Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

弱监督学习中的半监督技术

硕士研究生　谢崇玮

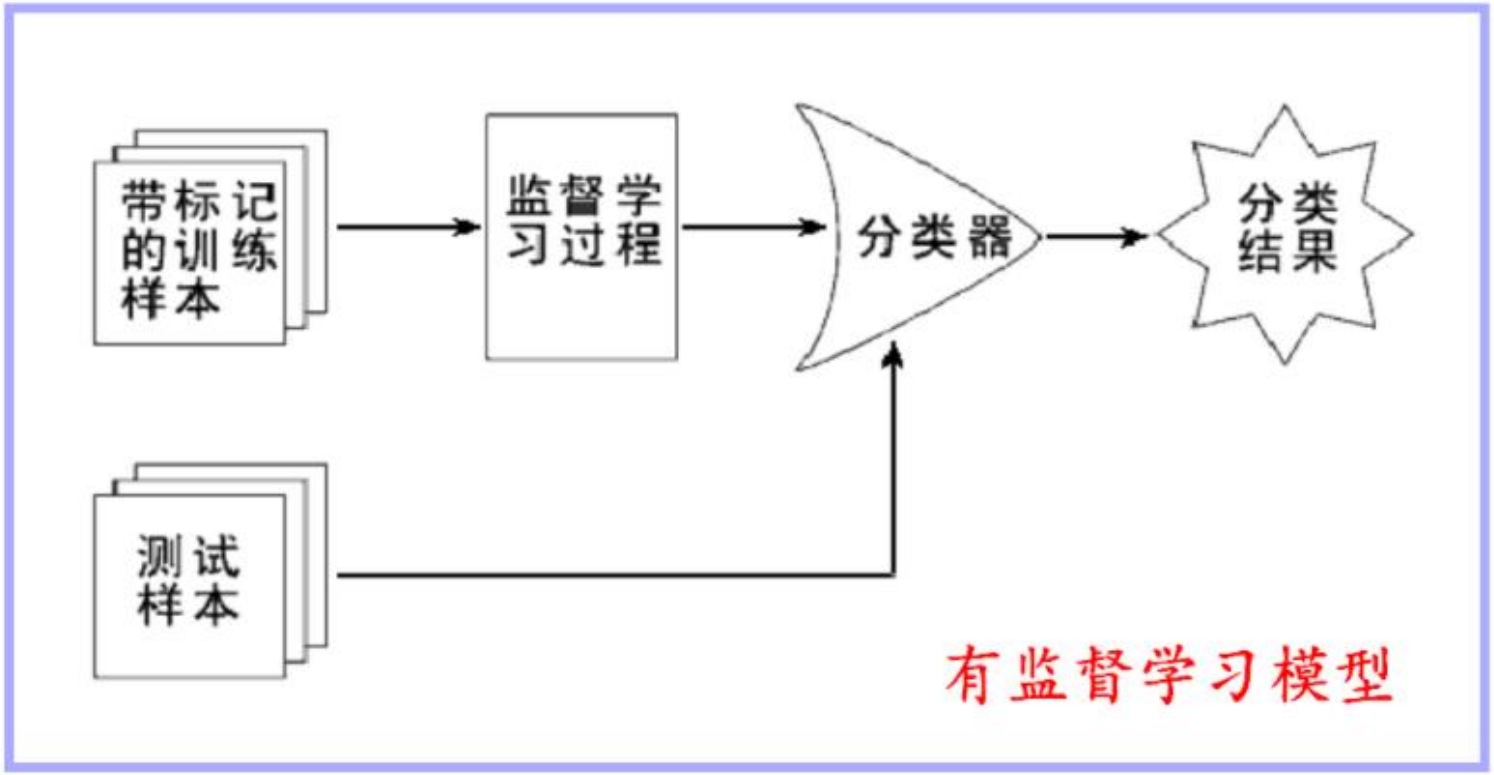
2022年02月13日

背景简介

- 预期收获
  - 1. 了解半监督技术的产生背景与发展历程
  - 2. 理解半监督技术的分类与基本原理
  - 3. 了解半监督技术的典型应用与未来发展

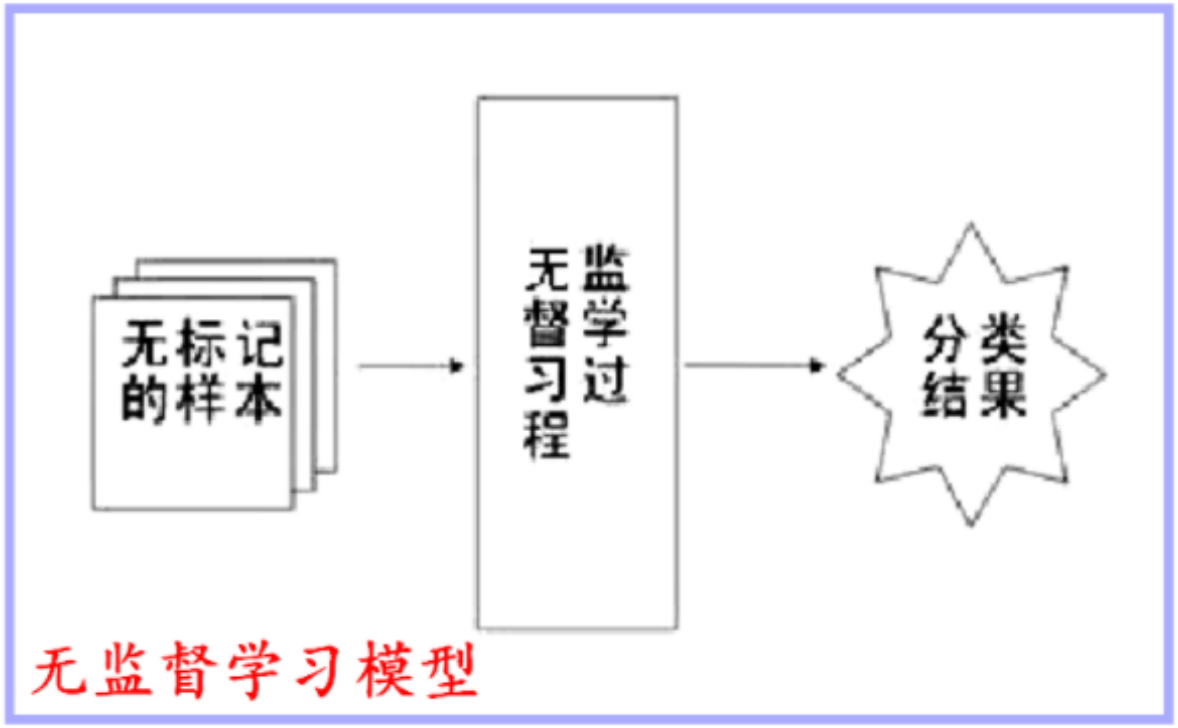
背景简介

- 监督学习（Supervised Learning）
  - 学习器通过对大量有标记的训练例进行学习，从而建立模型应用于预测未见示例的标记。但很难获得大量标记样本（耗时、耗钱）



背景简介

- 无监督学习（Unsupervised Learning）
  - 无训练样本，仅根据测试样本在特征空间分布情况来进行标记。但准确性差



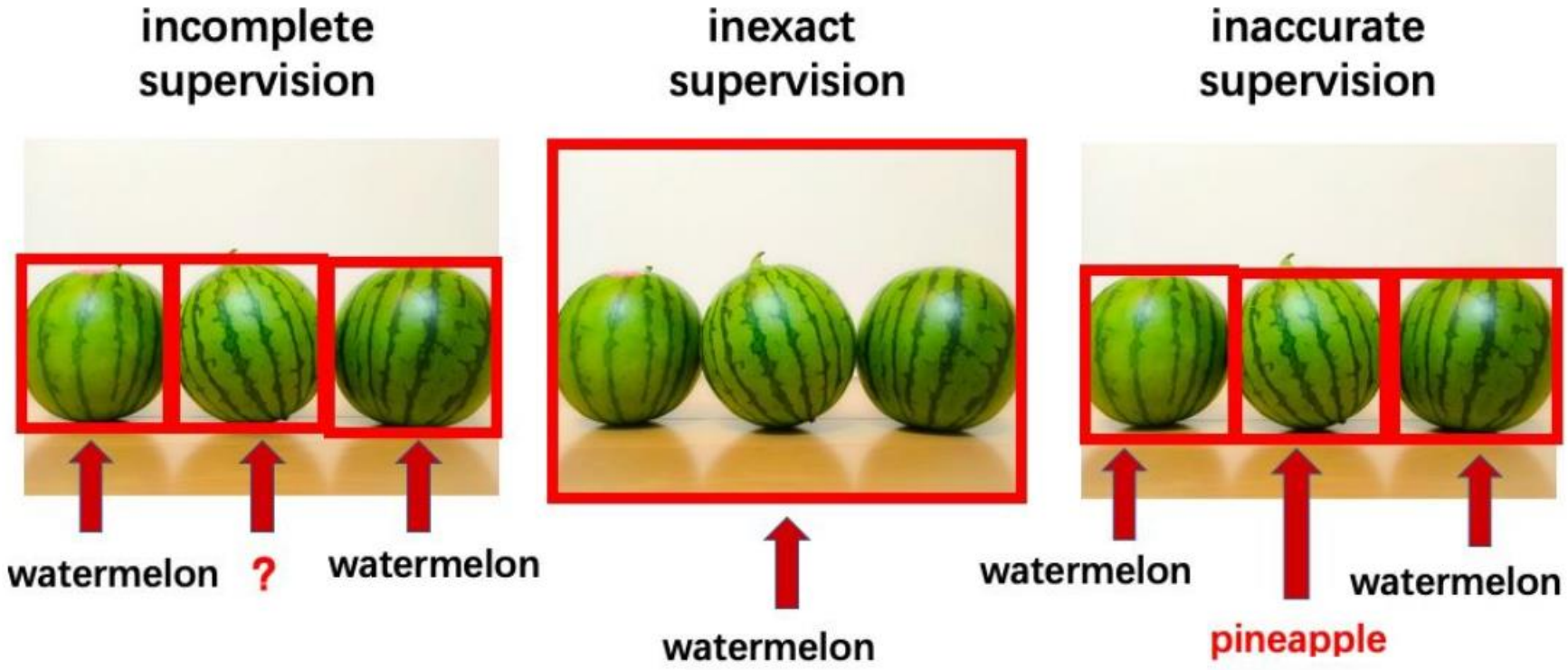
背景简介

- 弱监督学习（Weakly supervised Learning）
  - 有少量训练样本，学习机以从训练样本获得的知识为基础，结合测试样本的分布情况逐步修正已有知识，并判断样本的类别。

基本概念

基本概念

- 三种典型的类型（不完全监督、不确切监督、不精确监督）



It is worth mentioning that in real practice they often occur at the same time.

基本概念

- 实际问题（不完全监督、不确切监督、不精确监督）

“weak supervision” is very common

- incomplete

Image classification

It is easy to get a huge number of images from the Internet, but only a small subset of images can be annotated due to the human cost.

- inexact

Important target detection

Usually we only have image-level labels rather than object-level labels.

