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

Name: ____________________

NetID: ___________________

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 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)

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

$$Centroid_1 = \frac{(1,1)+(1,2)+(2,2)}{3} = \frac{(4,5)}{3} = (\frac{4}{3},\frac{5}{3})$$

$$Centroid_2 = \frac{(6,6)+(6,7)+(7,7)}{3} = \frac{(19,20)}{3} = (\frac{19}{3},\frac{20}{3})$$

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

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.
   Stopping/Convergence criterion:
      1. No (or minimal) re-assignments of data points to different clusters.
      2. No (or minimal) change of centroids.

Grade:__________

3. No decrease in the sum of squared error (SSE).

Because centroids have be changed in the latest iteration, the algorithm has not converged yet. (You can check any of the above convergence criteria, but the conclusion should be the algorithm has not converged.)

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.

Change K-Means to K-Medians.

Compared with mean value, median is less sensitive to outliers. How far away from the median the outliers lie is inconsequential. For example, the median of the following two data sets are identical: (1) 1,2,3,4,5; (2) 1,2,3,4,150. We can see that although (2) has an outlier 150, its median is still 3. In comparison, the mean value will shift tremendously from 3 in (1) to 32 in (2).

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 $a \sim N(3,1)$.

$$p(x=1|a) = \frac{1}{\sqrt{2\pi\sigma_a^2}}\exp\left(-\frac{(x-\mu_a)^2}{2\sigma_a^2}\right) = \frac{1}{\sqrt{2\pi(1)}}\exp\left(-\frac{(1-(3))^2}{2(1)}\right) = 0.0540$$

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

In a similar fashion to question 1, we calculate $p(x_i|b)$

$$p(x=1|b) = \frac{1}{\sqrt{2\pi\sigma_b^2}}\exp\left(-\frac{(x-\mu_b)^2}{2\sigma_b^2}\right) = \frac{1}{\sqrt{2\pi(1)}}\exp\left(-\frac{(1-(4))^2}{2(1)}\right) = 0.00443$$

Homework 2 **Name: ____________________**

ECE/CS 498 DS Spring 2020

Issued: 02/24/20 **NetID: ___________________**

Due: 03/02/20

Then, the posterior of $p(a|x_i)$ is calculated by

$$p(a|x=1) = \frac{P(x=1|a) \times P(a)}{P(x=1|a) \times P(a) + P(x=1|b) \times P(b)} = \frac{0.5 \ \times 0.054}{0.5 \times (0.054 + 0.00443)} = 0.924$$

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 $P(a \mid x_i)$	Posterior $P(b \mid x_i)$
1	0.9241418199787564	0.07585818002124355
1	0.9241418199787564	0.07585818002124355
2	0.8175744761936437	0.18242552380635635
6	0.07585818002124355	0.9241418199787564
6	0.07585818002124355	0.9241418199787564
7	0.02931223075135632	0.9706877692486436

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

Using the posteriors calculated from the previous question,

$$\mu_a = \frac{\sum_{i=1}^{6} P(a|x_i)x_i}{\sum_{i=1}^{6} P(a|x_i)} = 1.61541952$$

$$\sigma_a^2 = \frac{\sum_{i=1}^{6} P(a|x_i)(x_i-\mu_a)^2}{\sum_{i=1}^{6} P(a|x_i)} = 1.611$$

$$\mu_b = \frac{\sum_{i=1}^{6} P(b|x_i)x_i}{\sum_{i=1}^{6} P(b|x_i)} = 5.83584601$$

$$\sigma_b^2 = \frac{\sum_{i=1}^{6} P(b|x_i)(x_i-\mu_b)^2}{\sum_{i=1}^{6} P(b|x_i)} = 2.40950419$$

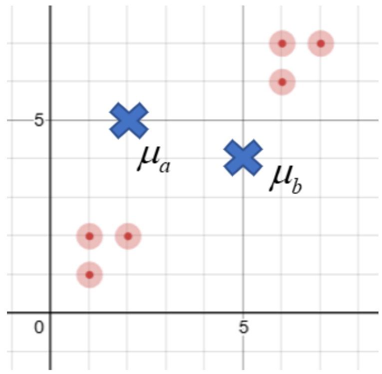
Task 2 2-D GMM:

Now consider clustering the original data given in Task 1 using 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.

