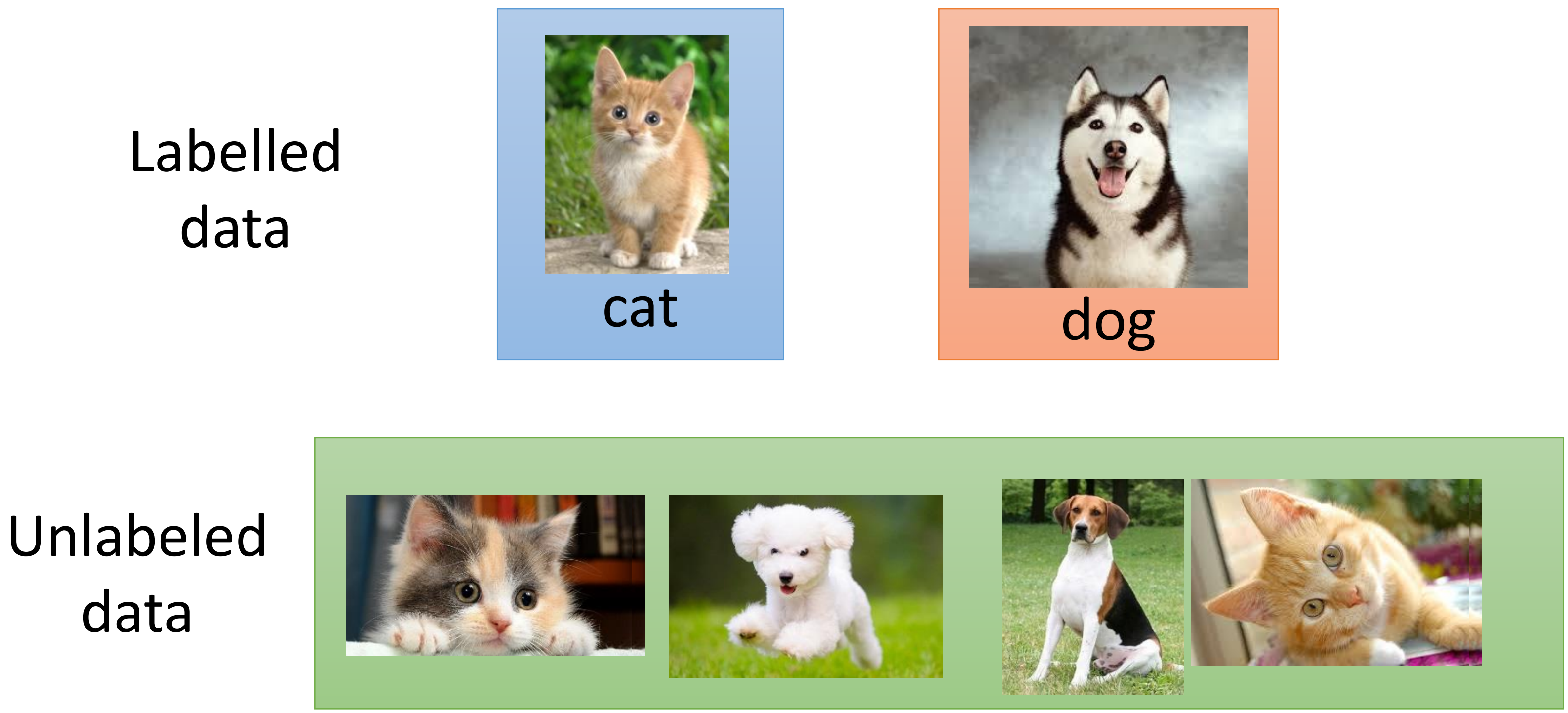
Semi-supervised Learning

Introduction



(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

有一堆unlabel data，只有一點label data

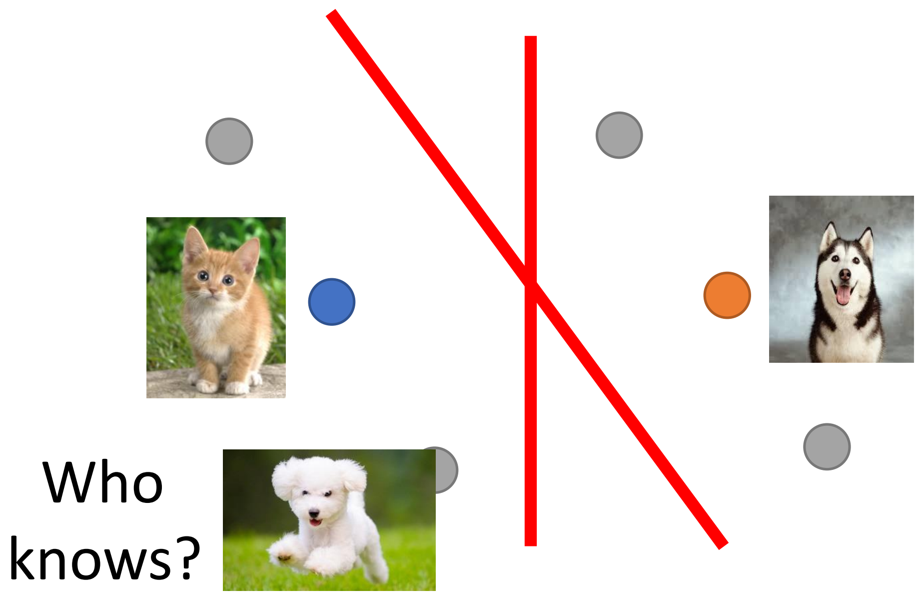
- Semi-supervised learning: $\{(x^r, \hat{y}^r)\}_{r=1}^{R}$, $\{x^u\}_{u=R}^{R+U}$
  - A set of unlabeled data, usually U >> R
  - Transductive learning: unlabeled data is the testing data
  - Inductive learning: unlabeled data is not the testing data

