CS540 Introduction to Artificial Intelligence Lecture 11

Young Wu
Based on lecture slides by Jerry Zhu and Yingyu Liang

June 24, 2019

Supervised Learning

Review

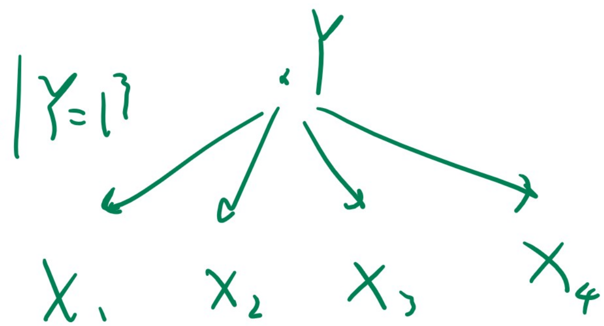
- Given training data and label.
- Discriminative: estimate $\hat{\mathbb{P}}\{Y = y | X = x\}$ to classify.
- Generative: estimate $\hat{\mathbb{P}}\{X = x | Y = y\}$ and Bayes rule to classify.

Naive Bayes

Review

- Naive Bayes: $X_j \leftarrow Y$.

$\hat{\Pr}\{X_1 = x_1 | Y = 1\}$



$X_1 | Y$

$\Pr\{Y = 0 |$ ... $\}$

$$\mathbb{P}\{Y = 1 | X_1 = x_1, ..., X_m = x_m\}$$

$$= \frac{\hat{\mathbb{P}}\{Y = 1\} \prod_{j=1}^{m} \hat{\mathbb{P}}\{X_j = x_j | Y = 1\}}{\mathbb{P}\{X_1 = x_1, ..., X_m = x_m\}} = \Pr\{Y=1\}\Pr\{X = x | Y = 1\} + \Pr\{Y = 0\}\Pr\{X = x | Y = 0\}$$

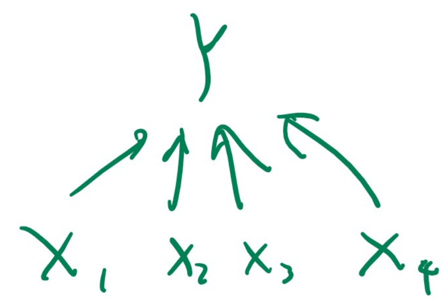
$$= \frac{1}{1 + \exp\left(\underbrace{-\log\left(\frac{\mathbb{P}\{Y = 1\}}{\mathbb{P}\{Y = 0\}}\right)}_{b} - \underbrace{\sum_{j=1}^{m} \log\left(\frac{\mathbb{P}\{X_j = x_j | Y = 1\}}{\mathbb{P}\{X_j = x_j | Y = 0\}}\right)}_{w^T x}\right)}$$

Logistic Regression

Review

$$\frac{1}{1+\exp\left(-\underbrace{\log\left(\frac{\hat{\mathbb{P}}\{Y=1\}}{\hat{\mathbb{P}}\{Y=0\}}\right)}_{b}-\sum_{j=1}^{m}\underbrace{\log\left(\frac{\hat{\mathbb{P}}\{X_j=x_j|Y=1\}}{\hat{\mathbb{P}}\{X_j=x_j|Y=0\}}\right)}_{\vec{w}^T x}\right)}$$

- Logistic Regression: $X_j \to Y$.



$$\tilde{\mathbb{P}}\{Y=1|X_1=x_1,\ldots,X_m=x_m\}$$

$a \in [0,1]$

$$=\frac{1}{1+\exp\left(-\left(b+\sum_{j=1}^{m} w_j x_j\right)\right)}$$

Naive Bayes v Logistic Regression Derivation

Review

Generative Adversial Network

Review

- Generative Adversial Network (GAN): two competitive neural networks.

1. Generative network input random noise and output fake images.
2. Discriminative network input real and fake images and output label real or fake.

Generative Adversial Network Diagram

Review

Midterm

Admin

$\boxed{W1 \rightarrow W5}$ Supervised.

