CS540 Introduction to Artificial Intelligence
Lecture 11

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

June 27, 2019

Supervised Learning

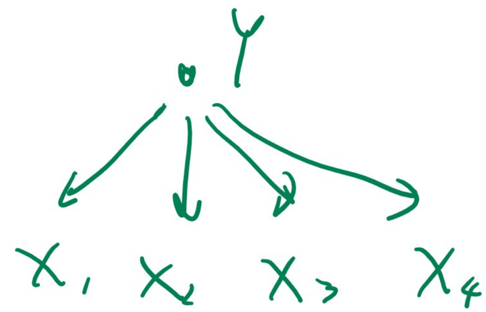
Review

- 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: $X_j \leftarrow Y$.



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

$$= \frac{\mathbb{P}\left\{Y = 1\right\} \prod_{j=1}^{m} \mathbb{P}\left\{X_j = x_j | Y = 1\right\}}{\mathbb{P}\left\{X_1 = x_1, ..., X_m = x_m\right\}}$$

CPT

$\Pr\{Y, X\}$

$\sim \Pr\{Y=1\} \cdot \Pr\{X = x | Y=1\} + \Pr\{Y=0\} \cdot \Pr\{X=x|$

$Y=0\}$

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

Logistic Regression

Review

$$\dfrac{1}{1+\exp\left(-\log\left(\dfrac{\mathbb{P}\{Y=1\}}{\mathbb{P}\{Y=0\}}\right)-\displaystyle\sum_{j=1}^{m}\log\left(\dfrac{\mathbb{P}\{X_j=x_j|Y=1\}}{\mathbb{P}\{X_j=x_j|Y=0\}}\right)\right)}$$

Naive Bayes

- Logistic Regression: $X_j \to Y$.

$a_i = \Pr\{Y=1 | x\}$

$$\tilde{\mathbb{P}}\{Y=1|X_1=x_1,...,X_m=x_m\} = \dfrac{1}{1+\exp\left(-\left(b+\displaystyle\sum_{j=1}^{m}w_jx_j\right)\right)}$$

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

Cumulative Distribution Inversion Method, Part I

Discussion

- Most programming languages have a function to generate a random number $u \sim$ Unif $[0, 1]$.
- If there are $m = 2$ tokens in total and the conditional probabilities are p and $1 - p$. Then the following distributions are the same.

$$z_N = \begin{cases} 0 & \text{with probability } p \\ 1 & \text{with probability } 1 - p \end{cases} \Leftrightarrow z_N = \begin{cases} 0 & \text{if } 0 \leqslant u \leqslant p \\ 1 & \text{if } p < u \leqslant 1 \end{cases}$$

Cumulative Distribution Inversion Method, Part II

Discussion

- In the general case with m tokens with conditional probabilities $p_1, p_2, ..., p_m$ with $\sum_{j=1}^{m} p_j = 1$. Then the following distributions are the same.

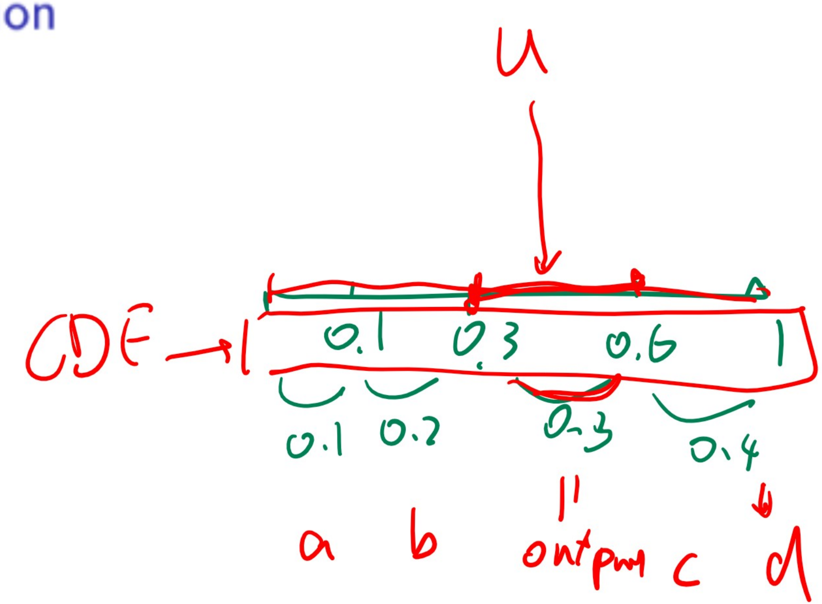
$$z_N = j \text{ with probability } p_j \Leftrightarrow z_N = j \text{ if } \sum_{j'=1}^{j-1} p_{j'} < u \leqslant \sum_{j'=1}^{j} p_{j'}$$

