Wednesday, March 11, 2020

ECE/CS 498 DSE
Midterm Exam
Spring 2020

SOLUTIONS

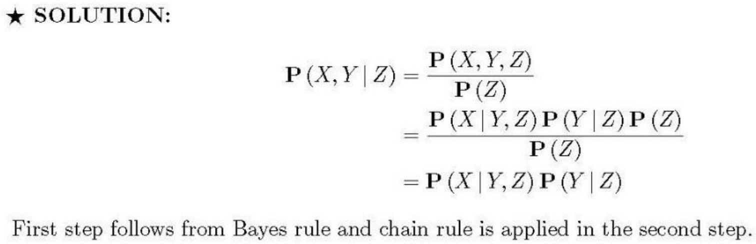
- Write your NetID at the top of each page
- This is a closed book exam
- You are allowed ONE double-sided 8.5” x 11” sheet of notes
- Absolutely no interaction between students is allowed
- Show all your work. Answers without appropriate justification will receive very little or no credit.
- If you need extra space, use the back of the previous page

Problem	**Points Possible**	**Points Earned**
1	17	
2	28	
3	17	
4	18	
5	12	
Total	**92**	

Problem 1 (17 points) – Short Answer Questions:

1. Conditional Probabilities
   a. Prove or disprove (by providing a counter-example) the following property: $P(X, Y|Z) = P(X|Y, Z)P(Y|Z)$ **(2 points)**

   Answer:



★ SOLUTION:

$$\begin{aligned}\mathbf{P}(X, Y \mid Z) &= \frac{\mathbf{P}(X, Y, Z)}{\mathbf{P}(Z)} \\ &= \frac{\mathbf{P}(X \mid Y, Z)\,\mathbf{P}(Y \mid Z)\,\mathbf{P}(Z)}{\mathbf{P}(Z)} \\ &= \mathbf{P}(X \mid Y, Z)\,\mathbf{P}(Y \mid Z)\end{aligned}$$

First step follows from Bayes rule and chain rule is applied in the second step.

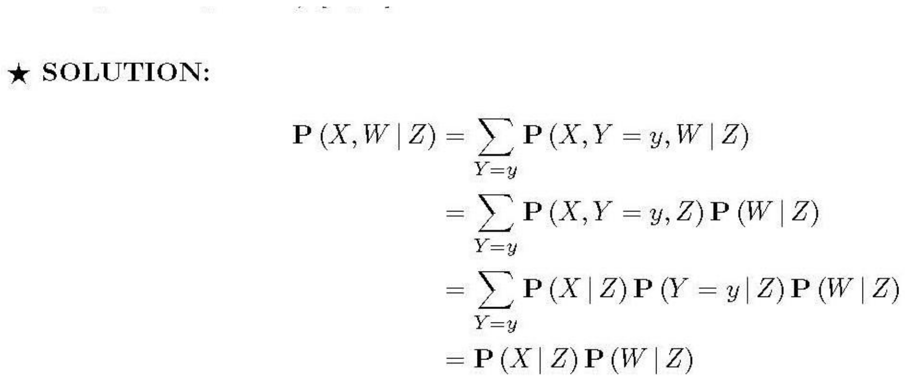
   b. You are given 4 random variables W, X, Y and Z

   - X and Y are independent given Z ($X \perp \mathrm{Y|Z}$)
   - X and Y are jointly independent of W given Z ($(X, Y) \perp W|Z$)

   Show that X is independent of W given Z (i.e., that $X \perp W \mid Z$).

   [HINT: $(A \perp B\,|C) \Rightarrow P(A, B|C) = P(A|C)P(B|C)$] **(3 points)**

   Answer:



★ SOLUTION:

$$\begin{aligned}\mathbf{P}(X, W \mid Z) &= \sum_{Y=y} \mathbf{P}(X, Y = y, W \mid Z) \\ &= \sum_{Y=y} \mathbf{P}(X, Y = y, Z)\,\mathbf{P}(W \mid Z) \\ &= \sum_{Y=y} \mathbf{P}(X \mid Z)\,\mathbf{P}(Y = y \mid Z)\,\mathbf{P}(W \mid Z) \\ &= \mathbf{P}(X \mid Z)\,\mathbf{P}(W \mid Z)\end{aligned}$$

2. A patient goes to the doctor for a medical condition, and the doctor suspects three diseases as the cause of the condition. The three diseases are D1, D2, D3, which are independent of each other. The doctor wants to check four symptoms S1, S2, S3, S4 in order to find the most probable cause of the condition. The symptoms are dependent on the three diseases as follows: S1 depends only on D1, S2 depends on D1 and D2. S3 depends on D1 and D3, whereas S4 depends only on D3. Assume all random variables are Boolean - they are either ‘true’ or ‘false’.

   a. Draw the Bayesian network for this problem. **(2 points)**
   Answer:

★ **SOLUTION:** The Bayesian network is shown in Figure 2.