兩種要不要把testing data包含在training 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?

semi-supervise learning常常伴隨一些假設



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) share gaussian performance比較好
  - P(x|C_i) is a Gaussian parameterized by μ^i and Σ

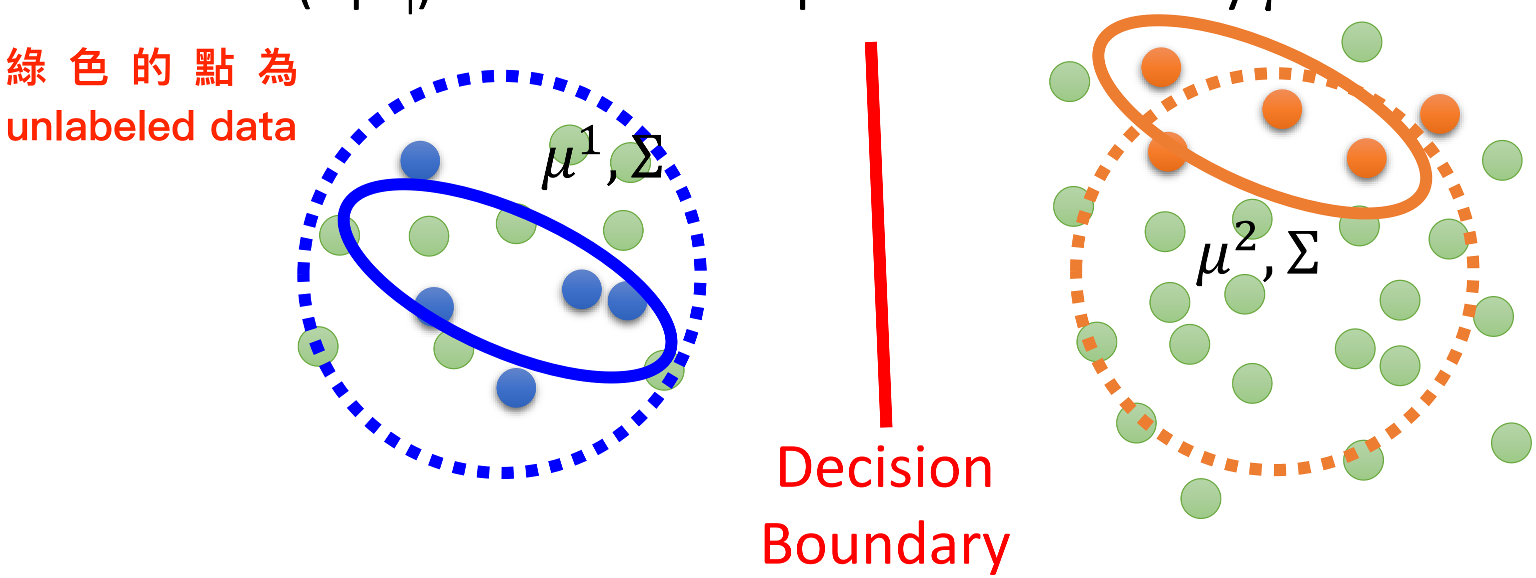
With $P(C_1)$, $P(C_2)$,μ^1,μ^2, Σ

maximum likelihood

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

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 $P(C_1), P(C_2), \mu^1, \mu^2, \Sigma$

Semi-supervised Generative Model

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

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

E

- Step 1: compute the posterior probability of unlabeled data

只考慮labeled data

posterior prob.

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

iteration到收斂

Back to step 1

M

- Step 2: update model labeled+unlabeled data一起考慮

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

unlabel data的預測結果

N: total number of examples

N_1: number of examples belonging to C_1

只考慮labeled data

每筆unlabeled data屬於class i 的機率

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

normalized

Why?

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

- Maximum likelihood with labelled data

$$logL(\theta) = \sum_{(x^r, \hat{y}^r)} logP_\theta(x^r | \hat{y}^r)$$

- Maximum likelihood with labelled + unlabeled data

$$logL(\theta) = \sum_{(x^r, \hat{y}^r)} logP_\theta(x^r | \hat{y}^r) + \sum_{x^u} logP_\theta(x^u)$$

unlabeled data出現的機率

Solved iteratively

$$\boxed{P_\theta(x^u) = P_\theta(x^u | C_1)P(C_1) + P_\theta(x^u | C_2)P(C_2)}$$

(x^u can come from either C_1 and C_2)

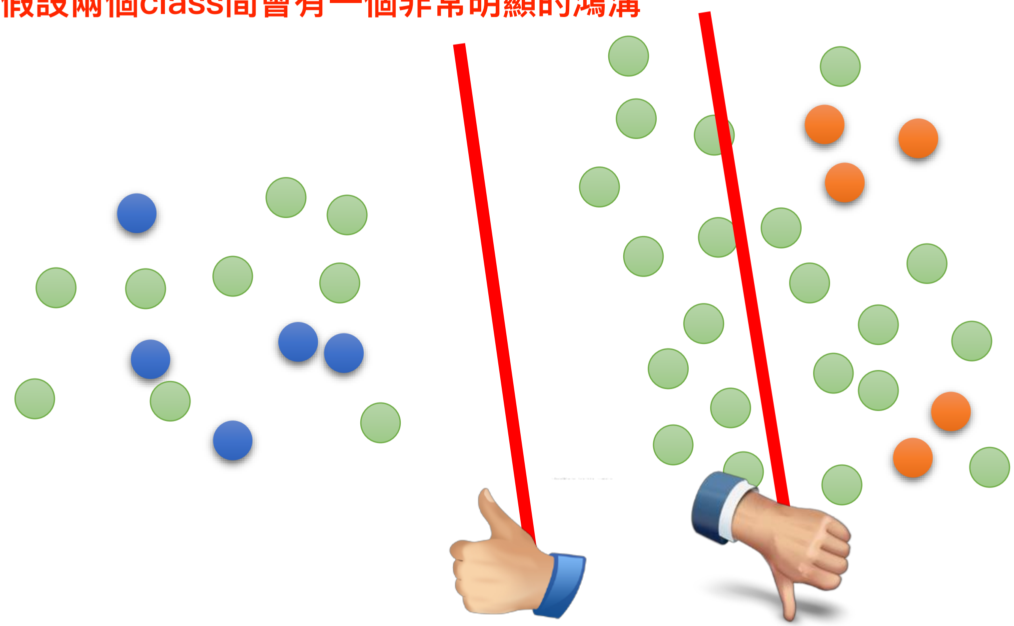
xu這筆data從c1,c2生成的機率

Semi-supervised Learning
Low-density Separation

假設兩個class間會有一個非常明顯的鴻溝

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:

