

ECE/CS 498 DSE
Midterm Exam
Spring 2018

Name: _____

NetID: _____

- Write your NetID at the top of each page
- This is a closed book exam
- You are allowed ONE 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	Grade	Total
1		15
2		20
3		25
4		20
5		20
Bonus		20
Total		100 + 20

Problem 1 (15 points) – Short Answer Questions

1. Three random variables have the following joint distribution:

$$P(X,Y,Z)=P(X)P(Y|X)P(Z|Y)$$

Show that X and Z are conditionally independent given Y .

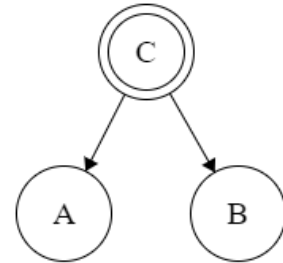
2. Explain conceptually using a single dimension the idea of Mahalanobis distance.

The advantage of using Mahalanobis distance over Euclidean distance is (circle all that apply):

- a) It is unitless
- b) It is inexpensive to compute compared to the Euclidean distance
- c) It is scale invariant
- d) It gives more importance to features with larger variances

3. What is the goal of Principal Component Analysis?
Use a figure to illustrate your answer.

4. Keeping Naïve Bayes classification in mind, write the definition of class conditional independence.



5. Define single-link and complete-link similarity functions.
6. In the lecture notes, we used the Expectation Maximization (EM) algorithm to learn the parameters of a Gaussian Mixture Model that enables soft-clustering of the observations (x_1, x_2, \dots, x_n) .
- a. Write down the steps of the EM algorithm for learning the parameters of Gaussian Mixture Model.

- b. During the in-class activity, your partner makes the following statement:

“Expectation Maximization algorithm can be used to learn parameters of any mixture models as long as the mixtures are represented by a valid distribution.”

S/he came with the following functions, $F(x)$ to represent mixtures. Circle all the choices that are valid $F(x)$ to represent mixture models.

Explain your selected choice(s).

- i. S/he is wrong. Only Gaussian distribution can be used for soft-clustering using Expectation Maximization.
- ii. Gamma distribution
- iii. Beta distribution

iv.
$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ |\sqrt{x}| & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

v.
$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{x^2 - 2x + 3}{3} & \text{if } 0 \leq x \leq 2 \\ 1 & \text{if } x > 2 \end{cases}$$

- vi. Only (ii), (iii), (iv) and (v)
- vii. Only (ii), (iii) and (iv)

Problem 2 (20 points)

1. Find the parameters of the linear regression (α, β) that best represents the following training data:

$$\{(0,0), (1,2), (2,4), (4,8), (10,20)\}$$

2. What is the coefficient of determination r^2 of your trained model?

3. What will be the quadratic least squares fit ($\|predicted - actual\|_2$) for $f(x) = a + bx + cx^2$ when applied to the same training data?

$$a = \underline{\hspace{2cm}}$$

$$b = \underline{\hspace{2cm}}$$

$$c = \underline{\hspace{2cm}}$$

4. We are trying to learn regression parameters for a dataset which we know was generated from a polynomial of a certain degree, but we do not know the degree of the polynomial that was used to generate the dataset.

Assume the data was generated from a polynomial of degree 5 with some added Gaussian noise (that is $y = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + w_5x^5 + \varepsilon, \varepsilon \sim N(0,1)$).

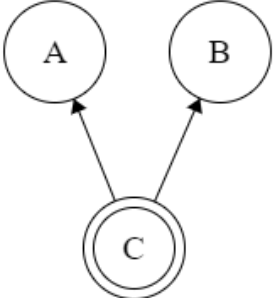
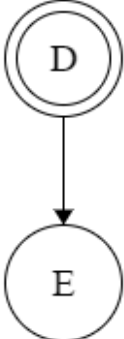
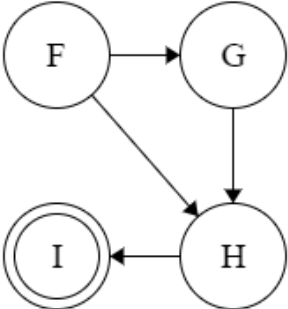
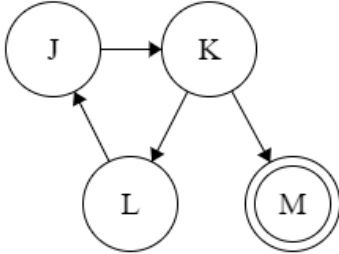
For training we have 100 (x, y) pairs and for testing we are using an additional set of 100 (x, y) pairs. Since we do not know the degree of the polynomial we learn two models from the data. Model A learns parameters for a polynomial of degree 4 and model B learns parameters for a polynomial of degree 6.

Which of these two models is likely to fit the *test* data better?

