Homework 2
ECE/CS 498 DS Spring 2020
Issued: 02/24/20
Due: 03/02/20

Name: GOWTHAM KUNTUMALLA
NetID: GOWTHAM 4

Note: Only Problem 1,3 will be graded. Problem 2 is optional.

You can either type your solutions or write your solutions by hand. You should make sure your manuscript is legible if you choose to write your solutions.

Problem 1

Task 1 K-Means:

Suppose there are 6 data points in 2D space. The coordinates of these data points are given in the following table:

$\vec{x}_1$	$\vec{x}_2$	$\vec{x}_3$	$\vec{x}_4$	$\vec{x}_5$	$\vec{x}_6$
(1,1)	(1,2)	(2,2)	(6,6)	(6,7)	(7,7)

Our task is to cluster them using K-means algorithm. **We set K=2 and use Euclidean distance.** In the beginning, the centroids are initialized to (3,2), (5,7).

1. Fill in the tables below to complete the first iteration:
   Step 1: Assign data points to the nearest centroid. Fill each data point in one of the columns below.

	Centroid (3,2)	Centroid (5,7)
Assigned Data points	(1,1) (1,2) (2,2)	(6,6), (6,7) (7,7)

$S(x_1, x_2, x_3)$ $S(x_4, x_5, x_6)$

Step 2: Update centroid. Calculate new centroids base on the data assignment in Step 1.

New Centroid I	New Centroid II
$\left(\frac{4}{3}, \frac{5}{3}\right)$	$\left(\frac{19}{3}, \frac{20}{3}\right)$

2. In addition, state at least two convergence criteria for K-Means, choose one criterion, and judge whether the K-Means algorithm has converged after the above step.

1) Convergence (in terms of assignments) — clusters remain unchanged
2) Number of iterations = 1

If we choose this, then [Yes] it has converged

Grade:_____

Homework 2
ECE/CS 498 DS Spring 2020
Issued: 02/24/20
Due: 03/02/20

Name: _____

NetID: GOWTHAM4

3. Suppose you have a dataset which has some outliers that will affect the K-means' clustering result if included. To achieve a better clustering result, you must reduce the effects of these outliers on the clusters. However, suppose you don't want to remove these outliers from the dataset when doing clustering, but you only want to make one subtle change to the K-Means algorithm. **Describe how you can achieve this, and why**. Hint: the new algorithm has the name beginning with K. Hint, Hint: You may want to update centroid by using other statistics instead of mean.

① K-Medoids algorithm. It minimises Pairwise dissimilarities [as opposed to sum of Euclidean distances in K-means algo]

② (or) K-Medians algorithm. Median isn't affected by outliers.

Task 2 1-D GMM:

Consider applying EM to train a Gaussian Mixture Model (GMM) to cluster the data in the above task into two clusters. First, we want to apply 1-D GMM. Therefore, we project these data to the horizontal axis by ignoring the second dimension. This leads **to 6 points at 1,1,2,6,6,7**. The initial Gaussian Components are $a \sim N(3,1), b \sim N(4,1)$.

1.For the point x=1, calculate $P(x = 1|a)$, which is the probability density of observing 1 when sampled from distribution $a \sim N(3,1)$. Show your work. $\mu_a = 3, \sigma_a^2 = 1$; $\mu_b = 4, \sigma_b^2 = 1$

$$P(x_i|a) = \frac{1}{\sqrt{2\pi\sigma_a^2}} \exp\left(-\frac{(x_i-\mu_a)^2}{2\sigma_a^2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_i-3}{1}\right)^2\right); \quad P(x=1|a) = \frac{1}{\sqrt{2\pi}} \exp(-2)$$

$P(x=1|a) = 0.054$

2. Calculate $P(a|x = 1)$, which is the posterior probability of distribution $\boldsymbol{a}$ given sample x=1. Show your work. Assume that prior $P(a) = P(b) = 0.5$.

similarly, $P(x=1|b) = 0.0044$

$$P(a|x=1) = \frac{P(x=1|a)\,P(a)}{P(x=1|a)P(a)+P(x=1|b)P(b)} = \frac{P(x=1|a)}{P(x=1|a)+P(x=1|b)} = 0.924$$

3. Follow your steps in the above question, calculate posterior $P(a|x_i)$ and $P(b|x_i)$ for all the data points. You might want to write a Python program to solve this question.

data point x_i	Posterior $P(a\,\|\,x_i)$	Posterior $P(b\,\|\,x_i)$
1	0.924	0.076
1	0.924	0.076
2	0.818	0.182
6	0.076	0.924
6	0.076	0.924
7	0.024	0.971

Probably 'a'

Probably 'b'

Grade: _____

Homework 2
ECE/CS 498 DS Spring 2020
Issued: 02/24/20
Due: 03/02/20

Name: _____

NetID: Gowtham4

4. Based on your results in the previous questions, calculate new means and variances for the two new Gaussian Components.