  f*: DNN,decision tree,SVM…都可以

  - Train model f^* from labelled data set

    regression沒用

    You can use any model here. Regression?

  - Apply f^* to the unlabeled data set 根據train好的model去預測unlabel data
    - 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 ex. confidence高的unlabel data加入到labeled data中，可依據這樣給出每個unlabeled data一個weight

    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

給訂label　　給訂label的機率
- Hard label v.s. Soft label

  Considering using neural network

  θ^* (network parameter) from labelled data

$x^u \rightarrow \boxed{\theta^*} \rightarrow \begin{bmatrix} 0.7 \\ 0.3 \end{bmatrix}$

Hard

Class 1 → New target for x^u is $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$

It looks like class 1, then it is class 1.

沒有幫助的！

Soft

70% Class 1
30% Class 2 → New target for x^u is $\begin{bmatrix} 0.7 \\ 0.3 \end{bmatrix}$

Doesn't work …

同一組參數當然output會一樣啊！

在做neural　network的classification中一定要用hard label，“非黑即白”

generative model因為會腦補所以可以用soft label

Entropy-based Regularization

看distribution的entropy，越小越好！！

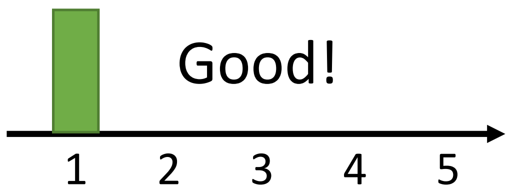
欲求

$$x^u \rightarrow \boxed{\theta^*} \rightarrow y^u$$

Distribution

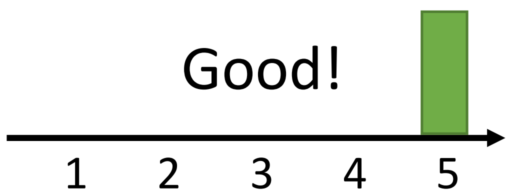
Entropy of y^u :
Evaluate how concentrate the distribution y^u is

y^u Good!



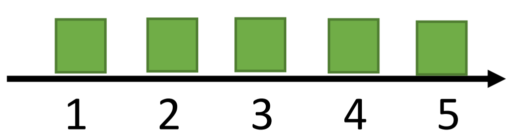
$$E(y^u) = 0$$

y^u Good!



$$E(y^u) = 0$$

y^u Bad!