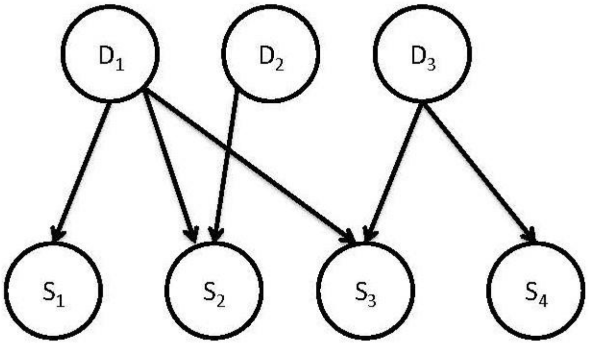


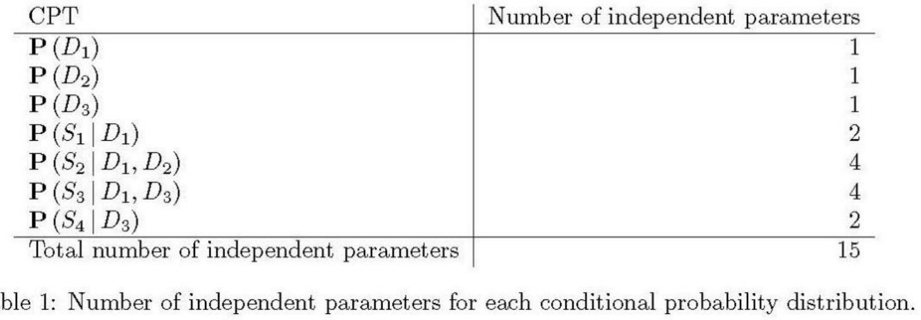
Figure 2: The Bayesian network for disease symptom problem.

   b. Write down the expression for the joint probability distribution as a product of conditional probabilities after applying local semantics. **(2 points)**
   Answer:

★ **SOLUTION:**

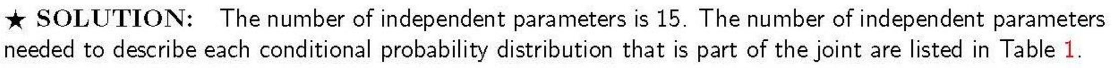
$$\mathbf{P}\left(D_1, D_2, D_3, S_1, S_2, S_3, S_4\right) = \mathbf{P}\left(D_1\right)\mathbf{P}\left(D_2\right)\mathbf{P}\left(D_3\right)\mathbf{P}\left(S_1 \,|\, D_1\right)\mathbf{P}\left(S_2 \,|\, D_1, D_2\right)\mathbf{P}\left(S_3 \,|\, D_1, D_3\right)\mathbf{P}\left(S_4 \,|\, D_3\right)$$

   c. What is the number of independent parameters that is required to describe this joint distribution using the Bayesian Network in (a)? **(2 points)**
   Answer:



CPT	Number of independent parameters
$\mathbf{P}(D_1)$	1
$\mathbf{P}(D_2)$	1
$\mathbf{P}(D_3)$	1
$\mathbf{P}(S_1 \mid D_1)$	2
$\mathbf{P}(S_2 \mid D_1, D_2)$	4
$\mathbf{P}(S_3 \mid D_1, D_3)$	4
$\mathbf{P}(S_4 \mid D_3)$	2
Total number of independent parameters	15

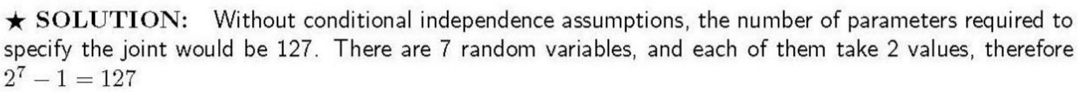
Table 1: Number of independent parameters for each conditional probability distribution.



★ **SOLUTION:** The number of independent parameters is 15. The number of independent parameters needed to describe each conditional probability distribution that is part of the joint are listed in Table 1.

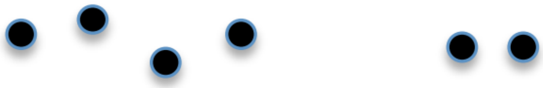
d. If there were no conditional independence relationships between the variables, then how many independent parameters would be required to specify the joint distribution? **(2 points)**

Answer:

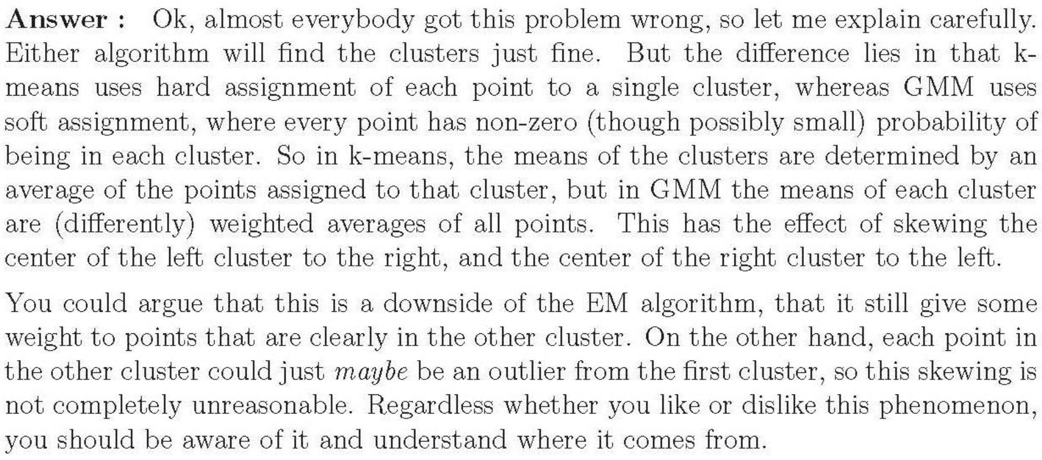


★ **SOLUTION:** Without conditional independence assumptions, the number of parameters required to specify the joint would be 127. There are 7 random variables, and each of them take 2 values, therefore $2^7 - 1 = 127$

3. Consider the set of training data below, and two clustering algorithms: K-Means with K=2, and a Gaussian Mixture Model (GMM) with 2 Gaussian components trained using EM. Will the centroids produced by K-Means be identical to the means of the Gaussian components after convergence? In one to two sentences, explain why or why not. [The scale along the horizontal and vertical axis is the same] **(2 points)**



Answer:



Answer : Ok, almost everybody got this problem wrong, so let me explain carefully. Either algorithm will find the clusters just fine. But the difference lies in that k-means uses hard assignment of each point to a single cluster, whereas GMM uses soft assignment, where every point has non-zero (though possibly small) probability of being in each cluster. So in k-means, the means of the clusters are determined by an average of the points assigned to that cluster, but in GMM the means of each cluster are (differently) weighted averages of all points. This has the effect of skewing the center of the left cluster to the right, and the center of the right cluster to the left.

You could argue that this is a downside of the EM algorithm, that it still give some weight to points that are clearly in the other cluster. On the other hand, each point in the other cluster could just *maybe* be an outlier from the first cluster, so this skewing is not completely unreasonable. Regardless whether you like or dislike this phenomenon, you should be aware of it and understand where it comes from.

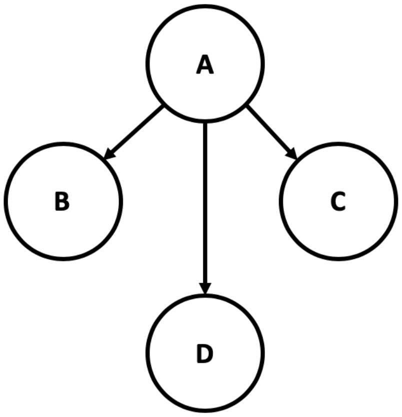
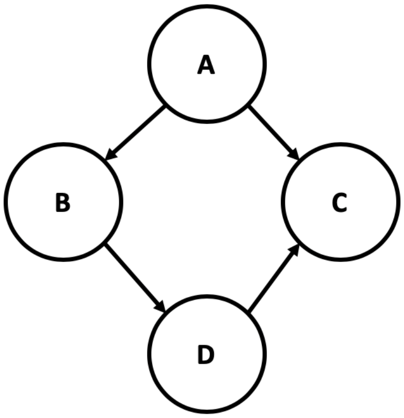
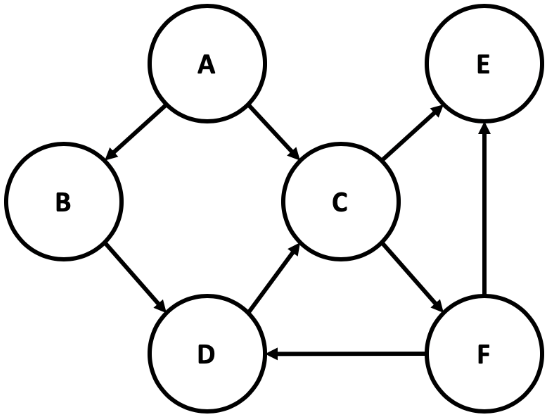
4. As an aspiring zoologist (who studies animals), you want to learn if there is a relationship between an elephant's brain weight, trunk weight, and overall weight. You summarize your results into the dataset shown below, where each sample is a row consisting of three features/dimensions. If you need to perform PCA on this dataset, would it be more appropriate to use the correlation or the covariance matrix? Explain your answer in a few sentences. **(2 points)**