- inaccurate

Crowdsourcing data analysis

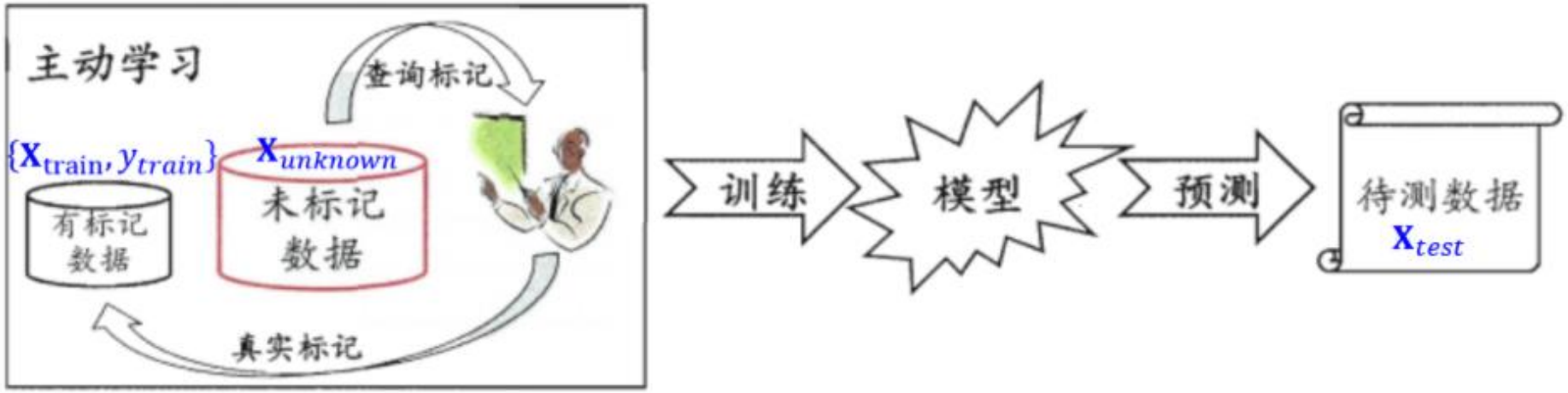
when the image annotator is careless or weary, or some images are difficult to categorize.