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

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

regression

As small as possible

loss function: gradient descend

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

cross entropy

labelled data

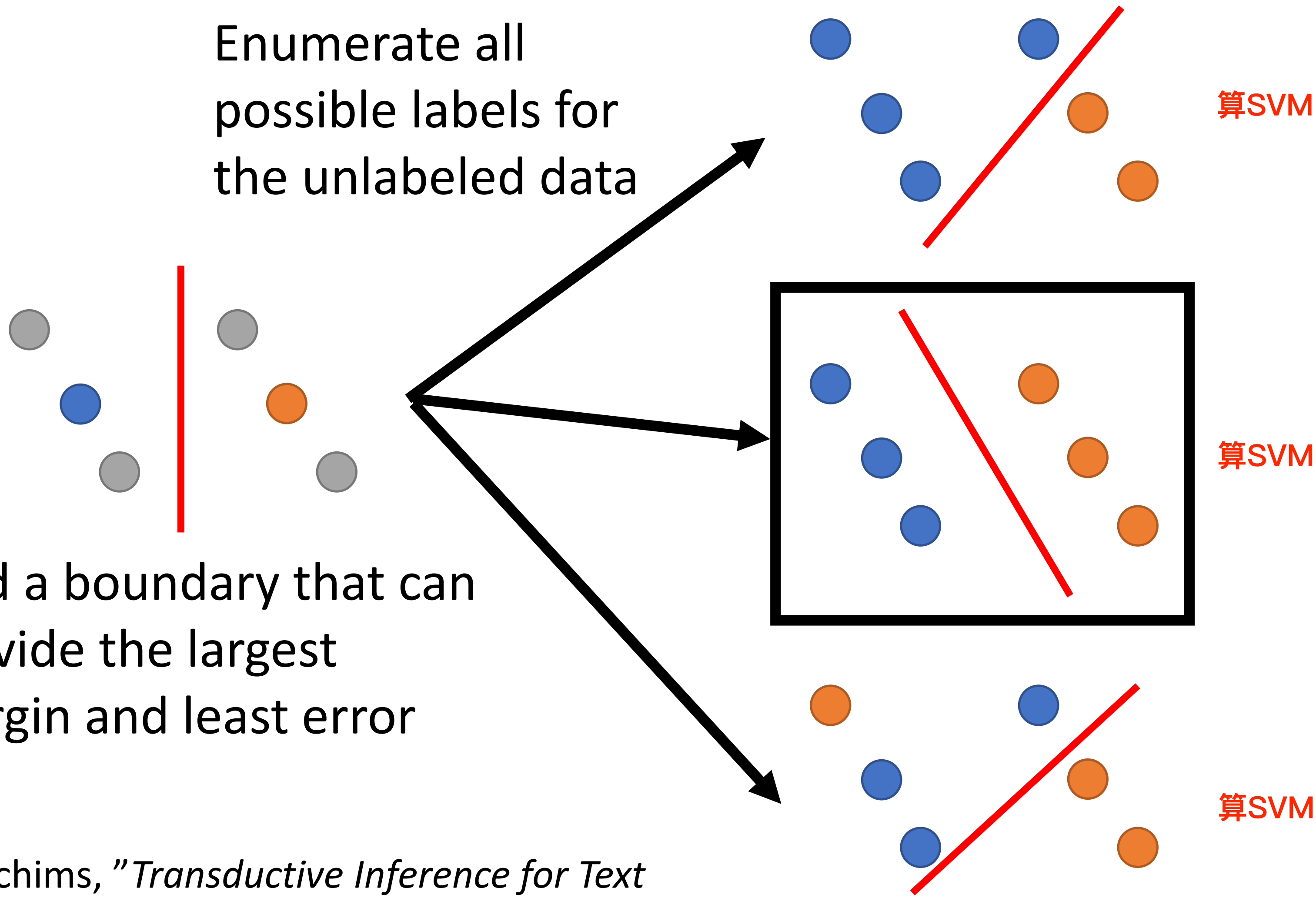
$$+\lambda \sum_{x^u} E(y^u)$$

unlabeled data

窮舉所有unlabeled data 可能的assignment

Outlook: Semi-supervised SVM

Enumerate all possible labels for the unlabeled data



Find a boundary that can provide the largest margin and least error

Thorsten Joachims, ”*Transductive Inference for Text Classification using Support Vector Machines”,* ICML, 1999

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: 精確的假設
  - 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

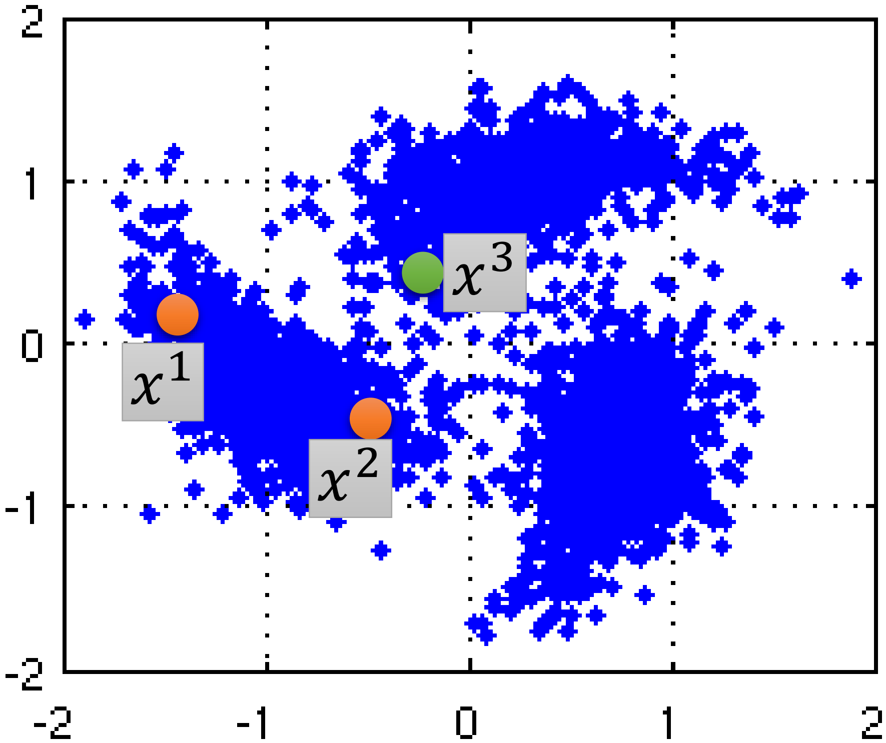
公館 v.s. 科技大樓

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

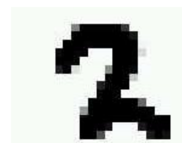
Source of image: http://hips.seas.harvard.edu/files/pinwheel.png

Smoothness Assumption

Not similar?

similar?