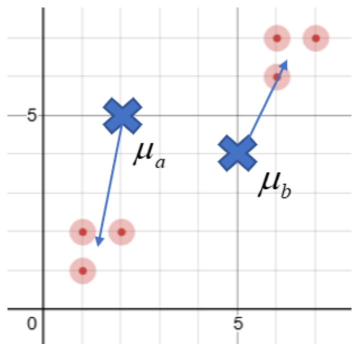
Grade:__________

Homework 2 **Name: ____________________**
ECE/CS 498 DS Spring 2020
Issued: 02/24/20 **NetID: ___________________**
Due: 03/02/20

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.



Solution:



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.

Since $\Sigma_a = \Sigma_b = I$, we know that the determinants $|\Sigma_a| = |\Sigma_b| = 1$. Thus,

$$P(\vec{x_i}|a) = \frac{1}{(2\pi)^{\left(\frac{2}{2}\right)}\sqrt{|\Sigma_a|}} \exp\left\{-\frac{1}{2}(\vec{x_i} - \vec{\mu_a})^T \Sigma_a^{-1} (\vec{x_i} - \vec{\mu_a})\right\}$$

$$= \frac{1}{2\pi\sqrt{(1)}} \exp\left\{-\frac{1}{2}\left(\begin{bmatrix}2\\2\end{bmatrix} - \begin{bmatrix}2\\5\end{bmatrix}\right)^T I \left(\begin{bmatrix}2\\2\end{bmatrix} - \begin{bmatrix}2\\5\end{bmatrix}\right)\right\} = 0.001768$$

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

We can calculate $P((2,2)|b)$ in a same manner we calculated $P((2,2)\,|\,a)$:

$$P(\vec{x_i}|b) = \frac{1}{(2\pi)^{\left(\frac{2}{2}\right)}\sqrt{|\Sigma_b|}} \exp\left\{-\frac{1}{2}(\vec{x_i} - \vec{\mu_b})^T \Sigma_b^{-1} (\vec{x_i} - \vec{\mu_b})\right\}$$

Grade:__________

Homework 2 **Name: ____________________**
ECE/CS 498 DS Spring 2020
Issued: 02/24/20 **NetID: ___________________**
Due: 03/02/20

$$= \frac{1}{2\pi\sqrt{(1)}} \exp\left\{-\frac{1}{2}\left(\begin{bmatrix}2\\2\end{bmatrix} - \begin{bmatrix}5\\4\end{bmatrix}\right)^T I \left(\begin{bmatrix}2\\2\end{bmatrix} - \begin{bmatrix}5\\4\end{bmatrix}\right)\right\} = 0.0002392$$

Next, we know that

$$P\big(a|(2,2)\big) = \frac{P\big((2,2)|a\big) \times P(a)}{P\big((2,2)|a\big) \times P(a) + P\big((2,2)|b\big) \times P(b)}$$

Since we do not have other information for prior P(a) and P(b), we assigned both to be 0.5. Recall that in ICA 1, we also assign priors to be equal when there is no other information.