- Materials: END HERE
- Calculator: pay 2 points out of 40.
- Formula sheet: will post
- Additional formula sheet: 2 points each
- NO examples, quiz questions, homework questions: 2 points each

request room

Unsupervised Learning

Motivation

- Supervised learning: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.
- Unsupervised learning: $x_1, x_2, ..., x_n$.
- There are a few common tasks without labels.

1. Clustering: separate instances into groups.
2. Novelty (outlier) detection: find instances that are different.
3. Dimensionality reduction: represent each instance with a lower dimensional feature vector while maintaining key characteristics.

Unsupervised Learning Applications

Motivation

1. Google News
2. Google Photo
3. Image Segmentation
4. Text Processing

group similar words into one type.

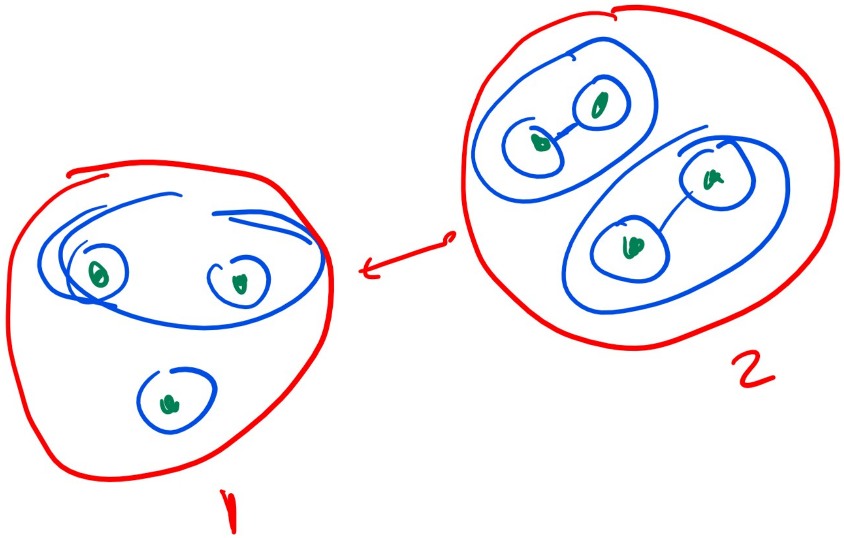
Hierarchical Clustering

Description

- Start with each instance as a cluster.
- Merge clusters that are closest to each other.
- Result in a binary tree with close clusters as children.

Hierarchical Clustering Diagram

Description



7 clusters

6

2

Clusters

Definition

- A cluster is a set of instances.

$$C_k \subseteq \{x_i\}_{i=1}^{n}$$

- A clustering is a partition of the set of instances into clusters.

$$C = C_1, C_2, ..., C_K$$

$$C_k \cap C_{k'} = \emptyset \; \forall \; k' \neq k, \bigcup_{k=1}^{K} C_k = \{x_i\}_{i=1}^{n}$$

Distance between Points

Definition

- Usually, the distance between two instances is measured by the Euclidean distance or L_2 distance.

$$\rho\left(x_i, x_{i'}\right) = \left\|x_i - x_{i'}\right\|_2 = \sqrt{\sum_{j=1}^{m}\left(x_{ij} - x_{i'j}\right)^2}$$

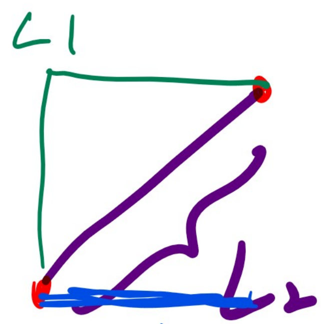
- Other examples include: L_1 distance and L_∞ distance.

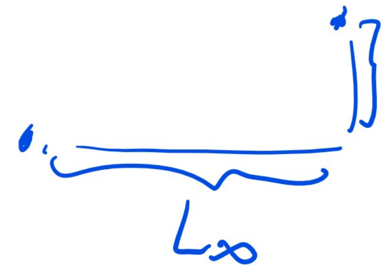
$$\rho_1\left(x_i, x_{i'}\right) = \left\|x_i - x_{i'}\right\|_1 = \sum_{j=1}^{m}\left|x_{ij} - x_{i'j}\right|$$

$$\rho_\infty\left(x_i, x_{i'}\right) = \left\|x_i - x_{i'}\right\|_\infty = \max_{j=1,2,\ldots,m}\left\{\left|x_{ij} - x_{i'j}\right|\right\}$$