基本概念

- 解决不完全监督

□ 主动学习（有人类干预）

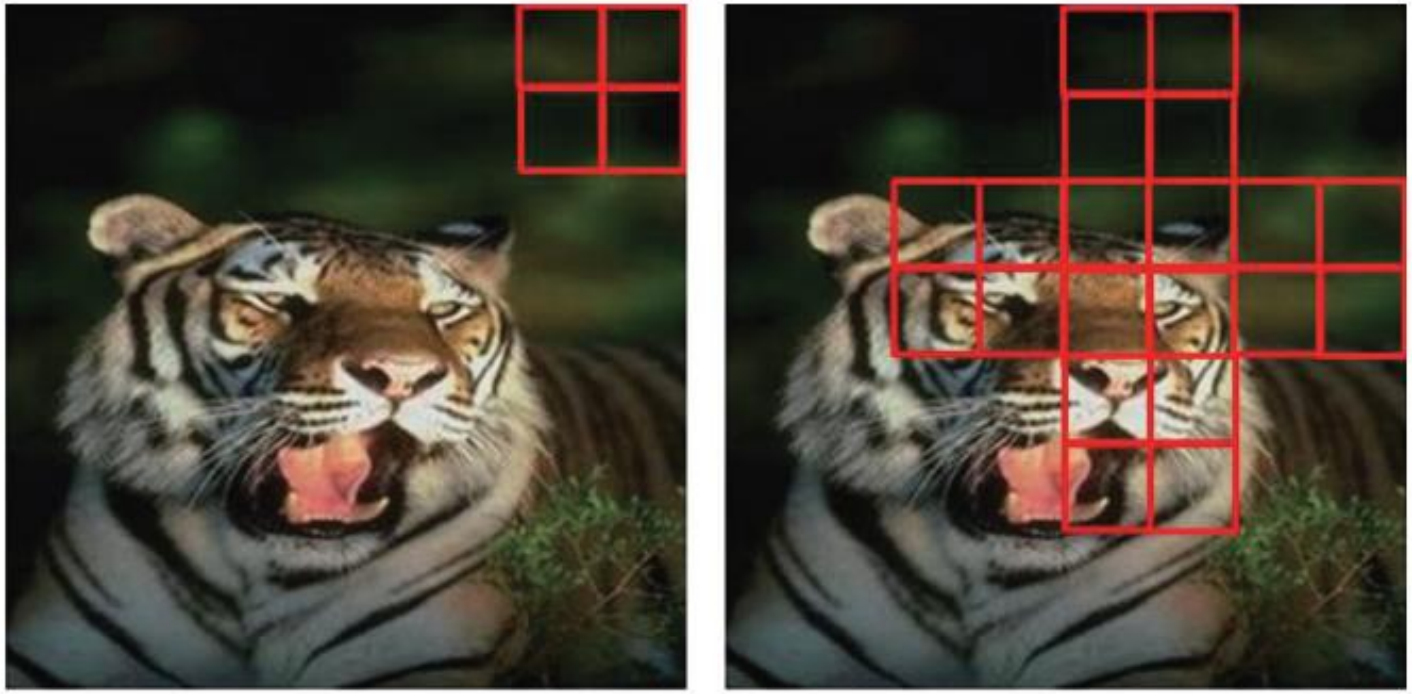
□ 半监督学习（没有人类干预）



基本概念

- 解决不确切监督

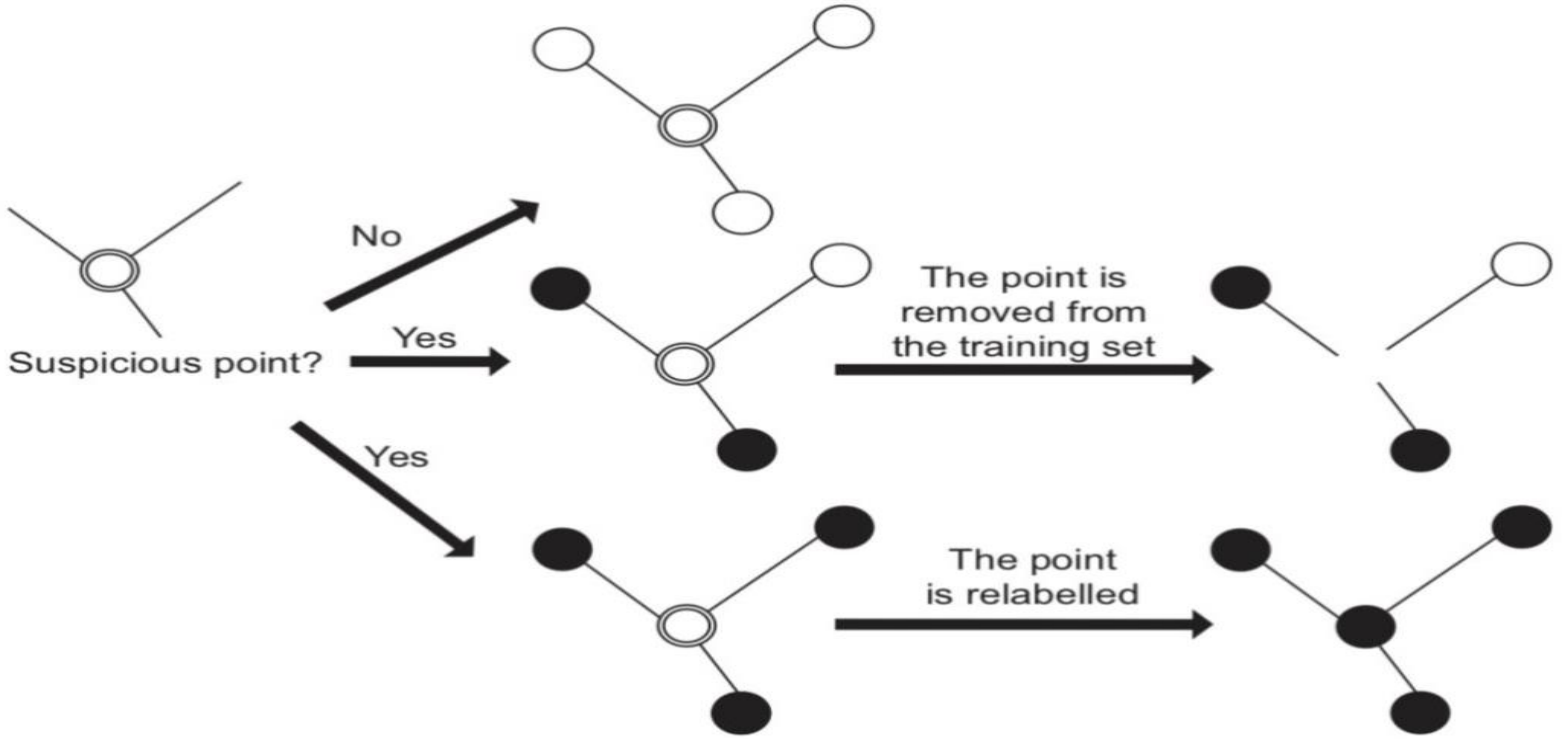
□ 多实例学习



基本概念

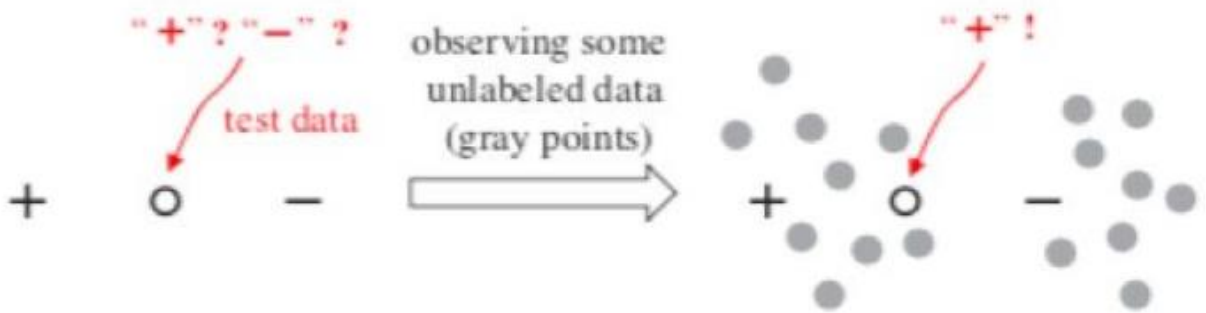
- 解决不精确监督

