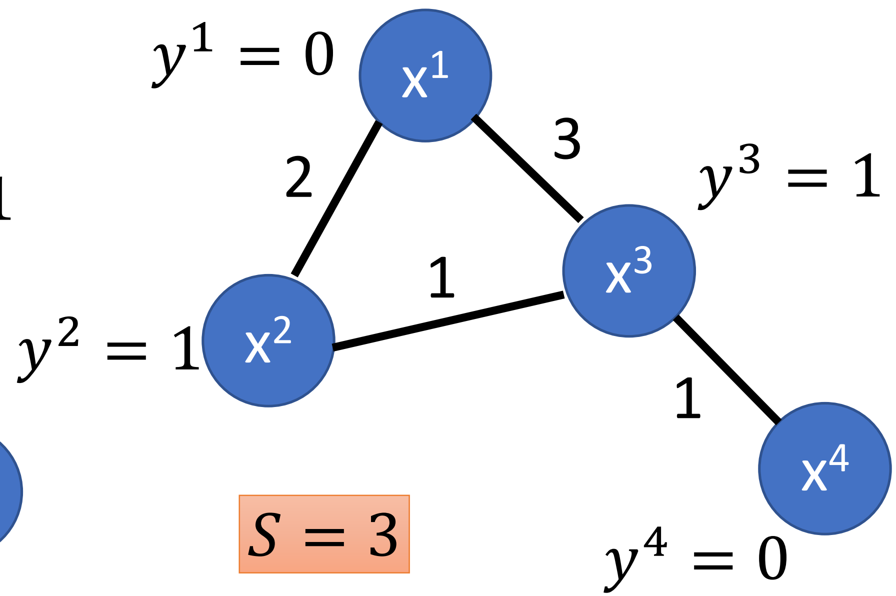
- Define the smoothness of the labels on the graph

$$S = \frac{1}{2}\sum_{i,j} w_{i,j}\left(y^i - y^j\right)^2$$

Smaller means smoother

For all data (no matter labelled or not)

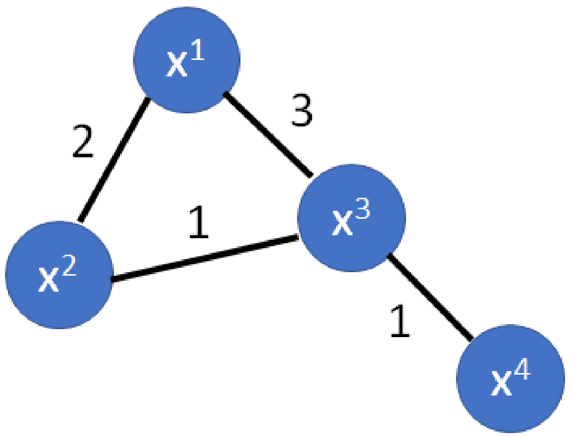




Graph-based Approach

- Define the smoothness of the labels on the graph

$$S = \frac{1}{2}\sum_{i,j} w_{i,j}\left(y^i - y^j\right)^2 = \boldsymbol{y}^T L \boldsymbol{y}$$



y: (R+U)-dim vector

$$\boldsymbol{y} = \left[\cdots y^i \cdots y^j \cdots\right]^{\boldsymbol{T}}$$

L: (R+U) x (R+U) matrix

Graph Laplacian

$$L = D - W$$

$$W = \begin{bmatrix} 0 & 2 & 3 & 0 \\ 2 & 0 & 1 & 0 \\ 3 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad \mathrm{D} = \begin{bmatrix} \boxed{5} & 0 & 0 & 0 \\ 0 & \boxed{3} & 0 & 0 \\ 0 & 0 & \boxed{5} & 0 \\ 0 & 0 & 0 & \boxed{1} \end{bmatrix}$$