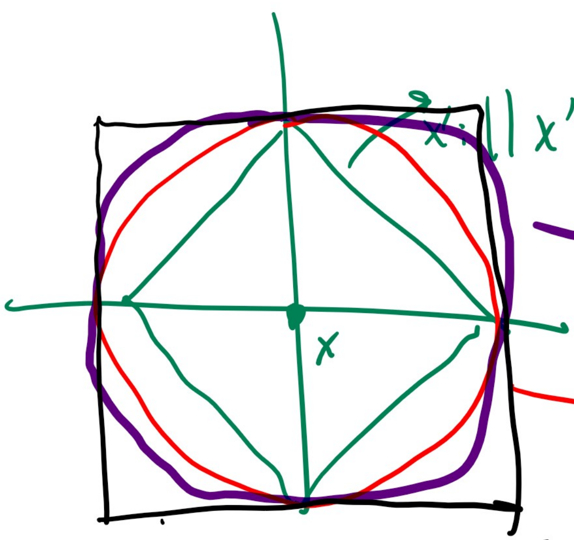
L_p Distance Diagram

Definition





$\|x' - x\|_p = 1 \quad p = 1, 2, \ldots, \infty$



$x' : \|x' - x\|_1 = 1$

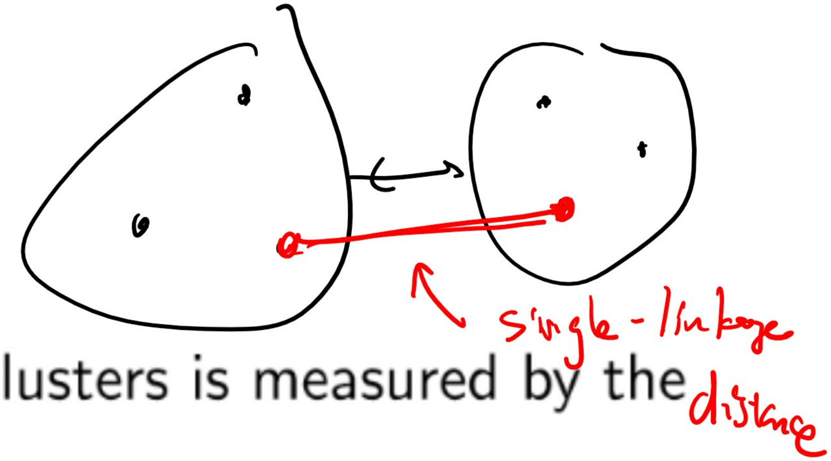
$x' : \|x' - x\|_3 = 1$

$x' : \|x' - x\|_2 = 1$

$x' : \|x' - x\|_\infty = 1$

Single Linkage Distance

Definition



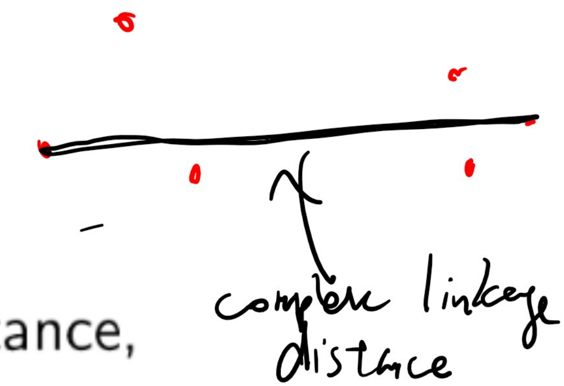
- Usually, the distance between two clusters is measured by the single-linkage distance.

$$\rho\left(C_k, C_{k'}\right) = \min\left\{\rho\left(x_i, x_{i'}\right) : x_i \in C_k, x_{i'} \in C_{k'}\right\}$$

- It is the shortest distance from any instance in one cluster to any instance in the other cluster.

Complete Linkage Distance

Definition



- Another measure is complete-linkage distance,

$$\rho\left(C_k, C_{k'}\right) = \max\left\{\rho\left(x_i, x_{i'}\right) : x_i \in C_k, x_{i'} \in C_{k'}\right\}$$

- It is the longest distance from any instance in one cluster to any instance in the other cluster.