带噪学习



基本概念

- 半监督学习



- 直推式学习和归纳式学习区别

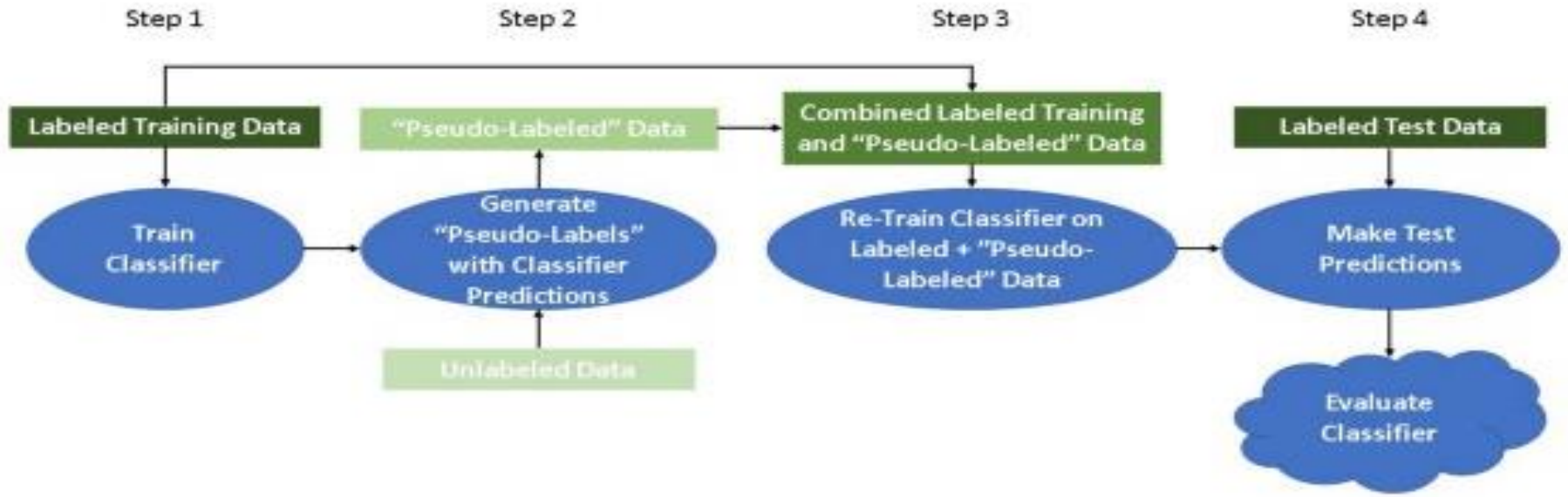
基本概念

- 半监督学习中两种基本的假设
  - 聚类假设（同一聚类中的样本点很可能具有同样的类别标记）
  - 流型假设（处于一个很小的局部邻域内的示例具有相似的性质）
- 主要目的
  - 利用隐藏在大量无标签样本中的数据分布信息来提升仅使用少量有标签样本时的学习性能