中間有過度的型態



中間沒有過度的型態

“indirectly” similar
with stepping stones

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

中間有過度的型態

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

Smoothness Assumption

文學分類

- Classify astronomy vs. travel articles

data量不夠多，中間沒有過度的狀態找不到overlap

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

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

Smoothness Assumption

- Classify astronomy vs. travel articles

data量要夠多才能辨別誰跟誰像，找到中間的過渡狀態

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

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

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

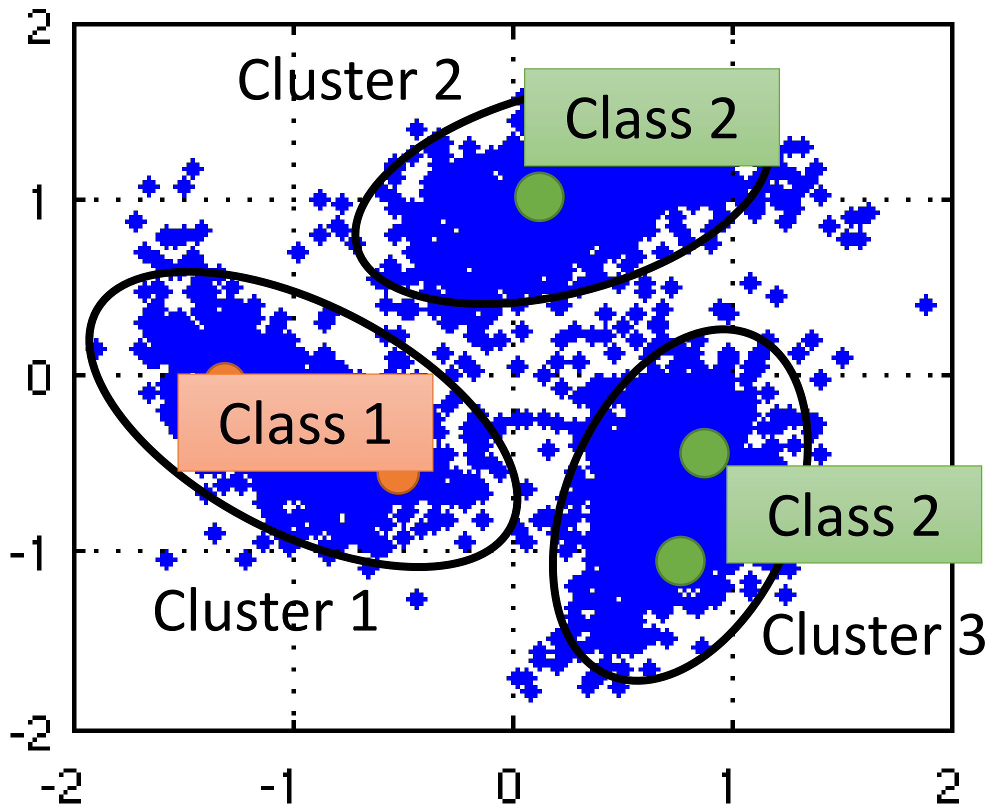
Cluster and then Label

看哪個distribution中哪個label比較多就依照這個來label

Class 1

Class 2

藍色是unlabel data



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

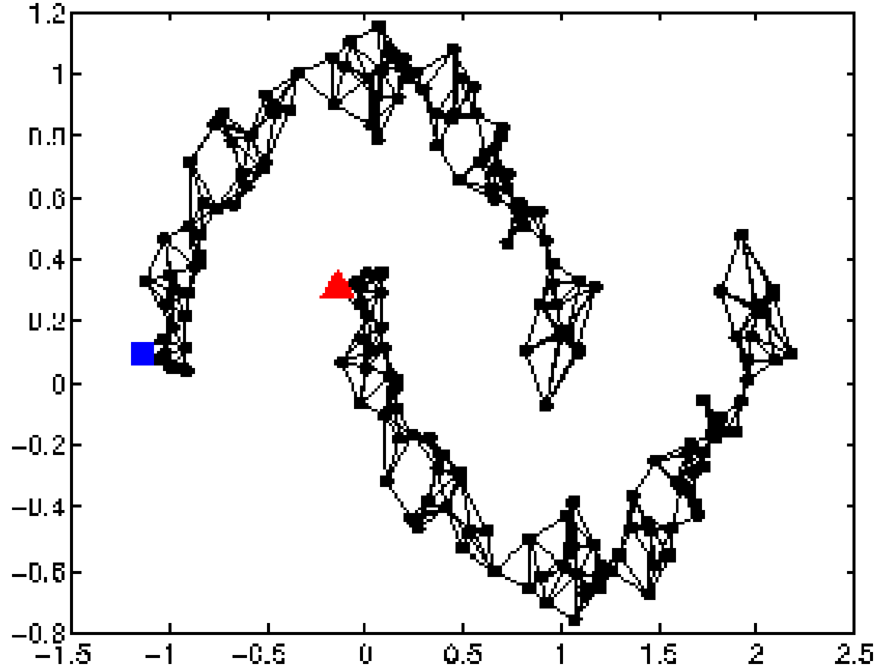
利用建立一個graph找出path，有path就是同一個class

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

Graph representation is nature sometimes.

E.g. Hyperlink of webpages, citation of papers

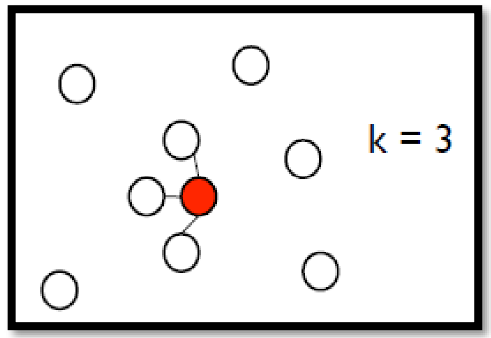
Sometimes you have to construct the graph yourself.

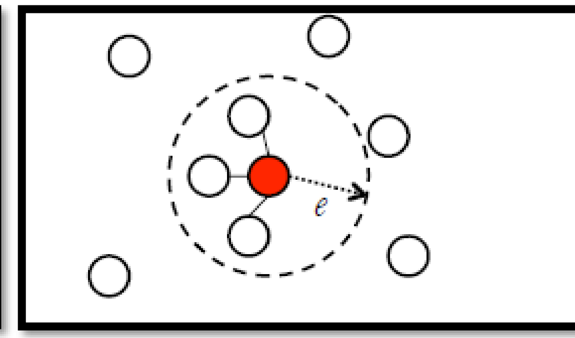


Graph-based Approach - Graph Construction

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 算相似度
- Add edge:
  - K Nearest Neighbor
    鄰近k個點
  - *e*-Neighborhood
    threshold e





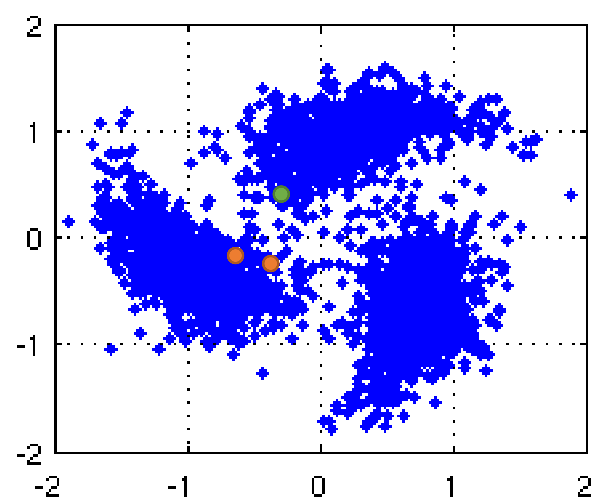
- Edge weight is proportional to $s(x^i, x^j)$
  每個edge給一個weight