new 'a'

$$\mu_a = \frac{\sum P(a|x_i)\,P(x_i)}{\sum P(a|x_i)}$$

$$\sigma_a^2 = \frac{\sum P(a|x_i)\cdot(x_i-\mu_a)^2}{\sum P(a|x_i)}$$

$\mu_a' = 1.615$

$\sigma_a^{2\prime} = 1.61$

new 'b'

similar formulas as 'a'

(used a python script to simplify calculations)

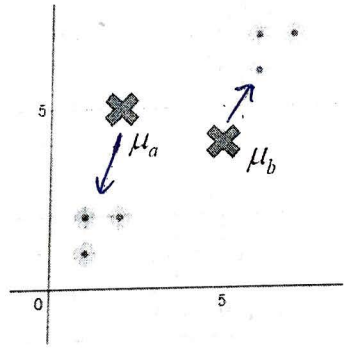
$\mu_b' = 5.83$

$\sigma_b^{2\prime} = 2.41$

Task 3 2-D GMM:

Now consider clustering the original data given in Task 2 using a 2-D GMM. We assign two initial Gaussian components to be $a \sim N((2,5), I)$, $b \sim N((5,4), I)$, where I means Identity matrix, which implies that the covariance between features is 0 and the variance of each feature is 1.

1.Draw on the figure the directions in which μ_a and μ_b will move in the next iteration. You can choose to copy or redraw the figure in your solution.



arrows are shown on the picture

---

work for Prob. 2.

$|\Sigma| = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1$; $(x_i - \mu_a) = (2,2) - (2,5) = (0,-3)$

also, $(x-\mu_a)^T \Sigma^{-1} (x-\mu_a) = \begin{bmatrix} 0 & -3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ -3 \end{bmatrix} = \begin{bmatrix} 0 & -3 \end{bmatrix} \begin{bmatrix} 0 \\ -3 \end{bmatrix} = 9$

2. For data point (2,2), calculate $P((2,2)|a)$, which is the probability density of observing (2,2) when sampled from distribution multivariate Gaussian distribution a. Show your work.

$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$\Sigma^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$$P(x_i|\mu_a, \Sigma_a) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_a)}} \exp\left(-\frac{1}{2} (x-\mu_a)^T \Sigma_a^{-1} (x-\mu_a) \right)$$

$$\Rightarrow P(x|\Sigma,\mu) = \frac{1}{\sqrt{(2\pi)^2}} \exp\left(-\frac{1}{2} \times 9\right) = \frac{1}{2\pi} \exp(-4.5)$$

$\Rightarrow$ **$P(2,2|a) = 0.00177$**

Grade: _____

Homework 2
ECE/CS 498 DS Spring 2020
Issued: 02/24/20
Due: 03/02/20

Name: _____

NetID: Gowthamy

3. Calculate $P(a|(2,2))$, which is the posterior probability of distribution a given sample (2,2). Show your work, and assume $P(a) = P(b) = 0.5$.

$P(a|x_i) = \frac{P(x_i|a)P(a)}{P(x_i|b)P(b)+P(x_i|a)P(a)} = \frac{P(x_i|a)}{P(x_i|a)+P(x_i|b)} = \frac{0.00177}{0.00177+0.00024}$

$P(x_i|b) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\begin{bmatrix}3 & 2\end{bmatrix}\begin{bmatrix}-3\\-2\end{bmatrix}\right)$

$= \frac{1}{2\pi}\exp\left(-\frac{1}{2}\times(9+4)\right)$

$= \frac{1}{2\pi}\exp(-6.5) = 0.00024$

$P(a|x_i) = 0.88$

4. Follow your steps in the above question, calculate posterior $P(a|x_i)$ and $P(b|x_i)$ for all the data points. You might want to write a Python program for this question.

soft!

data point x_i	Posterior $P(a\,|\,x_i)$	Posterior $P(b\,|\,x_i)$	
(1,1)	0.982	0.018	a>b
(1,2)	0.993	0.007	a>b
(2,2)	0.88	0.12	a>b
(6,6)	0.0025	0.9975	b>a
(6,7)	0.0066	0.9933	b>a
(7,7)	0.000335	0.99966	b>a

soft!

* wrote a python script for these calculations

5. Based on your result in the previous questions, calculate new means and covariance matrices for two new Gaussian Components.