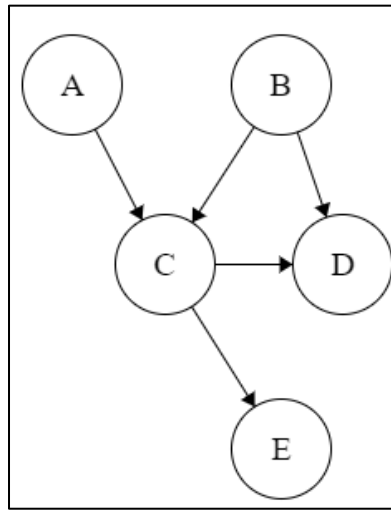
Problem 3 (25 points) – Bayesian Networks

1. For each of the following figures, circle **TRUE** or **FALSE**

Nodes with double circles represent the variable we are trying to predict.

<p>A.</p> 	<p>This figure represents a Naïve Bayes Network</p> <p>TRUE FALSE</p> <p>This figure represents a Bayesian Network</p> <p>TRUE FALSE</p>
<p>B.</p> 	<p>This figure represents a Naïve Bayes Network</p> <p>TRUE FALSE</p> <p>This figure represents a Bayesian Network</p> <p>TRUE FALSE</p>
<p>C.</p> 	<p>This figure represents a Naïve Bayes Network</p> <p>TRUE FALSE</p> <p>This figure represents a Bayesian Network</p> <p>TRUE FALSE</p>
<p>D.</p> 	<p>This figure represents a Naïve Bayes Network</p> <p>TRUE FALSE</p> <p>This figure represents a Bayesian Network</p> <p>TRUE FALSE</p>

2. Consider the following Bayesian Network. Each variable indicated below can take on *True* or *False* values, i.e. they are binary variables.



- i. If C is observed, are **D and E** independent?
- ii. If C is observed, are **A and B** independent?
- iii. Assuming C has not been observed, are **A and B** independent?
- iv. Calculate the probability of H_2 and apply the MAP decision rule to the 4 hypotheses below.

Note that X represents $X = T$ and \bar{X} represents $X = F$.

Hypothesis		Probability	Decision
H_0	$P(A, B E)$	0.082	
H_1	$P(A, \bar{B} E)$	0.020	
H_2	$P(\bar{A}, B E)$		
H_3	$P(\bar{A}, \bar{B} E)$	0.188	

- v. Apply the chain rule to calculate H_2 : $P(\bar{A}, B \mid E)$.

$$P(\bar{A}, B \mid E) =$$

Problem 4 (20 points) – k -means Clustering

1. You have a dataset with n variables/features (e.g., height, weight and calories consumed by the student).
 - i. What do you understand about the standardization of the dataset with respect to performing k -means clustering.
 - ii. How would you ensure these dimensions are comparable, so that you can run k -means?
 - iii. Give an example of another 3-dimensional dataset for which **standardization might** be necessary to get satisfactory results from k -Means clustering.
2. The log-likelihood of the data maximizes when every observation in an arbitrary given dataset forms a unique cluster with centroid being the observation point itself (i.e. n observations form n unique clusters). How will this solution perform with respect to a new test dataset that has the same number and types of features? Explain your answer.

3. Consider this training data set. Observations are $A - F$, and the single feature is X .

Observation	Feature Value (X)
A	0.1
B	0.6
C	0.8
D	2.0
E	3.0
F	4.0

You are told that this dataset forms two natural clusters.

Randomly, you choose observation A to initialize cluster #1 and observation B to initialize cluster #2. Assume your SSE calculations are based on the Euclidian distance function.

- i. Write down the cluster assignments that result from applying k -means to this data. Write C , D , E and F in the blanks below according to which cluster they are assigned (A and B are already assigned).

cluster #1: (A , _____) cluster #2: (B , _____)

- ii. After assigning examples to clusters in 3.i, you will recompute the cluster centroids for the next iteration of k -means. Calculate and write below the new centroids of the two clusters. You can write the answer in fractions.

Problem 5 (20 points) – GMM Clustering

1. If we were to replace the Gaussian Mixture Model with a Gamma Mixture Model, what would the new Expectation Maximization (EM) steps be for the new model?

2. Assume each data point $X_i \in \mathbb{R}^+$ ($i = 1 \dots n$) is drawn from the following process:

$$Z_i \sim \text{Multinomial}(\pi_1, \pi_2, \dots, \pi_K)$$

$$X_i \sim \text{Gamma}(2, \beta_{Z_i})$$

The probability density function of $\text{Gamma}(2, \beta)$ is $P(X = x) = \beta^2 x e^{-\beta x}$.

- a) Assume $K = 3$ and $\beta_1 = 1, \beta_2 = 2, \beta_3 = 4$. What is $P(Z = 1 | X = 1)$?

b) Describe the Expectation step. Write an equation for each value being computed.

c) For each of the following statements, select **TRUE** or **FALSE**. Provide a one sentence explanation for each.

i. Gamma mixture model can capture overlapping clusters, like Gaussian mixture model.

TRUE

FALSE

ii. As you increase K , you will always get better likelihood of the data.

TRUE

FALSE

5. Consider the training dataset in Problem 4 Part 3 (Page 10). You found the cluster assignments for the observations $A - F$. If the cluster assignment in iteration 2 is different than the cluster assignment in iteration 1 (found by you).

Is it possible for the k-means algorithm to revisit a centroid assignment in future iterations?

Justify how your answer proves that the k means algorithm converges in a finite number of steps.