Monday, March 12, 2018

ECE/CS 498 DSE
Midterm Exam
Spring 2018

SOLUTIONS

- 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.

$$\begin{aligned}
\because \mathcal{P}(X,Y,Z) &= \mathcal{P}(X)\,\mathcal{P}(Y \mid X)\,\mathcal{P}(Z \mid Y) \\
\mathcal{P}(X,Z \mid Y) &= \frac{\mathcal{P}(X,Y,Z)}{\mathcal{P}(Y)} \\
&= \frac{\mathcal{P}(X)\,\mathcal{P}(Y \mid X)\,\mathcal{P}(Z \mid Y)}{\mathcal{P}(Y)} \\
&= \frac{\mathcal{P}(X)\,\mathcal{P}(Y \mid X)}{\mathcal{P}(Y)}\,\mathcal{P}(Z \mid Y) \\
&= \mathcal{P}(X \mid Y)\,\mathcal{P}(Z \mid Y) \\
\therefore X \perp Z \mid Y &
\end{aligned}$$

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

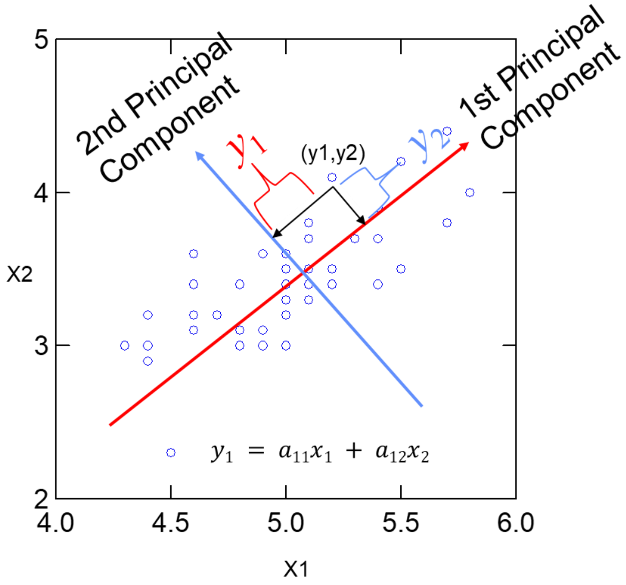
Mahalanobis distance between point x_1 and the mean μ of a distribution with standard deviation σ is $\frac{x_1-\mu}{\sigma}$. It represents the number of standard deviations away the point is from the mean.

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

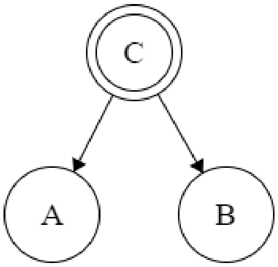
a) and c)

3. What is the goal of Principal Component Analysis?
Use a figure to illustrate your answer.
To find a set of orthonormal vectors such that each successive vector (principal component) accounts for the maximum possible unexplained variance in the data.



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

$$P(A,B|C) = P(A|C)P(B|C)$$



5. Define single-link and complete-link similarity functions.

Single-link: $d_{min}(C_i, C_j) = \min_{p \in C_i, q \in C_j} d(p,q)$

Complete link: $d_{max}(C_i, C_j) = \max_{p \in C_i, q \in C_j} d(p,q)$

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.

If there are K mixture components $c_1, \dots, c_K$, then

Expectation step:

For all $j \in \{1, \dots, K\}$

$$P(c_j|x_i) = \frac{p(x_i|c_j)P(c_j)}{\sum_{k=1}^{K} p(x_i|c_k)P(c_k)}$$

Maximization step:

For all $j \in \{1, \dots, K\}$, let the j^{th} Gaussian distribution have parameters μ_j, σ_j^2 and $P(c_j)$. Then,

$$\mu_j = \frac{\sum_{i=1}^{n} P(c_j|x_i)x_i}{\sum_{i=1}^{n} P(c_j|x_i)}$$

$$\sigma_j^2 = \frac{\sum_{i=1}^{n} P(c_j|x_i)\ (x_i - \mu_j)^2}{\sum_{i=1}^{n} P(c_j|x_i)}$$

$$P(c_j) = \frac{\sum_{i=1}^{n} P(c_j|x_i)}{n}$$

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: Yes, it is a valid distribution.

iii. Beta distribution : Yes, it is a valid distribution.

iv. $F(x) = \begin{cases} 0 & if\ x < 0 \\ \left|\sqrt{x}\right| & if\ 0 \leq x \leq 1 \\ 1 & if\ x > 1 \end{cases}$: Yes, it is a valid distribution.