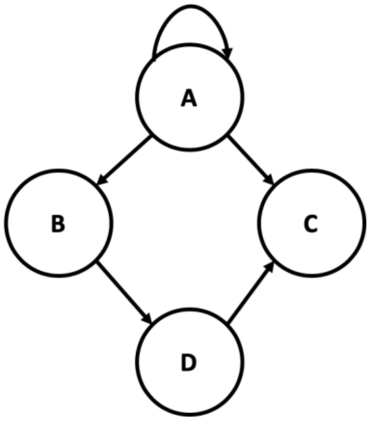
Brain Weight (lbs)	Trunk Weight (kg)	Overall Weight (tons)
9.7	120	4.1
9.3	113	3.3
10.1	137	4.5
11.5	145	5.2

It would be more appropriate to perform PCA on this dataset using the **correlation matrix**. Since the three different features report weight using incomparable units, the covariance matrix would be too sensitive to scaling among the features and would artificially emphasize the trunk weight feature (which has a range of 32 kg). For example, the range of the overall weight feature is 1.9 tons = 1723 kg, which is actually much larger than that of the trunk. Since the correlation matrix is the normalized covariance matrix, it accounts for all of the changes in units and scale between the features.

Problem 2 (28 points) – Bayesian Networks:

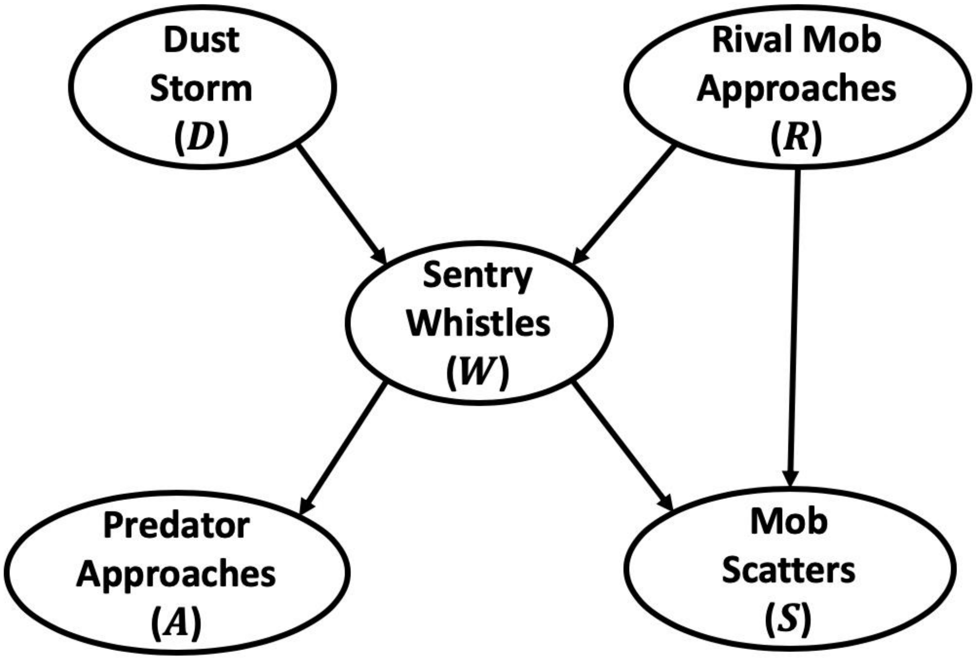
1. For each of the following networks, circle either **TRUE** or **FALSE** **(8 points - 1 point for each correct choice)**

	This figure represents a Naïve Bayes Network [✓ TRUE] FALSE This figure represents a Bayesian Network [✓ TRUE] FALSE
	This figure represents a Naïve Bayes Network TRUE [✓ FALSE] This figure represents a Bayesian Network [✓ TRUE] FALSE
	This figure represents a Naïve Bayes Network TRUE [✓ FALSE] This figure represents a Bayesian Network TRUE [✓ FALSE]

	This figure represents a Naïve Bayes Network TRUE [✓ FALSE] This figure represents a Bayesian Network TRUE [✓ FALSE]

2. Meerkats live in communities called **mobs**. In any mob, during the daytime, there is always one adult meerkat who serves as a **sentry**. The sentry climbs to a point of high visibility, and then whistles whenever it detects danger (such as an incoming dust storm or a rival mob). When the sentry whistles, the mob scatters into its burrows to hide from the imminent threat. The mob also scatters if they directly observe a rival mob approaching before the sentry sees it and whistles. Sometimes, when the sentry whistles, nearby predators (such as snakes) hear the noise and then approach the mob to investigate.