Gaussian Radial Basis Function:

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

euclidean distance

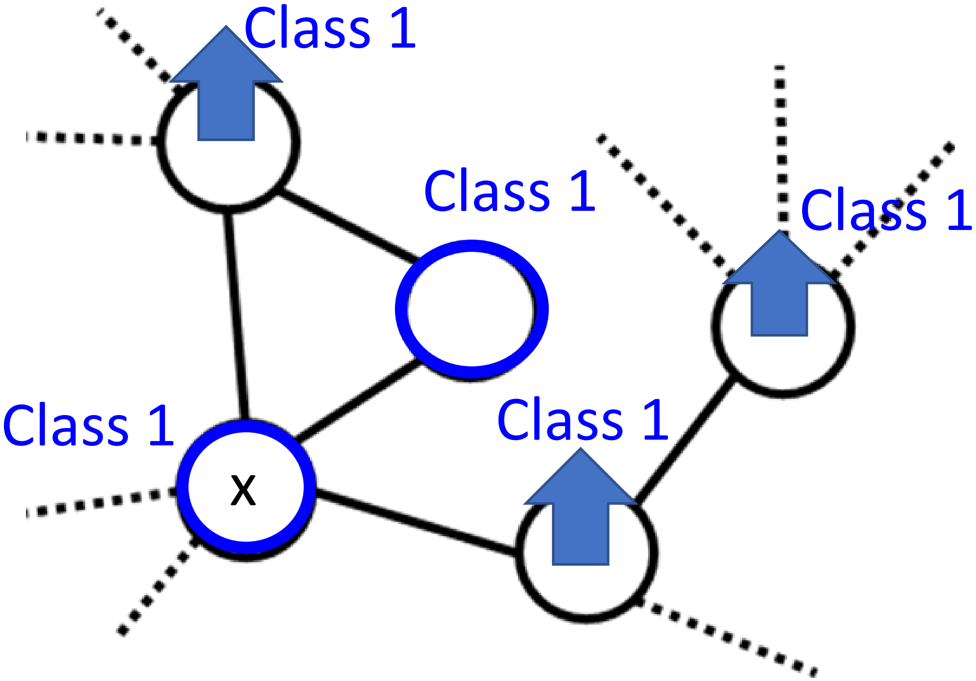
只要差一點距離就會被這個function拉大

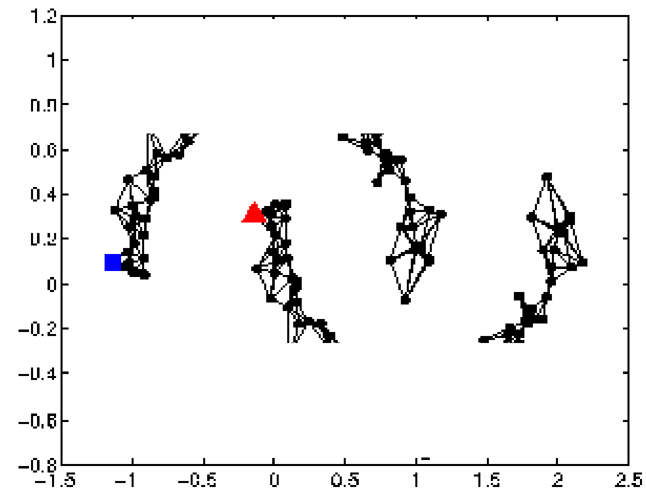


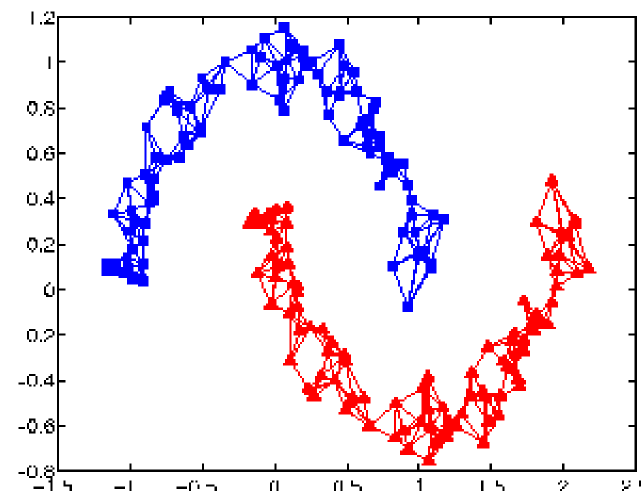
建好graph後根據labeled data散步道unlabeled data

Graph-based Approach

前提： labeled data要夠多！！！







The labelled data influence their neighbors.

Propagate through the graph

Graph-based Approach

evaluation major

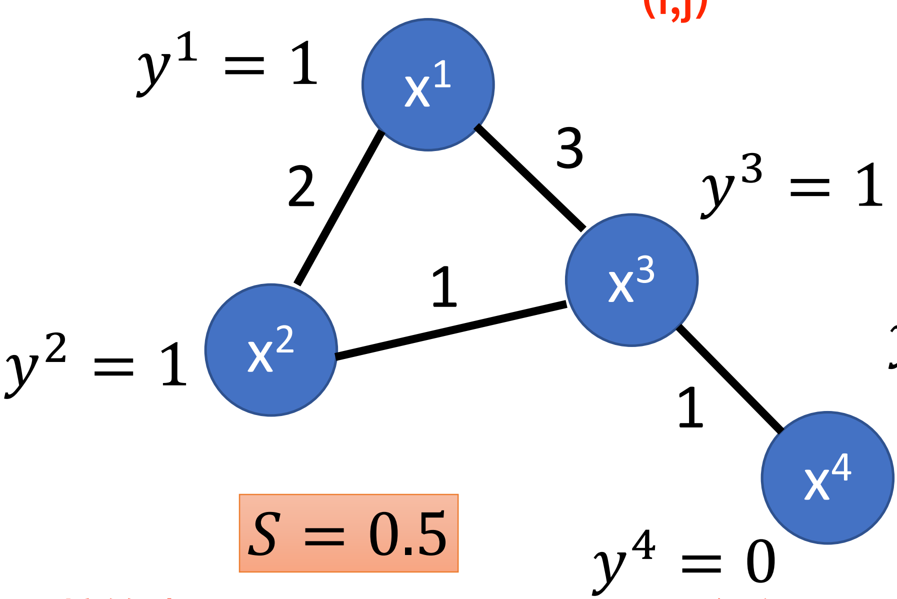
- Define the smoothness of the labels on the graph