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

No, it is not a valid distribution because it is not non-decreasing.

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

vii. Only (ii), (iii) and (iv) : Yes, the options mentioned here are valid distributions.

Problem 2 (20 points)

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

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

Method 1: Apply the formula from the lecture slides or HW3 to get the answer. Formulae are:

$$\beta = \frac{N(\sum xy) - (\sum x) \cdot (\sum y)}{N(\sum x^2) - (\sum x)^2}$$

$$\alpha = \frac{(\sum y) \cdot (\sum x^2) - (\sum x) \cdot (\sum xy)}{N(\sum x^2) - (\sum x)^2}$$

Method 2: By visual inspection, the points lie on the line

y = 2x. Therefore $\alpha = 0,\ \beta = 2$.

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

r^2 gives the proportion of total variation that is explained by the regression model.

$$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Substitute the values from above to get the answer. Or, notice that since the regression line perfectly explains the data,

$$r^2 = 1$$

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 quadratic model is an extension of a linear model. Since the linear model perfectly explains the data, the coefficient of the quadratic term will be zero.

$a = 0$

$b = 2$

$c = 0$

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 $y = w0 + w1x + w2x^2 + w3x^3 + w4x^4 + w5x^5 + \varepsilon, \varepsilon \sim N(0,1)$).

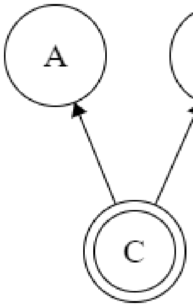
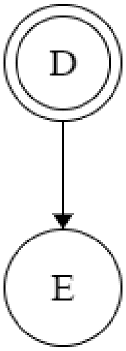
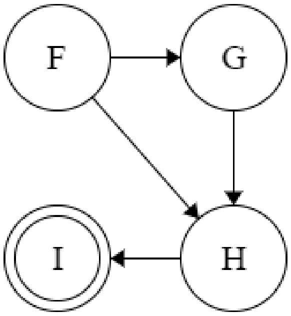
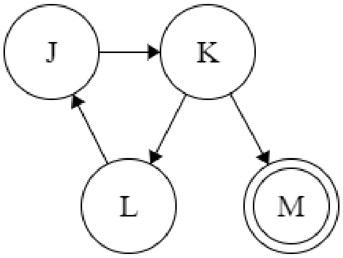
   For training we have 100 (x, y) pairs and for testing we are using an additional set of 100 (x, y) pairs. Since we do not know the degree of the polynomial we learn two models from the data. Model A learns parameters for a polynomial of degree 4 and model B learns parameters for a polynomial of degree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

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

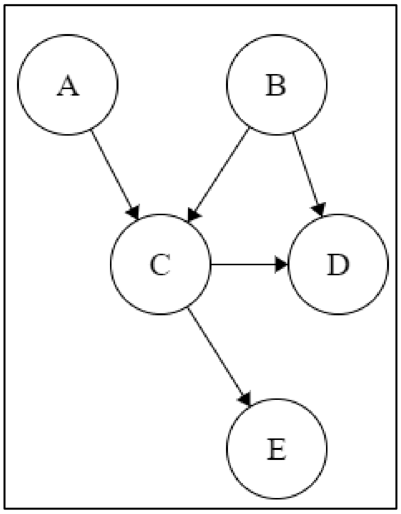
Model with degree 6 overfits and the model with degree 4 underfits.

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.

A.		This figure represents a Naïve Bayes Network [✓ **TRUE**] **FALSE** This figure represents a Bayesian Network [✓ **TRUE**] **FALSE**
B.		This figure represents a Naïve Bayes Network [✓ **TRUE**] **FALSE** This figure represents a Bayesian Network [✓ **TRUE**] **FALSE**
C.		This figure represents a Naïve Bayes Network **TRUE** [✓ **FALSE**] This figure represents a Bayesian Network [✓ **TRUE**] **FALSE**
D.		This figure represents a Naïve Bayes Network **TRUE** [✓ **FALSE**] This figure represents a Bayesian Network **TRUE** [✓ **FALSE**]

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?