   From your first few weeks ECE/CS 498 DS, you realize that you can model this situation using the Bayesian network drawn below. All of the variables are binary and can only take on values from $\{yes, no\}$.



a. For each node, list all of its non-descendants **(5 points - 1 points for each node)**

Node	**Non-descendants**
D	R
R	D
W	D, R
A	D, R, W, S
S	D, R, W, A

b. State the rule for local semantics **(2 points)**

Given its parents, a node is independent of its other non-descendants.

c. Suppose that it is known that $D = yes,\ W = yes,\ A = no,$ and $S = yes$. Based on this evidence, you need to decide whether or not there is a rival mob that is approaching.

i. Using the definition of conditional probability, rewrite $P(R = yes|D = yes, W = yes, A = no, S = yes)$ as a fraction. Do not factorize the numerator or denominator. **(2 points)**

$$P(R = yes|D = yes, W = yes, A = no, S = yes)$$
$$= \frac{\boldsymbol{P(R = yes, D = yes, W = yes, A = no, S = yes)}}{\boldsymbol{P(D = yes, W = yes, A = no, S = yes)}}$$

ii. Factorize the joint distribution $P(R, D, W, A, S)$ using local semantics with the provided Bayesian network. **(2 points)**

$$P(R, D, W, A, S) = \boldsymbol{P(S|W, R)P(W|D, R)P(A|W)P(D)P(R)}$$

iii. Consider the following expressions:

(1) $P(R = yes|D = yes, W = yes, A = no, S = yes)$
(2) $P(R = yes, D = yes, W = yes, A = no, S = yes)$

a. Classify each expression as a (i) joint probability, (ii) conditional probability, (iii) marginal probability, (iv) none of the above. **(2 points)**

Expression (1) is a conditional probability, and expression (2) is a joint probability.

b. Are the expressions the same? If not, briefly discuss in one to two sentences how they are different. **(3 points)**

The two expressions are not the same.

Expression (1) defines the probability of the event $R = yes$ in the reduced sample space where $D = yes,\ W = yes,\ A = no$, and $S = yes$.

By conditioning on the evidence, we restrict the probabilistic universe to only all possible outcomes where $D = yes,\ W = yes,\ A = no$, and $S = yes$, and then find the probability that $R = yes$ in the set of remaining outcomes.

Expression (2) defines the joint probability of the events $R = yes,\ D = yes,\ W = yes,$ $A = no$, and $S = yes$ in the total sample space.

In other words, it defines the probability of all of these events happening simultaneously in the set of all possible outcomes for R, D, W, A, and S.

d. Answer the following questions on independence and explain your answer.

i. Are $\boldsymbol{D}$ and $\boldsymbol{R}$ independent? **(2 points)**

Yes. There are two possible trails between D and R.

The trail $D \rightarrow W \leftarrow R$ consists of a v-structure where the collider node W is not observed, which means that this trail is not active.

The trail $D \rightarrow W \rightarrow S \leftarrow R$ consists of another v-structure where the collider node S is not observed, and thus is not active.

Since all trails between D and R are not active, there is no information flow between the two variables and the independence relationship holds.

ii. If ***S*** is observed, are ***D*** and ***R*** independent? **(2 points)**

No. The trail $D \rightarrow W \rightarrow S \leftarrow R$ consists of a v-structure where the collider node S is observed. Also, none of the non-collider nodes D, W, and R are observed, which means this trail is active. Thus, since there is an active trail between D and R, there is information flow between the two variables and the independence relationship does not hold.

Problem 3 (17 points) – PCA

1. **Short Answers**

a. Is the dimension of the projected data points on the principal components always less than the dimension original data points? **Select True or False. (1 point)**

True / False

b. Suppose you have calculated the covariance matrix Σ, how can you get the **total** variance from this matrix? Give a mathematical expression. **(2 points)**

Total variance is trace(Σ).

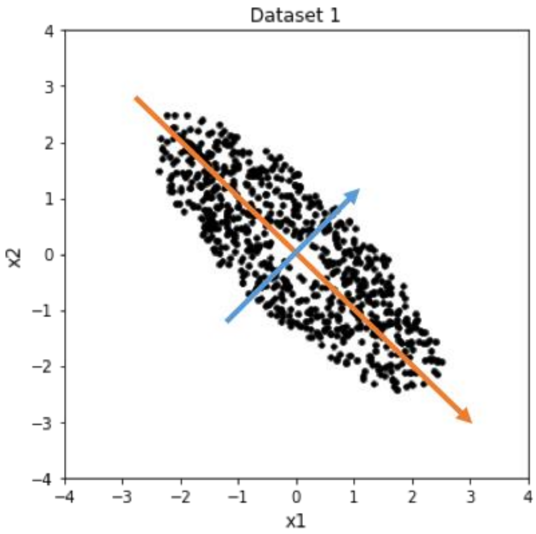
c. What information do the individual eigenvalues of the covariance matrix provide? **Select one below. (1 point)**