Graph-based Approach

- Define the smoothness of the labels on the graph

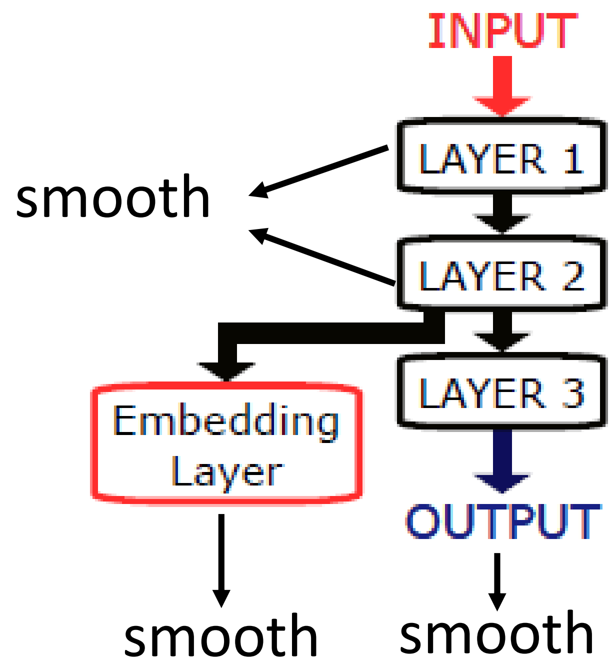
$$S = \frac{1}{2}\sum_{i,j} w_{i,j}\left(y^i - y^j\right)^2 = \boldsymbol{y}^T L \boldsymbol{y}$$