Features
0 → x-axis
1 → y-axis

New 'a'

$\mu_{a,0} = \sum_{i=1}^{N} \frac{P(a|\vec{x_i})\,x_{i,0}}{\sum_{i=1}^{N} P(a|\vec{x_i})}$

$\mu_{a,1} = \sum_{i=1}^{N} \frac{P(a|\vec{x_i})\,x_{i,1}}{\sum_{i=1}^{N} P(a|\vec{x_i})}$

$(\Sigma_a)_{j,k} = \sum_{i=1}^{N} \frac{P(a|\vec{x_i})(x_{i,j}-\mu_{a,j})(x_{i,k}-\mu_{a,k})}{\sum_{i=1}^{N} P(a|\vec{x_i})}$

axis $\in \{0,1\}$

↓ covariance matrix

New 'b'

Similar formulas as 'a'

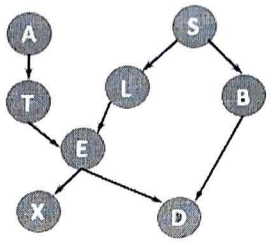
New 'a'	New 'b'
$\mu_a' = (1.32, 1.67)$	$\mu_b' = (6.12, 6.44)$
$\Sigma_a' = \begin{bmatrix}0.286 & 0.185\\0.185 & 0.31\end{bmatrix}$	$\Sigma_b' = \begin{bmatrix}1.11 & 1.06\\1.06 & 1.22\end{bmatrix}$

Grade: _____

Homework 2
ECE/CS 498 DS Spring 2020
Issued: 02/24/20
Due: 03/02/20

Name: Goutham Kunturmalla
NetID: gowtham4

Problem 3



The chest clinic network above concerns the diagnosis of lung disease (tuberculosis, lung cancer, or both, or neither). In this model, a visit to Asia is assumed to increase the probability of tuberculosis. We have the following binary variables:

Variable	
X	positive X-ray
D	dyspnea (shortness of breath)
E	either tuberculosis or lung cancer
T	tuberculosis
L	lung cancer
B	bronchitis
A	a visit to Asia
S	Smoker

A. What are the differences between a joint probability and a conditional probability? Briefly explain.

(A) Joint Probability (say P(A,B)) is a Probability /Count/ of chance both A,B occuring at the same time

(B) whereas conditional probability indicates the happening/chance of A such that 'B' has already happened. In general $P(A|B) \neq P(B|A)$

B. Write down the factorization of the joint probability P(A,T,E,X,L,S,B,D) based on the network.

start bottoms up for simplification

But $P(A,B) = P(A|B)\,P(B)$
$= P(B|A)\,P(A)$

$P(A,T,E,X,L,S,B,D)$

$= P(X|E)\,P(D|E,B)\,P(B|S)\,P(L|S) *$

$P(E|T,L)\,P(T|A)\,P(A)\,P(S)$

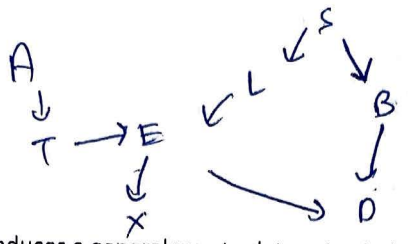
local semantics and chain rules are applied.

Grade:_____

Homework 2
ECE/CS 498 DS Spring 2020
Issued: 02/24/20
Due: 03/02/20

Name: _____

NetID: Gowtham4



We just need one active trail for independence

C. This video introduces a general way to determine independence relationships in Bayes Net: https://www.coursera.org/lecture/probabilistic-graphical-models/flow-of-probabilistic-influence-1eCp1.

Example: Is it true that tuberculosis ⫫ smoking | shortness of breath (given shortness of breath, tuberculosis and smoking are independent)?

Solution: There are two trails from T to S: (T, E, L, S) and (T, E, D, B, S). The trail (T, E, L, S) features a collider node E that is opened by the conditioning variable D. The trail is thus active and we do not need to check the second trail because for independence all trails needed to be blocked. The independence relationship does thus generally not hold.

Are the following conditional independence relationships true or false? Explain why.

1. either tuberculosis or lung cancer (E) ⫫ bronchitis (B) | smoking (S)

True, since 'S' is given, B is fixed and independent of other children/descedants of S.

2. positive x-ray (X) ⫫ smoking (S) | lung cancer (L)

True, Once 'L' is known, the trail is cutoff from 'S'. 'X' is then independent of influence of S.

3. a visit to Asia (A) ⫫ bronchitis(B) | lung cancer (L), shortness of breath (D)

$A \perp L, D$ & $B \not\perp L, D$

false, There is an active trail, A-T-E-D-B

D. Express the P(D) by marginalizing the joint probability, simplify it using your answer in B

$$P(D) = \sum_{A,T,E,X,L,S,B} P(D \mid A,T,E,X,L,S,B)\, P(A,T,E,X,L,S,B) = \sum_{A,T,E,X,L,S,B} P(\text{all vars})$$

$$= \sum_{A,T,E,L,S,B} P(X \mid E)\, P(D \mid E,B)\, P(B \mid S)\, P(L \mid S)\, P(E \mid T,L)\, P(T \mid A)\, P(A)\, P(S) \quad (P(X \mid E) = 1)$$

$$= \sum_{A,T,E,L,S,B} P(D \mid E,B)\, P(B \mid S)\, P(L \mid S)\, P(E \mid T,L)\, P(T \mid A)\, P(A)\, P(S)$$

Grade: _____