Therefore,

$$P\big(a|(2,2)\big) = \frac{P\big((2,2)|a\big) \times 0.5}{P\big((2,2)|a\big) \times 0.5 + P\big((2,2)|b\big) \times 0.5} = \frac{0.5 \ \times 0.001768}{0.5 \ \times (0.001768 + 0.0002392)} = 0.8808$$

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.

data point x_i	Posterior $P(a \mid x_i)$	Posterior $P(b \mid x_i)$
(1,1)	0.9820137900379085	0.017986209962091562
(1,2)	0.9933071490757152	0.006692850924284862
(2,2)	0.8807970779778824	0.11920292202211766
(6,6)	0.002472623156634776	0.9975273768433652
(6,7)	0.006692850924284862	0.9933071490757152
(7,7)	0.0003353501304664781	0.9996646498695335

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

Using posteriors calculated from the previous question,

$$\mu_a = \frac{\Sigma_{i=1}^{6} P(a|\vec{x_i})\vec{x_i}}{\Sigma_{i=1}^{6} P(a|\vec{x_i})} = [1.32406143, 1.6730262]$$

$$\Sigma_a = \frac{\Sigma_{i=1}^{6} P(a|\vec{x_i})(\vec{x_i} - \vec{\mu_a})(\vec{x_i} - \vec{\mu_a})^T}{\Sigma_{i=1}^{6} P(a|\vec{x_i})} = [[0.28652493, 0.18511676], [0.18511676, 0.31089694]]$$

Grade:__________

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

Name: ____________________

NetID: ___________________

$$\mu_b = \frac{\Sigma_{i=1}^{6} P(b|\vec{x_i})\vec{x_i}}{\Sigma_{i=1}^{6} P(b|\vec{x_i})} = [6.12744387, 6.44648611]$$

$$\Sigma_b = \frac{\Sigma_{i=1}^{6} P(b|\vec{x_i})(\vec{x_i} - \vec{\mu_b})(\vec{x_i} - \vec{\mu_b})^T}{\Sigma_{i=1}^{6} P(b|\vec{x_i})} = [[1.10802717, 1.05669068], [1.05669068, 1.22260849]]$$

Problem 2 (Optional)

Weather (sunny or rainy) and Location (town or highway) have the potential to cause disengagements of autonomous vehicles. These disengagements could lead to accidents. Given the Bayes Net in Figure 1, answer the following questions:

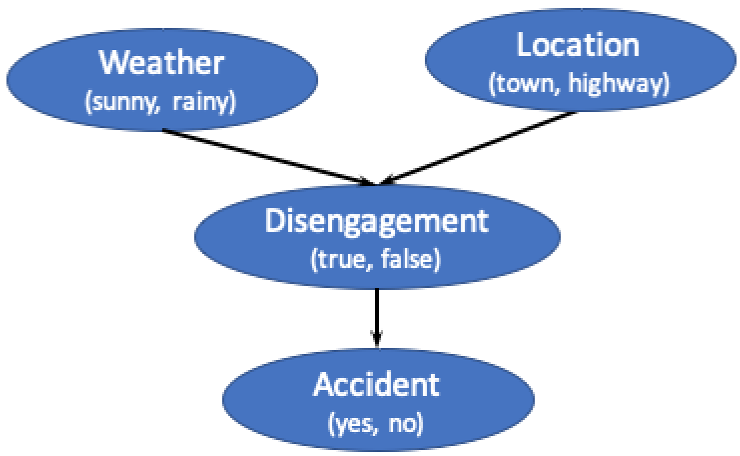


Figure 1

Weather	Probability
sunny	0.7
rainy	0.3
Location	**Probability**
town	0.8
highway	0.2

Disengagement Conditional Probability Table (CPT)			
Weather	**Location**	**Disengagement=true**	**Disengagement=false**
sunny	town	0.05	0.95
sunny	highway	0.01	0.99
rainy	town	0.15	0.85
rainy	highway	0.05	0.95

Grade:__________

Accident CPT		
Disengagement	**Accident=yes**	**Accident=no**
true	0.4	0.6
false	0.01	0.99

A. How many parameters are needed to define the conditional probability distribution of the Bayes Net given in Figure 1?

This question means to ask how many parameters are required to define the joint probability distribution using the Bayes Net given in Figure 1.