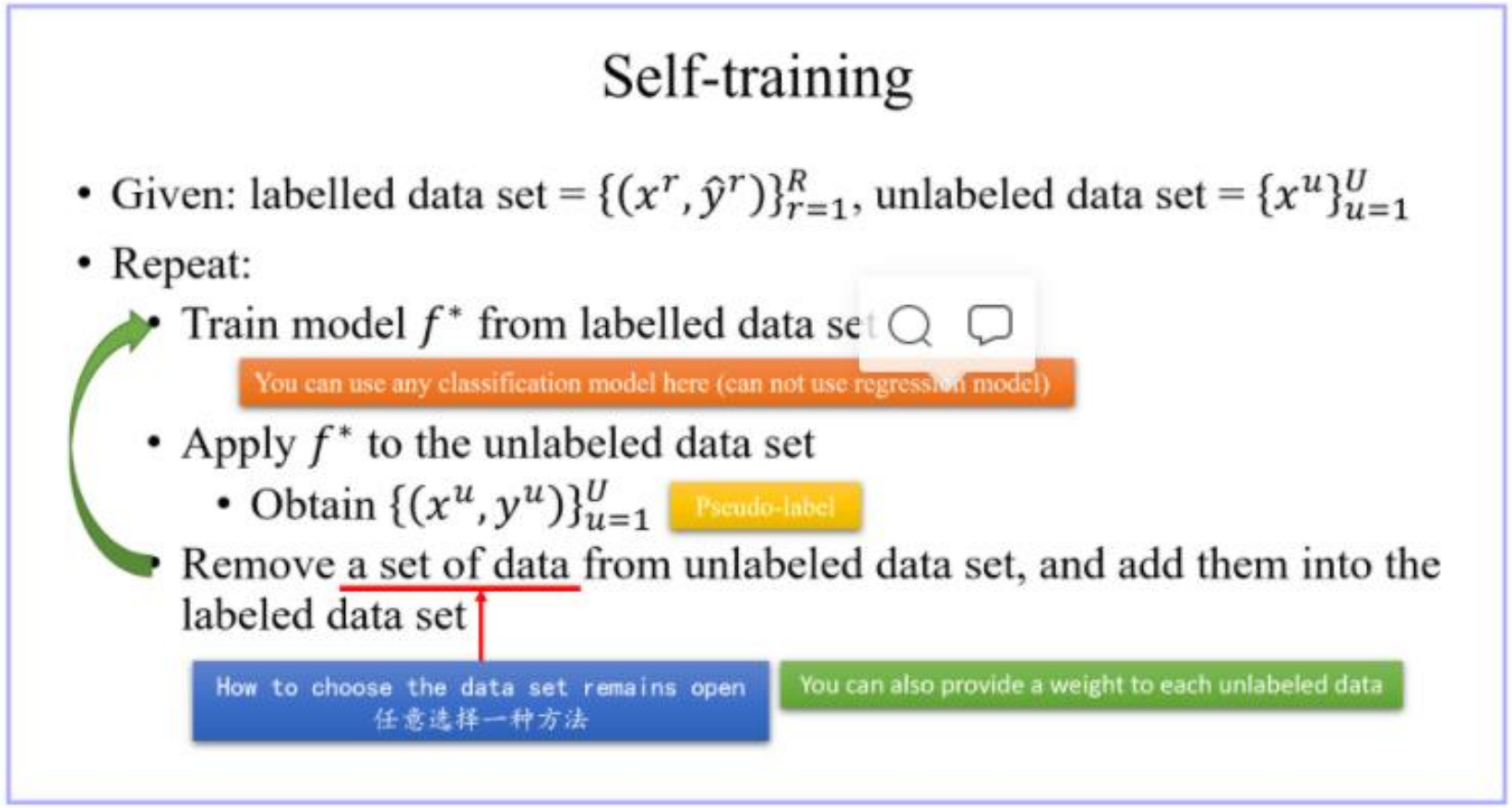
算法原理

算法原理

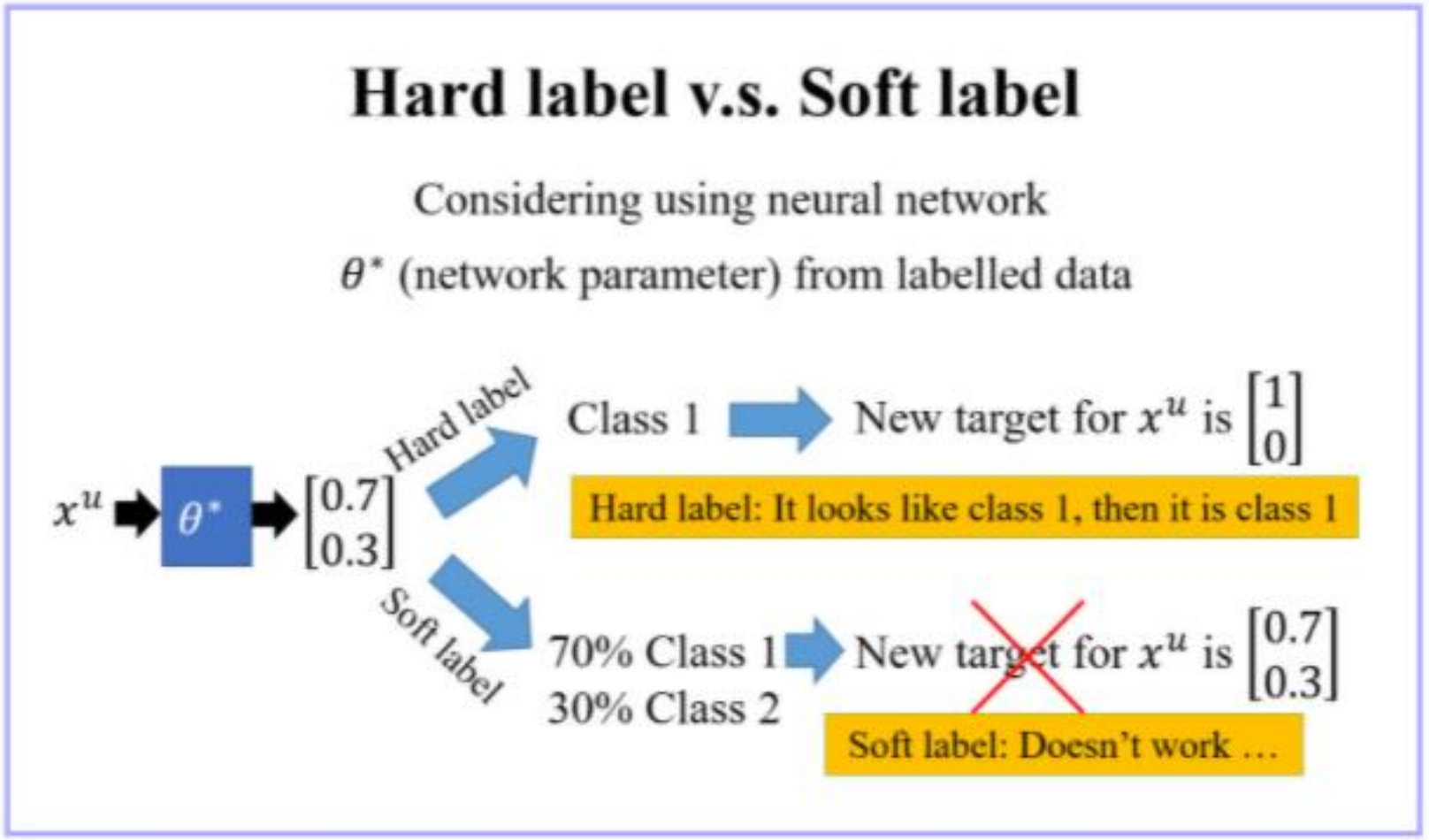
- 低密度分割算法(Low-density Separation)：自训练(Self-training)



算法原理


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 se
You can use any classification model here (can not use regression model)
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
任意选择一种方法
You can also provide a weight to each unlabeled data

算法原理


Hard label v.s. Soft label
Considering using neural network
θ^* (network parameter) from labelled data
x^u
θ^*
$\begin{bmatrix}0.7\\0.3\end{bmatrix}$
Hard label
Soft label
Class 1
New target for x^u is $\begin{bmatrix}1\\0\end{bmatrix}$
Hard label: It looks like class 1, then it is class 1
70% Class 1
30% Class 2
New target for x^u is $\begin{bmatrix}0.7\\0.3\end{bmatrix}$
Soft label: Doesn't work …

算法原理

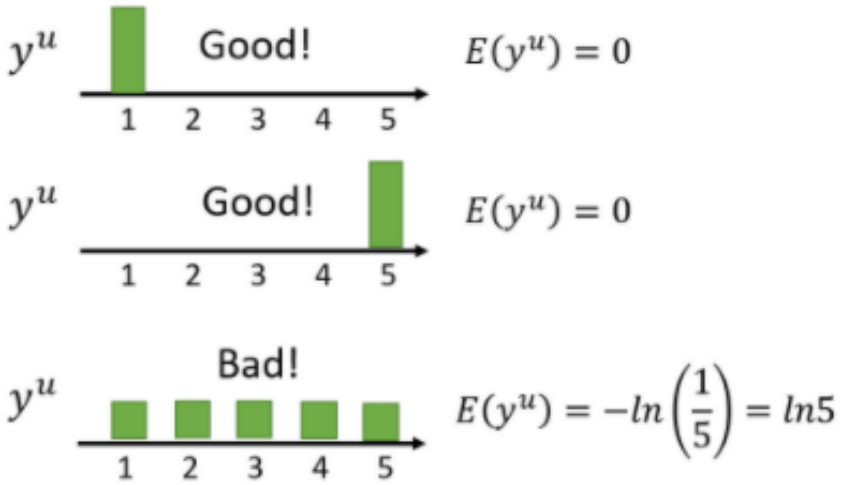
Self-training: Entropy-based Regularization