Depending on model parameters

$$L = \sum_{x^r} C\left(y^r, \hat{y}^r\right) \boxed{+\lambda S}$$

As a regularization term

J. Weston, F. Ratle, and R. Collobert, “Deep learning via semi-supervised embedding,” ICML, 2008

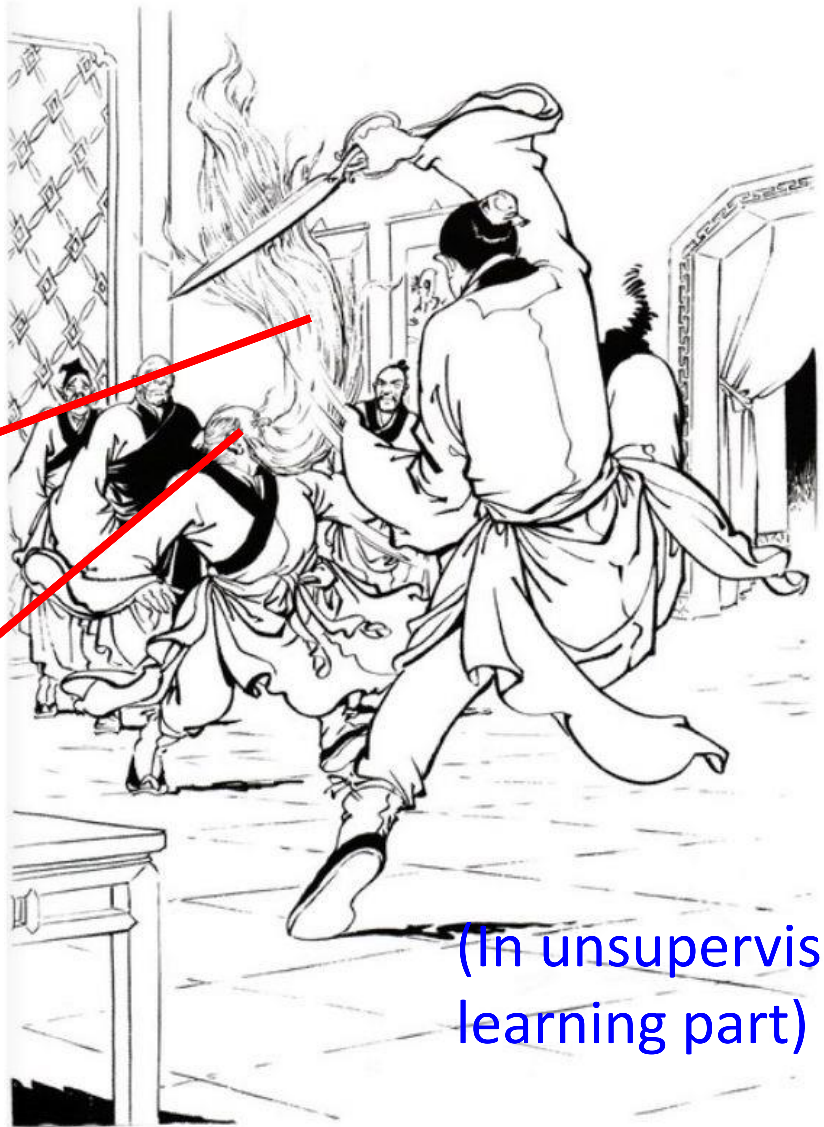


Semi-supervised Learning
Better Representation

去蕪存菁，化繁為簡

Looking for Better Representation

- Find a better (simpler) representations from the unlabeled data

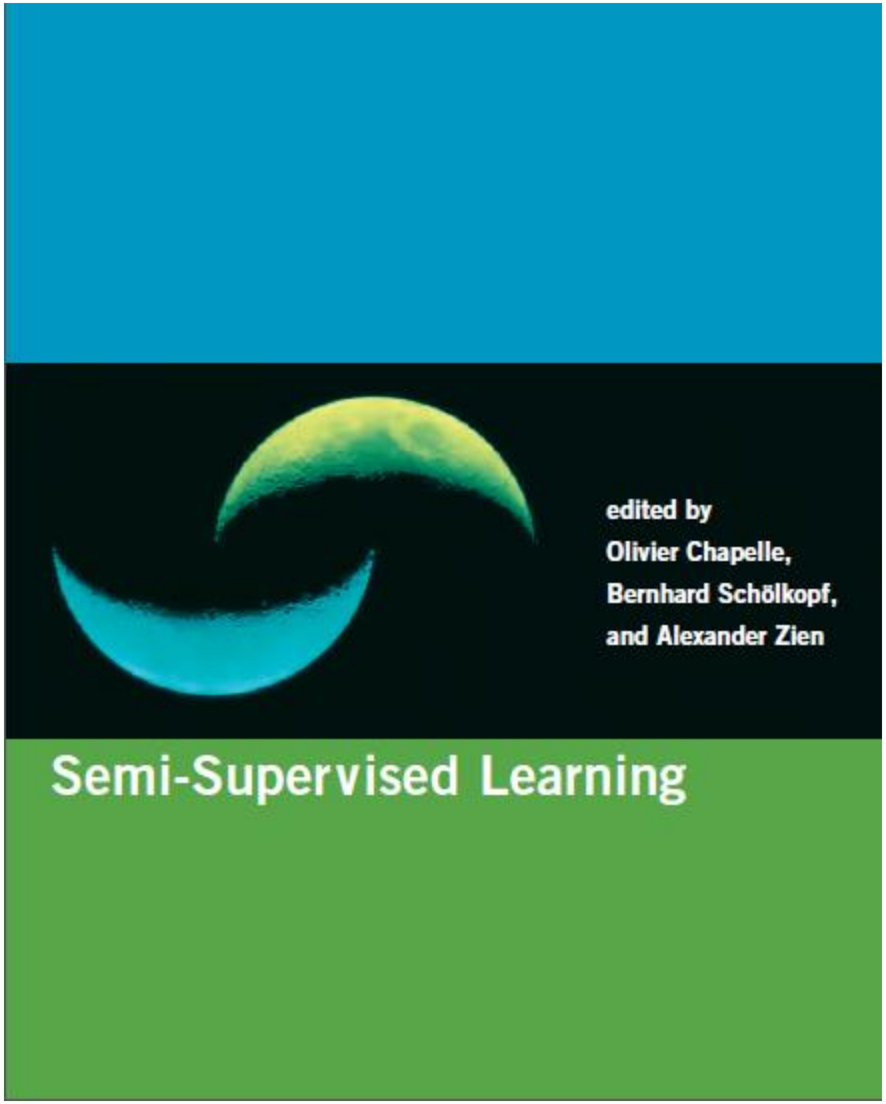


Original representation

Better representation

(In unsupervised learning part)

Reference



http://olivier.chapelle.cc/ssl-book/

Acknowledgement

- 感謝 劉議隆 同學指出投影片 上的錯字
- 感謝 丁勃雄 同學指出投影片 上的錯字