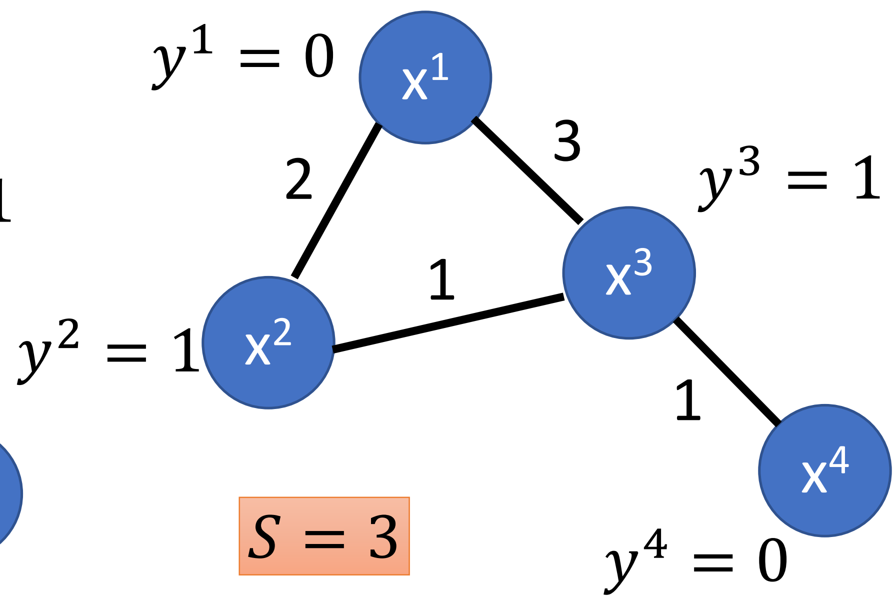
$$S = \frac{1}{2}\sum_{i,j} w_{i,j}\left(y^i - y^j\right)^2$$

Smaller means smoother

For all data (no matter labelled or not)

(i,j)





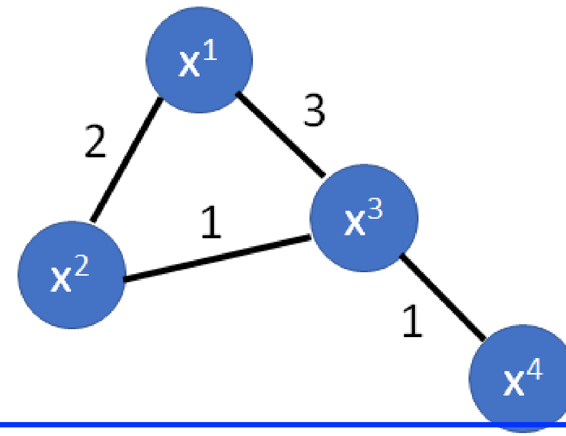
比較符合smoothness assumption假設

Graph-based Approach

- Define the smoothness of the labels on the graph

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

notation轉換



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

每個row的sum

Graph Laplacian

$$L = \underline{D} - \underline{W}$$

Laplacian

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

graph上edge的weight

Graph-based Approach

- Define the smoothness of the labels on the graph

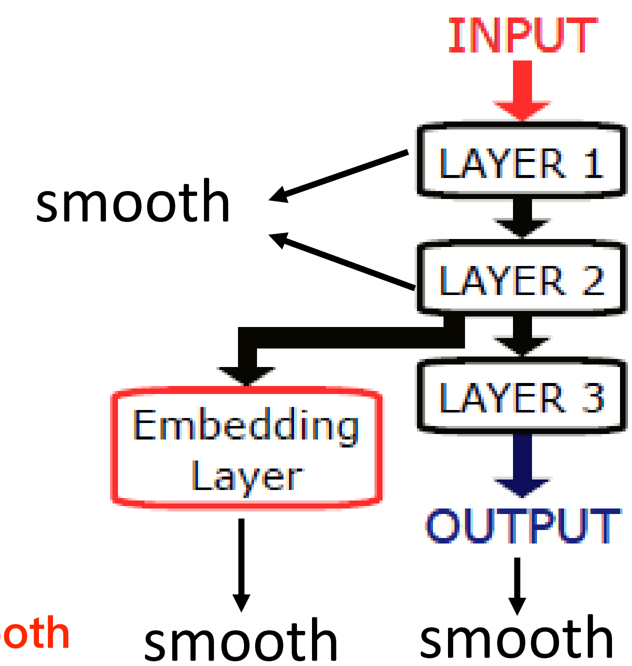
$$S = \frac{1}{2}\sum_{i,j} w_{i,j}\left(y^i - y^j\right)^2 = \boldsymbol{y}^T L \boldsymbol{y}$$