- This can be used to generate a random token from the conditional distribution.

CDF Inversion Method Diagram

Discussion

$$\hat{P}_r\{a, \underline{a}\} = 0.1$$
$$\{b, a\} = 0.2$$
$$\{c, a\} \quad 0.3$$
$$\{d, a\} \quad 0.4$$



Unsupervised Learning

Motivation

instances labels

- Supervised learning: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- Unsupervised learning: $\underbrace{x_1, x_2, \ldots, x_n}$.
- 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.

not feature selection, combining features to make new one.

Unsupervised Learning Applications

Motivation

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

trigram for words,
bigram

group similar tokens
into one cluster
one type

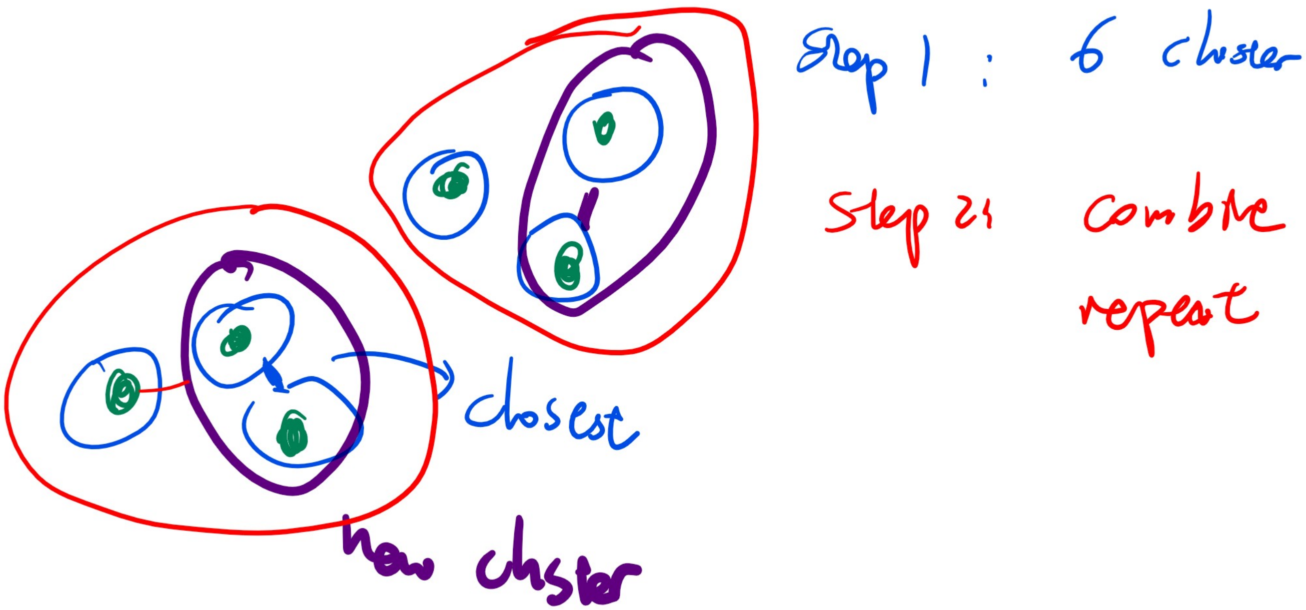
Hierarchical Clustering

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



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.

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

nothing in common

adding up all cluster

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

Distance between Points

Definition

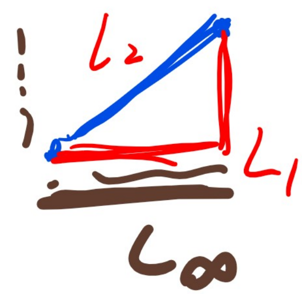
- Usually, the distance between two instances is measured by the Euclidean distance or L_2 distance.