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

$x^u \rightarrow \theta^* \rightarrow y^u$

Distribution of y^u

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



$$L = \sum_{x^r} L(y^r, \hat{y}^r) + \lambda \sum_{x^u} E(y^u)$$

labelled data　　unlabeled data

算法原理

- 先聚类后标注算法(Cluster and then Label)

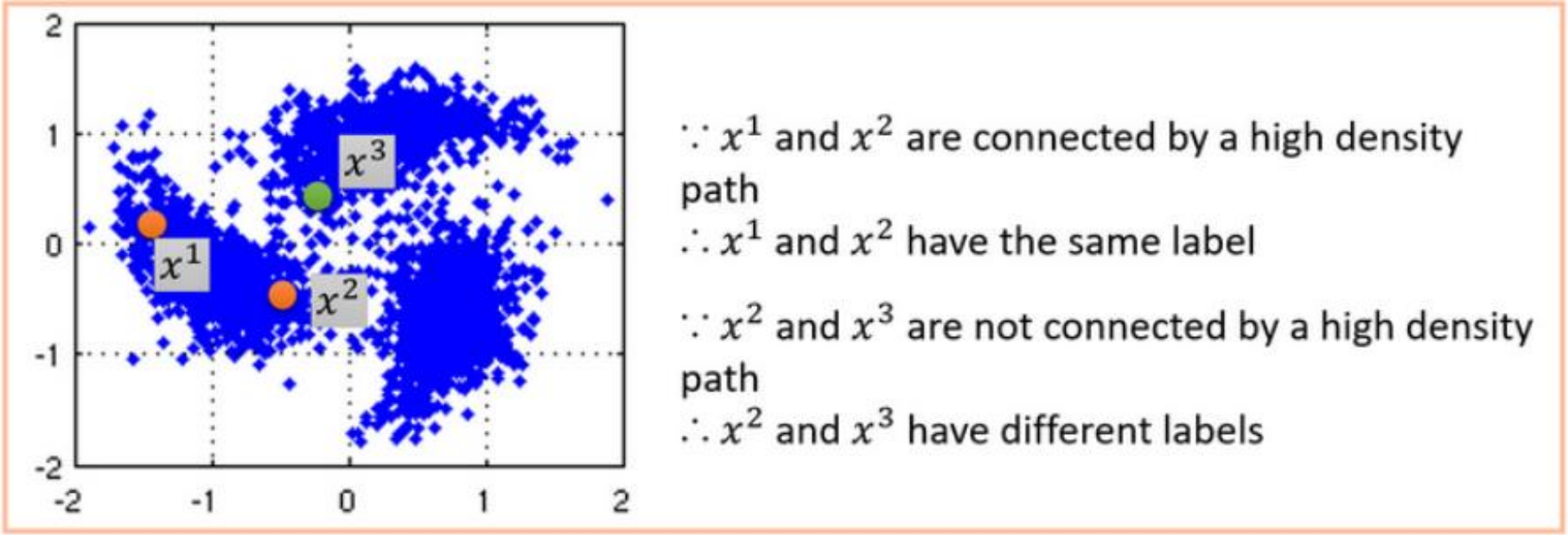
Smoothness Assumption

- 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

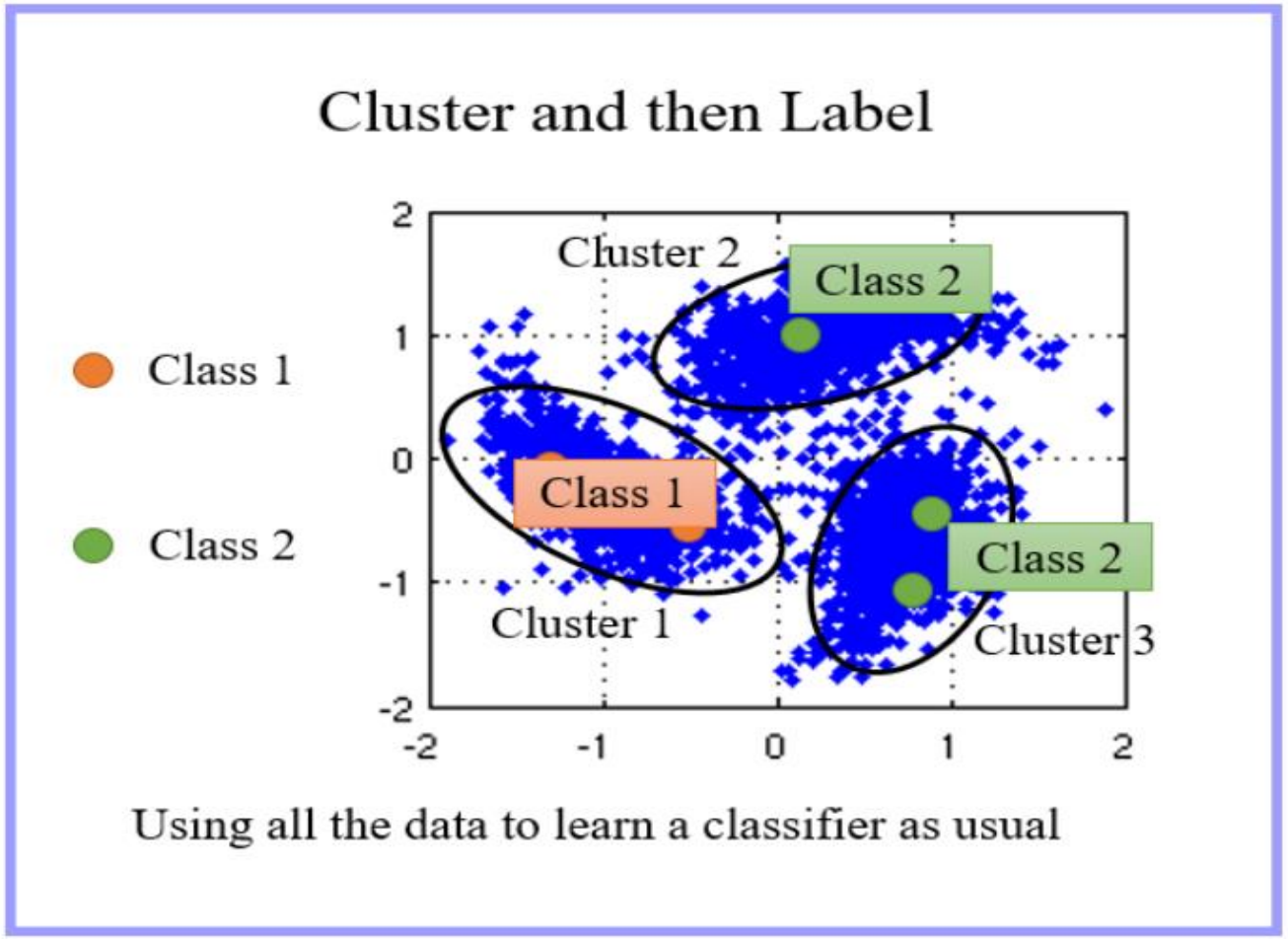
算法原理

- 先聚类后标注算法(Cluster and then Label)



算法原理

- 先聚类后标注算法(Cluster and then Label)



算法原理

- 基于图的方法(Graph-Based Approach)

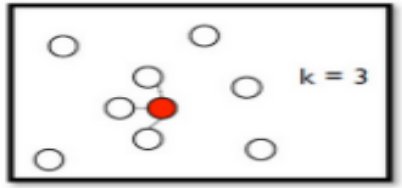
Graph Construction

- Define the similarity $s(x^i, x^j)$ between x^i and x^j

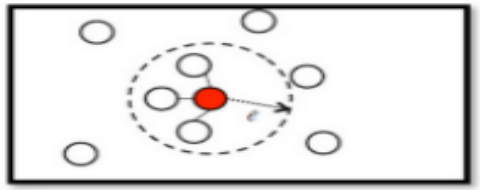
$$s(x^i, x^j) = exp\left(-\gamma \|x^i - x^j\|^2\right)$$ → Gaussian Radial Basis Function

- Add edge:

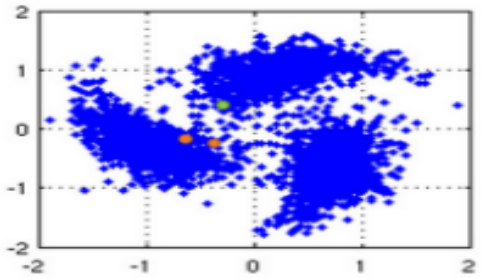
K Nearest Neighbor



e-Neighborhood



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



算法原理

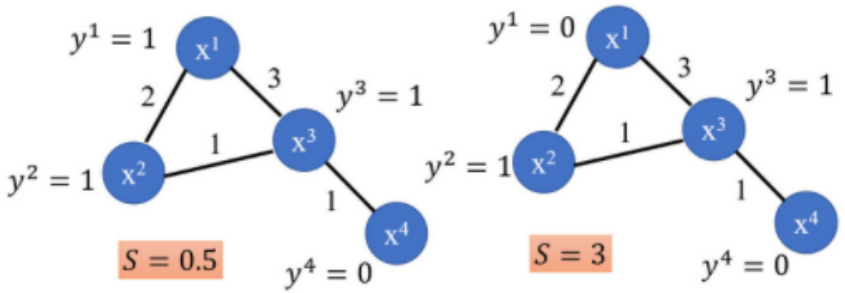
- 基于图的方法(Graph-Based Approach)

Graph-based Approach

Define the smoothness of the labels on the graph

$$S = \frac{1}{2}\sum_{i,j} w_{i,j}(y^i - y^j)^2$$ ← Smaller means smoother

For all data (no matter labelled or not)



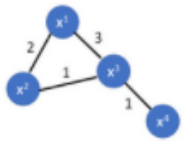