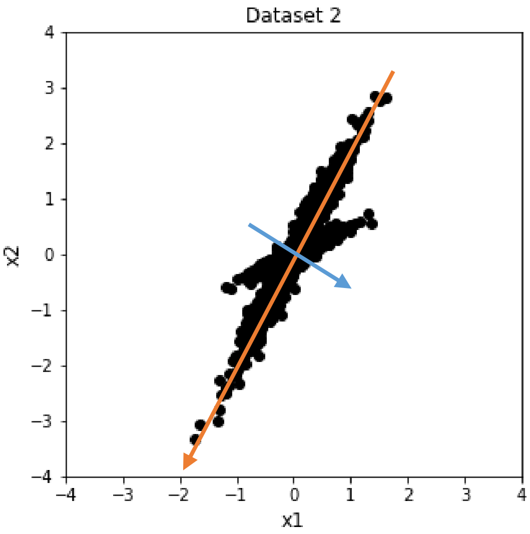
i. The fractional variance explained by the corresponding principal component
ii. The variance explained by the corresponding principal component
iii. The length of the corresponding principal component
iv. None of the above

d. Given a collection of data points in matrix X, we project all datapoints in this matrix onto the principal component space (the subspace represented by the principal components) and name the collection of datapoints after the projection Y. Let $\Omega = Y^\top Y$ be the covariance matrix of Y. **What are the values at the off-diagonal positions of Ω?** Explain your answer within two sentences. **(2 points)**

The covariance between the projected points onto two different PCs are zero because PCs are orthogonal to each other, as a result, Ω is a diagonal matrix.

2. For each of the plots below draw the **first and second** principal components on the plot. Use <u>length</u> to denote an estimate of the variances. **(4 points – 2 points for each plot)**





3. Suppose you have three data points (-√3, -2√3), (0, 0) and (√3, 2√3) and you want to find the principal components for them. You have put them into the matrix $X=\begin{bmatrix} \sqrt{3} & 2\sqrt{3} \\ 0 & 0 \\ -\sqrt{3} & -2\sqrt{3} \end{bmatrix}$.

a. Using the data matrix X above, calculate the variance-covariance matrix Σ. **(3 points)**

$$\Sigma = \frac{1}{3}\,X^{\top}X = \frac{1}{3}\begin{bmatrix} \sqrt{3} & 0 & -\sqrt{3} \\ 2\sqrt{3} & 0 & -2\sqrt{3} \end{bmatrix}\begin{bmatrix} \sqrt{3} & 2\sqrt{3} \\ 0 & 0 \\ -\sqrt{3} & -2\sqrt{3} \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 6 & 12 \\ 12 & 24 \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 4 & 8 \end{bmatrix}$$

b. Using the covariance matrix, find the **unit eigenvector** corresponding to the **first** principal component. Is using the first principal component enough for representing the data? **(4 points – 2 points for each question)**

Finding the eigenvalues and the eigenvectors:

$$\begin{vmatrix} 2-\lambda & 4 \\ 4 & 8-\lambda \end{vmatrix} = 0$$

$$(2-\lambda)(8-\lambda) - 16 = 0$$

$$\lambda^2 - 10\lambda = 0$$

$$\lambda = 0, \lambda = 10$$

Select the largest eigenvalue, which is 10, find the eigenvector:

$$\begin{bmatrix} 2-10 & 4 \\ 4 & 8-10 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{0}$$

$$\mathbf{4}\boldsymbol{x} - \mathbf{2}\boldsymbol{y} = \mathbf{0}$$

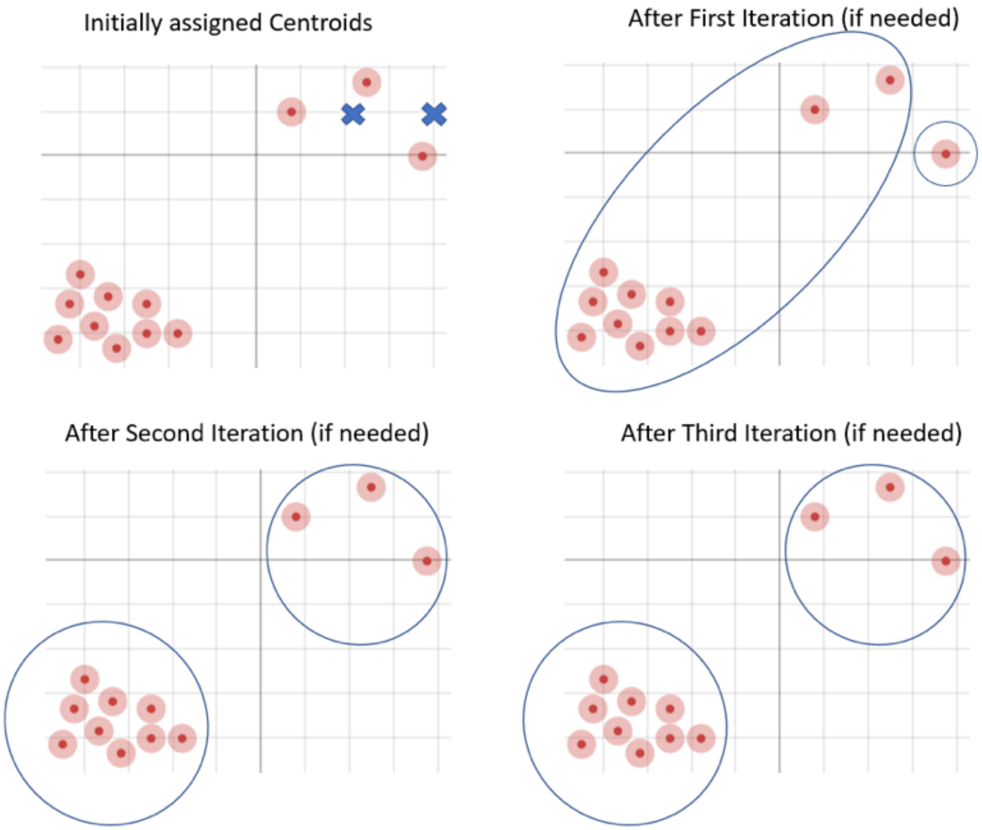
$$\mathbf{2}\boldsymbol{x} = \boldsymbol{y}$$