Average Linkage Distance Diagram

Definition

- Another measure is average-linkage distance.

$$\rho\left(C_k, C_{k'}\right) = \frac{1}{\left|C_k\right| \left|C_{k'}\right|} \sum_{x_i \in C_k, x_{i'} \in C_{k'}} \rho\left(x_i, x_{i'}\right)$$

- It is the average distance from any instance in one cluster to any instance in the other cluster.

Hierarchical Clustering Example 1 Part I

Quiz (Graded)

$$\sqrt{(x_1 - x_2)^2} = |x_1 - x_2|$$

Q2

- Spring 2018 Midterm Q5
- Given three clusters $A = \{0, 2, 6\}$, $B = \{3, 9\}$, $C = \{11\}$. What is the next iteration of hierarchical clustering with Euclidean distance and single linkage?
- A: Merge A and B.
- B: Merge A and C.
- C: Merge B and C.
- D: No change, E: Do not choose.

5

1

2

CS540S1

min distance between points in cluster

Hierarchical Clustering Example 1 Part II

Quiz (Graded)

Q4

- Spring 2018 Midterm Q5
- Given three clusters $A = \{0, 2, 6\}$, $B = \{3, 9\}$, $C = \{11\}$. What is the next iteration of hierarchical clustering with Euclidean distance and complete linkage?
- A: Merge A and B.
- B: Merge A and C.
- C: Merge B and C.
- D: No change, E: Do not choose.

11

9

8

max to find dist

merge min dist.

Hierarachical Clustering Example 2

Quiz (Participation)

- Spring 2017 Midterm Q4
- Given the distance between the clusters so far. Which pair (choose 2) of clusters will be merged using single linkage.

–	A	B	C	D	E
A	0	1075	2013	2054	996
B	1075	0	3272	2687	2037
C	2013	3272	0	808	1307
D	2054	2687	808	0	1059

merge CD

	A	B	CD	E
A	0	1075	2013	996
B	1075	0	2687	2037
CD	2013	2687	0	1059

Hierarchical Clustering

Algorithm

- Input: instances: $\{x_i\}_{i=1}^n$, the number of clusters K, and a distance function ρ.
- Output: a list of clusters $C = C_1, C_2, ..., C_K$.
- Initialize for $t = 0$.

$$C^{(0)} = C_1^{(0)}, ..., C_n^{(0)}, \text{ where } C_k^{(0)} = \{x_k\}, k = 1, 2, ..., n$$

- Loop for $t = 1, 2, ..., n - k + 1$.

$$(k_1^\star, k_2^\star) = \arg\min_{k_1, k_2} \rho\left(C_{k_1}^{(t-1)}, C_{k_2}^{(t-1)}\right)$$

$$C^{(t)} = \left(C_{k_1^\star}^{(t-1)} \cup C_{k_2^\star}^{(t-1)}\right), C_1^{(t-1)}, \text{... no } k_1^\star, k_2^\star ..., C_n^{(t-1)}$$

K Means Clustering

Description

- This is not K Nearest Neighbor.
- Start with random cluster centers.
- Assign each point to its closest center.
- Update all cluster centers as the center of its points.

K Means Clustering Diagram

Description

Center

Definition

- The center is the average of the instances in the cluster,

$$c_k = \frac{1}{|C_k|} \sum_{x \in C_k} x$$

Distortion

Distortion

- Distortion for a point is the distance from the point to its cluster center.
- Total distortion is the sum of distortion for all points.

all training data

cluster center

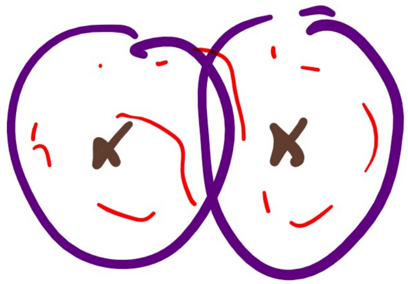
$$D_K = \sum_{i=1}^{n} \rho\left(x_i, \boxed{c_{k^\star(x_i)}\left(\cancel{x_i}\right)}\right)$$