$$S = \frac{1}{2}\sum_{i,j} w_{i,j}(y^i - y^j)^2 = \boldsymbol{y}^T L \boldsymbol{y}$$ ← Depending on model parameters

y: (R+U)-dim vector

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

L: (R+U) × (R+U) matrix — Graph Laplacian

$$L = D - W$$



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

$$L = \sum_{x^r} L(y^r, \hat{y}^r) \quad +\lambda S$$ ← As a regularization term

应用总结

应用总结

- 应用领域
  - 语音识别(Speech Recognition)
  - 文本分类(Text categorization)
  - 语义解析(Parsing)
  - 视频监控(Video surveillance)
  - 蛋白质结构预测(Protein structure prediction)
- 待研究的问题
  - 无标签样本的有效利用问题
  - 大量无标签样本的高效使用问题
  - 特征选择中的有效性问题

参考文献

[1] Samuli Laine, Timo Aila. TEMPORAL ENSEMBLING FOR SEMI-SUPERVISED LEARNING[C].ICLR, 2017.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow. MixMatch: A Holistic Approach to Semi-Supervised Learning[C]. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

[3] Zhi-Hua Zhou. A brief introduction to weakly supervised learning[C]. Advance access publication 25 August 2017.

[4] Rodrigo Benenson, Stefan Popov. Large-scale interactive object segmentation with human annotators[C]// Conference on Computer Vision and Pattern Recognition (CVPR),2019.

[5] Qizhe Xie, , Zihang Dai. Unsupervised Data Augmentation for Consistency Training[C]. 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

知人者智，自知者明。
胜人者有力，自胜者
强。知足者富。强行
者有志。不失其所者
久。死而不亡者，寿。

谢谢！