$$P(W, L, D, A) = P(W)P(L)P(D|W, L)P(A|D)$$

1+1+4+2=8 parameters

B. Construct the **joint probability distribution** of Weather, Location, and Disengagement

$$P(W, L, D) = P(D|W, L)P(W)P(L)$$

Weather	Location	Disengagement=true	Disengagement=false
sunny	town	0.028	0.532
sunny	highway	0.0014	0.1386
rainy	town	0.036	0.204
rainy	highway	0.003	0.057

C. Calculate the probability of the following hypotheses
Let
a. A = Accident
b. D = Disengagement
c. W = Weather
d. L = Location

$$P(W, L|A) = \frac{P(W, L, A)}{P(A)} = \frac{\sum_D P(W, L, A, D)}{P(A|D)P(D) + P(A|\bar{D})P(\bar{D})}$$

$$= \frac{\sum_D P(W)P(L)P(D|W, L)\, P(A|D)}{P(A|D)P(D) + P(A|\bar{D})P(\bar{D})}$$

$$= \frac{P(W)P(L)\sum_D P(D|W,L)P(A|D)}{P(A|D)P(D) + P(A|\bar{D})P(\bar{D})}$$

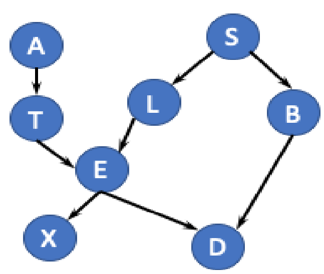
$P(D)$ and $P(\bar{D})$ can be found by summing up the individual columns from the joint probability table in part B, and the rest of the values can be found directly from the provided conditional probability and prior tables.

Hypothesis		Probability
H0	$P(W = sunny, L = town \mid A = yes)$	0.4504
H1	$P(W = sunny, L = highway \mid A = yes)$	0.0531
H2	$P(W = rainy, L = town \mid A = yes)$	0.4482
H3	$P(W = rainy, L = highway \mid A = yes)$	0.0483

D. Apply the MAP decision rule to the 4 hypotheses above.

Looking across all the Hypotheses, H0 has the maximum probability. Hence MAP decision rule will select H0.

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.

Consider two events A and B.

The joint probability $P(A, B)$ is the probability of both events happening at the same time. For a joint probability, the total sample space is the set of all possible outcomes for A and B.

The conditional probability $P(A|B)$ is the probability of event A occurring given that event B is known to have occurred. In other words, it is the probability of event A in the reduced sample space consisting only of outcomes where event B has occurred.

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

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

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)

• There are two trails from e to b: (e, l, s, b) and (e, d, b)

• The trail (e, l, s, b) is blocked by s (s is in a tail-tail configuration and part of the conditioning set)

• The trail (e, d, b) is blocked by the collider configuration for node d as none of d or its descendants are given.

• All trails are blocked, which means the independence relation holds.

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

• There are two trails from a to s: (x, e, l, s) and (x, e, d, b, s)

• The trail (x, e, l, s) is also blocked by l so no information flow.

• The trail (x, e, d, b, s) is blocked by the collider node d.

• All trails are blocked, so the independence relation holds.

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

• There are two trails from a to b: (a, t, e, l, s, b) and (a, t, e, d, b)

• The trail (a, t, e, l, s, b) features a collider node e that is opened by the conditioning variable d but the l node is closed by the conditioning variable l: the trail is blocked

• The trail (a, t, e, d, b) features a collider node d that is opened by conditioning on d. On this trail, e is not in a head-head (collider) configuration) so that all nodes are open and the trail active.

• Hence, the independence relation does not hold.

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

This question meant to ask "Express the probability P(D) by marginalizing the joint probability and then simplifying it using your answer in part B".

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

$$= \Sigma_{A,T,E,X,L,S,B} P(X|E)P(D|E,B)P(E|T,L)P(L|S)P(B|S)P(S)P(T|A)P(A)$$