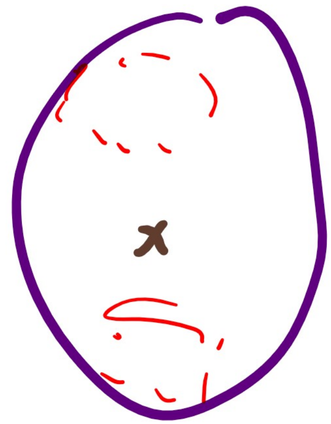
cluster x_i is assigned to

$$k^\star(x) = \arg \min_{k=1,2,\ldots K} \rho\left(x, c_k\right)$$

Objective Function

Definition





k means converge to local min

- This algorithm stop in finite steps.
- This algorithm is trying to minimize the total distortion but fails.

Objective Function Counterexample

Definition

Gradient Descent

Definition

- When ρ is the Euclidean distance. K Means algorithm is the gradient descent when distortion is the objective (cost) function.

$$\frac{\partial}{\partial c_k} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - c_k\|_2^2 = 0$$

gradient descent for total distortion as cost

$$\Rightarrow -2 \sum_{x \in C_k} (x - c_k) = 0$$

$$\Rightarrow \boxed{c_k = \frac{1}{|C_k|} \sum_{x \in C_k} x}$$

same as k means

Gradient Descent Derivation

Derivation

K Means Clustering Example Part I

Quiz (Graded)

Q5

- Spring 2018 Midterm Q5
- Given data $\{5, 7, 10, 12\}$ and initial cluster centers $c_1 = 3, c_2 = 13$, what is the initial clusters?
- A: $\{5, 7\}$ and $\{10, 12\}$
- B: $\{5\}$ and $\{7, 10, 12\}$
- C: $\{5, 7, 10\}$ and $\{12\}$
- D: none of the above, E: do not choose.

$\{5, 7\}$, $\{10, 12\}$

1 2

$c_1 = 6$, $11 = c_2$

K Means Clustering Example Part II

Quiz (Graded)

- Spring 2018 Midterm Q5
- Given data $\{5, 7, 10, 12\}$ and initial cluster centers $c_1 = 3, c_2 = 13$, what are the cluster in the next iteration?
- A: $\{5, 7\}$ and $\{10, 12\}$ 6. 11
- B: $\{5\}$ and $\{7, 10, 12\}$
- C: $\{5, 7, 10\}$ and $\{12\}$
- D: none of the above, E: do not choose.

K Means Clustering

Algorithm

- Input: instances: $\{x_i\}_{i=1}^n$, the number of clusters K, and a distance function ρ.
- Output: a list of clusters $C = C_1, C_2, ..., C_K$.
- Initialize $t = 0$.

$$c_k^{(0)} = K \text{ random points}$$

- Loop until $c^{(t)} = c^{(t-1)}$.

$$C_k^{(t-1)} = \left\{ x : k = \arg \min_{k' \in 1,2,\ldots,K} \rho\left(x, c_k^{(t-1)}\right) \right\} \rightarrow \text{label}$$

$$c_k^{(t)} = \frac{1}{\left|C_k^{(t-1)}\right|} \sum_{x \in C_k^{(t-1)}} x \qquad \rightarrow \text{recenter}$$

Number of Clusters

Discussion

- There are a few ways to pick the number of clusters K.

1. K can be chosen using prior knowledge about X.
2. ~~K can be the one that minimizes distortion?~~ No, when $K = n$, distortion $= 0$.
3. K can be the one that minimizes distortion + regularizer.

dimension of each x

$$K^{\star} = \arg\min_{k} \left(D_k + \lambda \cdot m \cdot k \cdot \log n \right)$$

# of x instances

large cost for large # of cluster k

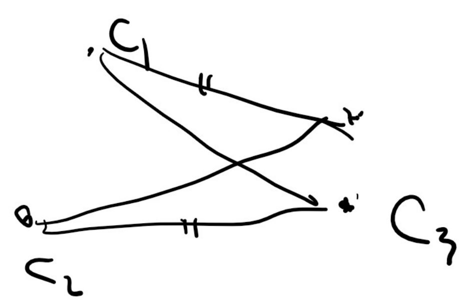
- λ is a fixed constant chosen arbitrarily.