The **first** principle component is therefore $\begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}$. It is enough because the other eigenvector is zero so no variance is explained by the second principal component, we can therefore drop it.

Problem 4 (18 Points) – Clustering

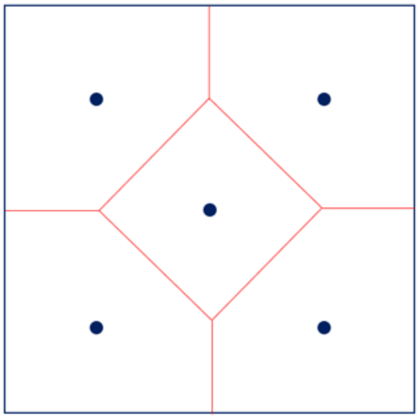
1. **K-Means and GMM**

   a. You are given the following dataset, where dots are the data and crosses are the initial centroids. For each iteration of K-means, circle the two clusters that are assigned. Recall that in a given iteration, cluster assignment is performed before updating the centroid location. The resulting clusters after the first iteration are provided to you. How many iterations do we need until we can confirm convergence (which in this case is when the cluster assignments remain the same)? **(4 points – 1 point for each iteration, 2 points for the number of iterations needed until convergence)**


Initially assigned Centroids
After First Iteration (if needed)
After Second Iteration (if needed)
After Third Iteration (if needed)

We will need 3 iterations in order to converge via K-Means. Although only 2 iterations are needed to arrive at the final cluster assignments, a third iteration is needed to confirm that we have in fact converged.

b. Suppose that the centroids found by K-Means for a different dataset are drawn in the following figure. Draw the decision boundaries between the clusters. [HINT: the decision boundaries are the lines that separate the regions for different cluster assignments] **(4 points)**



c. John is trying to cluster (1,2), (6,6), (2,4) into two clusters *a* and *b*. He is using the EM algorithm with two gaussian components. Suppose in an E-step, John calculates the posteriors for cluster *a* as follows:

	x= (1,2)	x= (6,6)	x= (2,4)
P(*a*|x)	0.7	0.6	0.7

Calculate the new mean of Gaussian Component *b* in the following M-step. **(4 points)**

	x= (1,2)	x= (6,6)	x= (2,4)
P(b|x) = 1 - P(a|x)	0.3	0.4	0.3

$$\mu_b = \frac{\sum P(b \mid x_i) x_i}{\sum P(b \mid x_i)} = \frac{(0.3\times 1 + 0.4\times 6 + 0.3\times 2, 0.3\times 2 + 0.4\times 6 + 0.3\times 4)}{0.3+0.4+0.3} = (3.3, 4.2)$$

2. Hierarchical Clustering

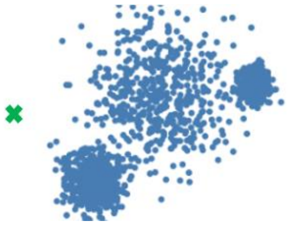
a. In lecture, we explored Novel HIV Drug Resistance Mutations using clustering. In this exam, we are going to cluster some of the mutations using complete-link hierarchical clustering. The mutations and their distances are presented in the following distance matrix.

	L210W	M41L	T215Y	L228R	K219R
L210W	0	0.66	0.65	0.05	0.23
M41L	0.66	0	0.74	0.1	0.17
T215Y	0.65	0.74	0	0.09	0.15
L228R	0.05	0.1	0.09	0	0.19
K219R	0.23	0.17	0.15	0.19	0

Rewrite the distance matrix after first merge. Remember to fill out the shaded boxes as well! **(5 points)**