Depending on model parameters

smoothness越小越好

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

As a regularization term



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

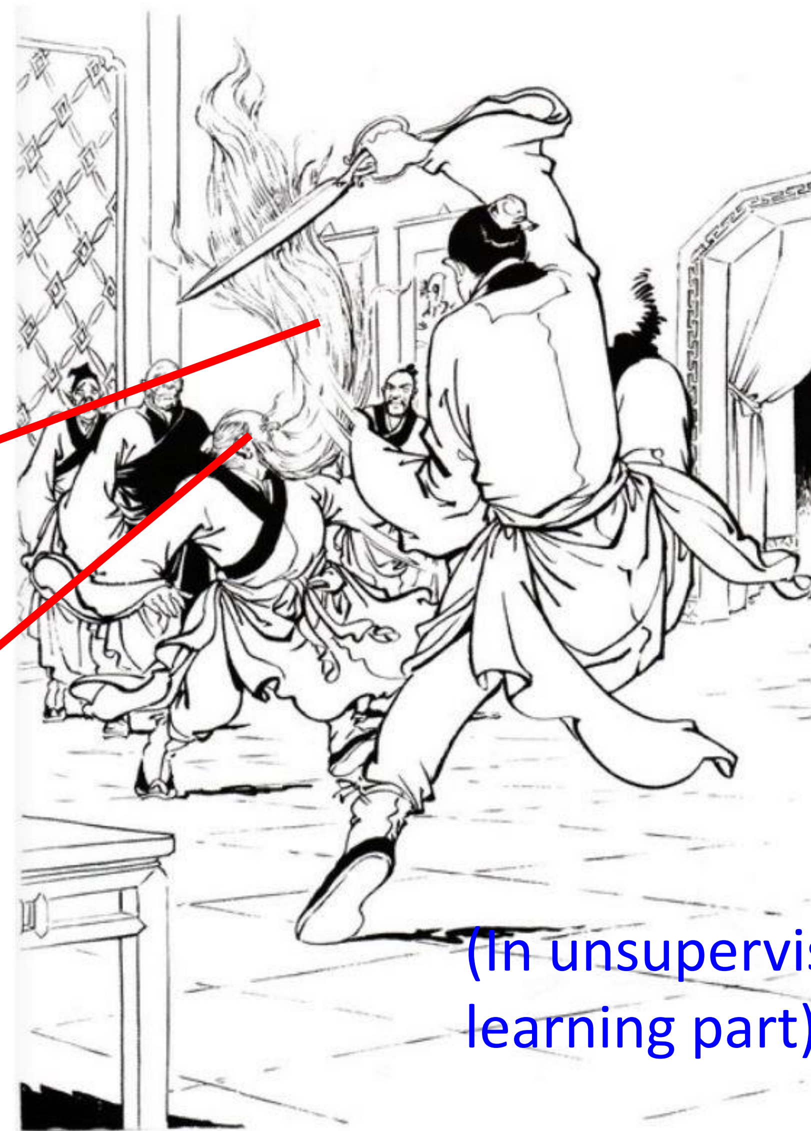
每個layer都可以符合smooth

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

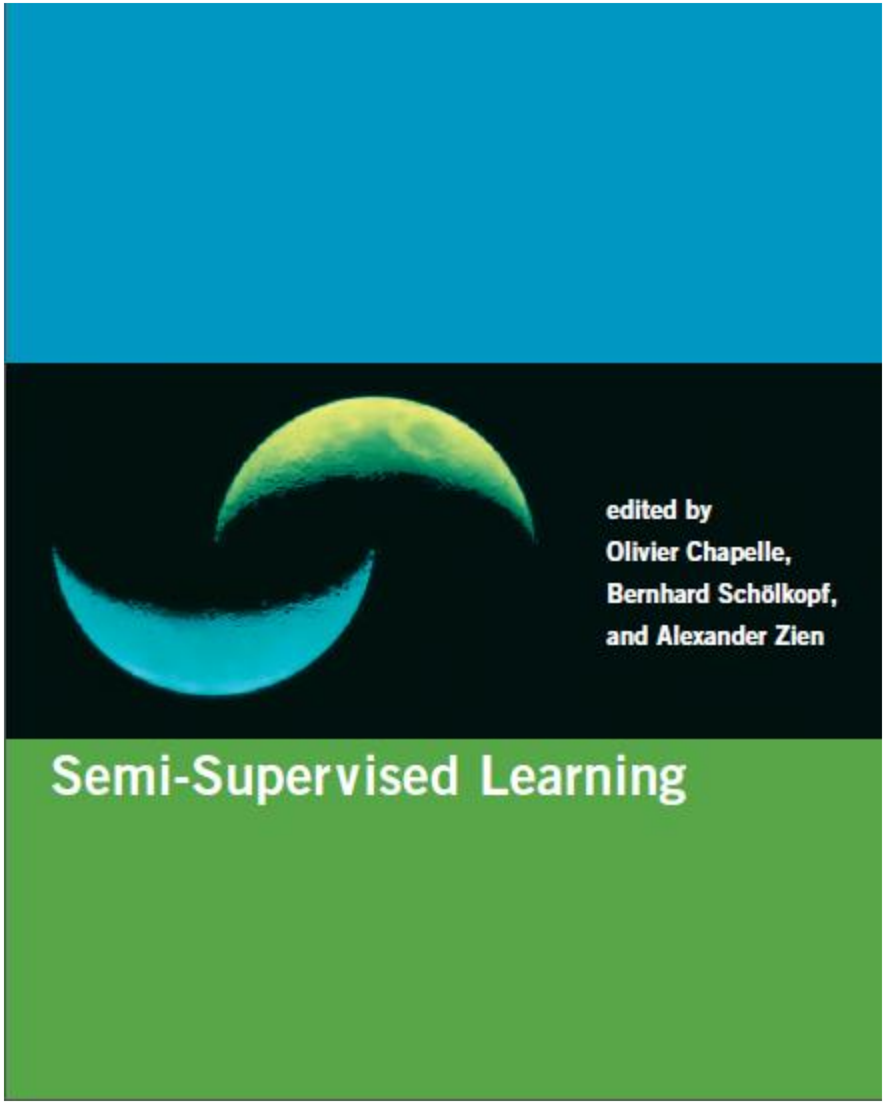


Original representation

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