Initial Clusters

Discussion

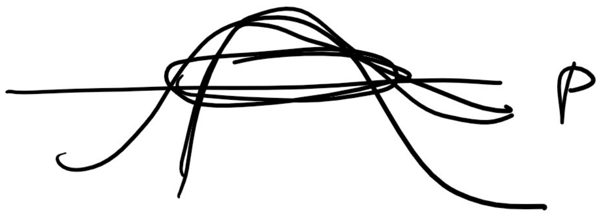
- There are a few ways to initialize the clusters.

1. K uniform random points in $\{x_i\}_{i=1}^n$.
2. 1 uniform random point in $\{x_i\}_{i=1}^n$ as $c_1^{(0)}$, then find the farthest point in $\{x_i\}_{i=1}^n$ from $c_1^{(0)}$ as $c_2^{(0)}$, and find the farthest point in $\{x_i\}_{i=1}^n$ from the closer of $c_1^{(0)}$ and $c_2^{(0)}$ as $c_3^{(0)}$, and repeat this K times.

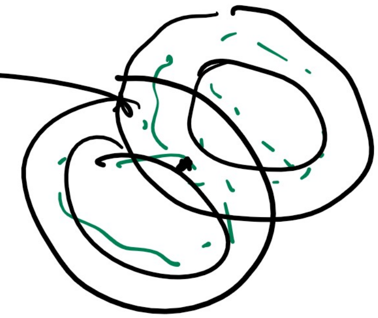


Gaussian Mixture Model

Discussion



prob belong to C_1 and C_2



- In K means, each instance belong to one cluster with certainty.
- One continuous version is called the Gaussian mixture model: each instance belongs to one of the clusters with a positive probability.
- The model can be trained using Expectation Maximization Algorithm (EM Algorithm).

EM Algorithm, Part I

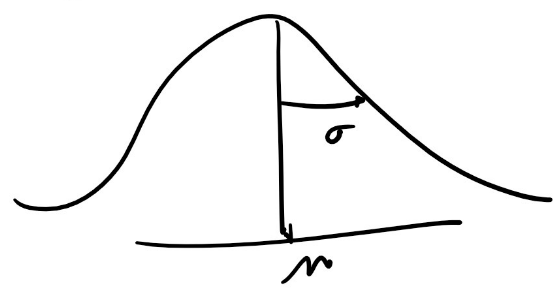
Discussion

- The means μ_k and variances σ_k^2 for each cluster need to be trained. The mixing probability π_k also needs to be trained.

$$\left(\mu_1, \sigma_1^2, \pi_1\right), \left(\mu_2, \sigma_2^2, \pi_2\right), ..., \left(\mu_K, \sigma_K^2, \pi_K\right)$$

σ^2

- Initialize by random guesses of clusters means and variances.



EM Algorithm, Part II

Discussion

- Expectation Step. Compute responsibilities for $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, K$.

prôb of point i assigned to cluster k.

$$\hat{\gamma}_{i,k} = \frac{\hat{\pi}_k \phi_k(x_i)}{\sum_{k'=1,2,\ldots,K} \hat{\pi}_{k'} \phi_{k'}(x_i)}$$

0, 1

$$\phi_k(x) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_k} \exp\left(-\frac{(x-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2}\right)$$

EM Algorithm, Part III

Discussion

- Maximization Step. Compute means and variances for each $k = 1, 2, ..., K$.

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} \hat{\gamma}_{i,k} x_i}{\sum_{i=1}^{n} \hat{\gamma}_i}, \text{ and } \hat{\sigma}_k^2 = \frac{\sum_{i=1}^{n} \hat{\gamma}_{i,k} (x_i - \hat{\mu}_k)^2}{\sum_{i=1}^{n} \hat{\gamma}_i}$$

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{\gamma}_{i,k}$$

- Repeat until convergent.

$$\mu_1 = \frac{(0.9 \cdot x_1 + 0.9 x_1 + 0.9 x_3 + 0.1 x_4)}{0.9 + 0.9 + 0.9 + 0.1}$$

$$C_1 = \frac{X_1 + X_2 + X_3}{3}$$

Gaussian Mixture Model Diagram

Discussion