	L210W and L228R	M41L	T215Y	K219R
L210W and L228R	0	0.66	0.65	0.23
M41L	0.66	0	0.74	0.17
T215Y	0.65	0.74	0	0.15
K219R	0.23	0.17	0.15	0

b. Suppose we have applied hierarchical clustering on the following dataset, and they have been clustered into 2 classes. The data point represented as a cross is in its own cluster and all the other data points represented by dots are assigned to the second cluster. Note that the horizontal and vertical scales in the figure are the same. Determine which inter-class distance metric was used. **Select one below. (1 point)**



i. Single link
ii. Average link
iii. Complete link

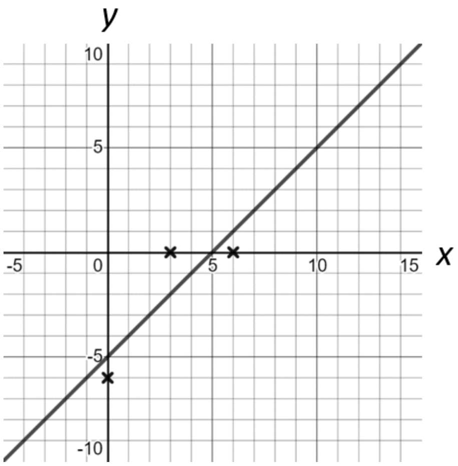
Problem 5 (12 points) – Regression

You are given the following data points $D = \{(0, -6), (3, 0), (6, 0)\}$. All points are referred to as (x, y), where x is the independent parameter and y is the dependent one. Performing the least squares linear regression to this dataset gives us the following regression parameters:

$$\{\alpha\ (intercept) = -5, \qquad \beta\ (slope) = 1\}$$

i.e. The linear regression line is $\hat{y} = x - 5$. The coefficient of determination for the fit is $r^2 = \frac{3}{4}$

The graph below shows the datapoints and the best fit line.



1. We try to fit a 2nd order polynomial curve to the datapoints specified above. Suppose we calculate the coefficient of determination r^2 for the quadratic regression model in the same manner as for the linear regression model. What value will r^2 take for the quadratic regression model? Will it be better than that of the linear regression model? Explain your answer within three sentences. **(3 points)**

   Since there are only 3 non-collinear points in the dataset, we can fit a quadratic curve perfectly to 3 points, resulting in $r^2 = 1$. So yes, the coefficient of determination r^2 for the quadratic regression model will be better than the r^2 for the linear regression model.

2. You are given another datapoint (-1, 4). Compute the linear regression parameters (α, β) for the line that best fits this new dataset: **(4 points- 2 points each for α and β)**

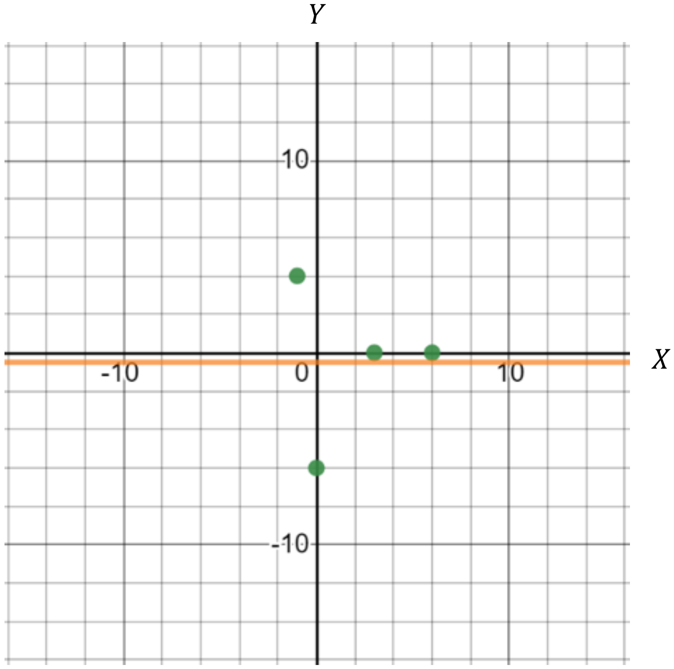
$$D_{new}(\boldsymbol{x}, \boldsymbol{y}) = \{(0, -6), (3, 0), (6, 0), (-1,4)\}$$

$$\beta = \frac{4 \times (-4) - (8) \times (-2)}{4 \times 46 - 64} = \frac{0}{120} = 0$$

$$\alpha = \frac{-2 \times (46) - (8) \times (-4)}{4 \times 46 - 64} = \frac{-60}{120} = -0.5$$

$$\hat{y} = 0x - 0.5$$

3. Plot the datapoints in D_{new} and the new least squares linear fit. **(2 points)**



4. You are told that the datapoint (-1, 4) is an outlier. Explain the effect of outliers on the least square estimation for linear regression. **(3 points)**

Linear regression is sensitive to outliers. When we add an outlier to the original dataset, it incurs a very large residual with respect to the original linear regression line in part 1. Since the least squares fit tries to minimize the square of the residuals over all data points (including the outlier), the new regression line will favor minimizing the large residual at the expense of increasing the residuals for inliers. This results in a worse fit for the inliers, which is supported by results in part 3 ($r_{new}^2 = 0 < r^2$, fit with slope 0).