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

- 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

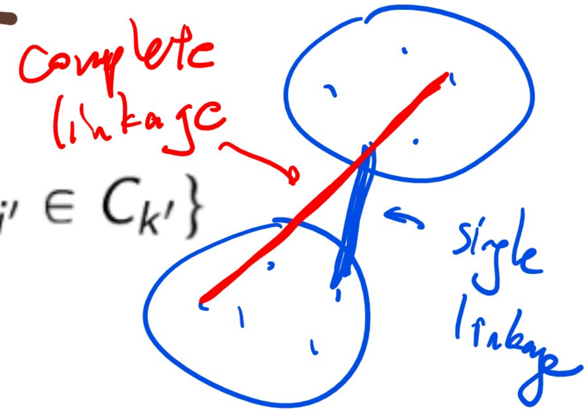
Definition

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)

Q1

- 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.

$d(A, B) = 1$

$d(B, C) = 2$

$d(A, C) = 5$

min distance among all pairs of points

distance between cluster

merge closest clusters

Hierarchical Clustering Example 1 Part II

Quiz (Graded)

Q3

$d(A, B) = 9$

$d(A, C) = 11$, $d(B, C) = 8$

- 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?

max distance

- A: Merge A and B.
- B: Merge A and C.
- C: Merge B and C.
- D: No change, E: Do not choose.

merge clusters with min distance

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.

CD

–	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
E	996	2037	~~[illegible]~~	1059	0

d(A, CD)

⇓

min{d(A, C), d(A, D)}

CD

A

d(C, D) is smallest

merge C, D

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

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

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

Number of Clusters

Discussion

- K can be chosen using prior knowledge about X.
- The algorithm can stop merging as soon as all the between-cluster distances are larger than some fixed R.
- The binary tree generated in the process is often called dendrogram, or taxonomy, or a hierarchy of data points.
- An example of a dendrogram is the tree of life in biology.

K Means Clustering

Description

back
6:32

- 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,

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

$$m \quad \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{im} \end{pmatrix} + \begin{pmatrix} x_{i'1} \\ x_{i'2} \\ \vdots \\ x_{i'm} \end{pmatrix}$$

Distortion

Distortion

≈ Cost

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

instance

cluster center this instance belongs to

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

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

Objective Function

Definition

$$\min_{C_k} \ \text{Cost} = \text{Distortion}$$

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

local minimums.

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.

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

gradient of distortion w.r.t. c_k

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

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

gradient descent step

Gradient Descent Derivation

Derivation

K Means Clustering Example Part I

Quiz (Graded)

Q4

- 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}

$c_1 = 6$ $c_2 = 11$

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 $\{10, 12\}$
- 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 $c^{(t)} = c^{(t-1)}$.

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

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

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.

dim of x_i, # feature

# of instances

K = ↓ distortion

$$K^{\star} = \arg\min_{k} \left(D_k + \underbrace{\lambda \cdot m \cdot \underline{k} \cdot \log n}_{\text{cost of adding clusters}} \right)$$

- λ is a fixed constant chosen arbitrarily.

Initial Clusters

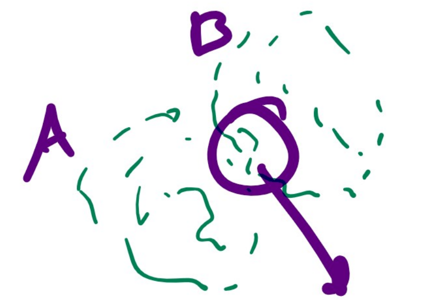
Discussion

- 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



$\frac{1}{2}$ A
$\frac{1}{2}$ B

- 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

k mean

center C_1 C_2 C_3

- 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.

size of cluster

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

centers

EM

- Initialize by random guesses of clusters means and variances.

EM Algorithm, Part II

Discussion

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

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

prob of instance i belongs to cluster k.

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

EM Algorithm, Part III

Discussion

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

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

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

- Repeat until convergent.

Gaussian Mixture Model Diagram

Discussion

7:15