Local semantics: Each node is conditionally independent of its non-descendants given its parents.

E: node of interest. C: parent. D: non-descendant. Therefore, D and E are independent given C.

ii. If C is observed, are ***A* and *B*** independent?
No.

iii. Assuming C has not been observed, are ***A* and *B*** independent?
Yes.

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	$P(A, B \mid E)$	0.082	Reject
H_1	$P(A, \bar{B} \mid E)$	0.020	Reject
H_2	$P(\bar{A}, B \mid E)$	0.710	Accept
H_3	$P(\bar{A}, \bar{B} \mid E)$	0.188	Reject

P(A,B|E) has four possible values since A and B both can take one of two values. Therefore, the column given above would sum to 1.

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

$$P(\bar{A}, B \mid E) = \frac{P(E, \bar{A}, B)}{P(E)} = \frac{\sum_{C,D} P(E, \bar{A}, B, C, D)}{\sum_{A,B,C,D} P(E, A, B, C, D)}$$

$$= \frac{\sum_{C,D} P(E|C)P(D|B,C)P(C|\bar{A},B)P(\bar{A})P(B)}{\sum_{A,B,C,D} P(E|C)P(D|B,C)P(C|\bar{A},B)P(\bar{A})P(B)}$$

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.

Make the data have zero mean and unit variance.

    ii. How would you ensure these dimensions are comparable, so that you can run k-means?

    By standardizing the variables.

    iii. Give an example of another 3-dimensional dataset for which **standardization might** be necessary to get satisfactory results from k-Means clustering.

Any example works fine. Weather data: Temperature, wind, humidity.

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.

It will not perform well because the clusters will be overfit.

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, _C,D,E,F_)

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.

Cluster 1: 0.1/1 = 0.1
Cluster 2: (0.6+0.8+2+3+4)/5 = 2.08

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?

Expectation step: Compute posterior probability

$$P(c_j|x_i) = \frac{p(x_i|c_j)P(c_j)}{p(x_i|c_1)P(c_1) + p(x_i|c_2)P(c_2) + p(x_i|c_3)P(c_3)}$$

Maximization step: Compute new/update parameter (in this case, the mean)

Mean for the gamma distribution with parameters α and β is $\frac{\alpha}{\beta}$. The formula is the same as calculating the mean of a Gaussian distribution.

$$\frac{\alpha_j}{\beta_j} = \frac{\sum_i P(c_j|x_i)x_i}{\sum_i P(c_j|x_i)}$$

$$P(c_j) = \frac{\sum_i P(c_j|x_i)}{N}$$

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

$$Z_i \sim Multinomial(\pi_i, \pi_2, \dots, \pi_K)$$

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

The probability density function of $Gamma(2, \beta)$ 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 $P(Z = 1\ |X = 1)$?

$$P(Z=1|X=1) \propto P(X=1|Z=1)P(Z=1) = \pi_1 e^{-1}$$
$$P(Z=2|X=1) \propto P(X=1|Z=2)P(Z=2) = \pi_2 4e^{-2}$$
$$P(Z=3|X=1) \propto P(X=1|Z=3)P(Z=3) = \pi_3 16e^{-4}$$

$$P(Z=1|X=1) = \frac{\pi_1 e^{-1}}{(\pi_1 e^{-1} + \pi_2 4e^{-2} + \pi_3 16e^{-4})}$$

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

For each $X = x$,

$$P(Z=k|X=x) = \frac{P(X=x|Z=k)P(Z=k)}{\sum_{k'} P(X=x|Z=k')P(Z=k')} = \frac{\beta_k^2 x e^{-\beta_k x} \pi_k}{\sum_{k'} \beta_{k'}^2 x e^{-\beta_{k'} x} \pi_{k'}}$$

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**

k-means cost functions (20 points - BONUS Question)

1. k-means is an algorithm that has an aim of finding k clusters with minimum SSE. Write down the equation for the SSE. Make sure to properly define any notation you use.

For Euclidean distance,

$$SSE = \sum_j \sum_{i \in C_j} (x_i - \mu_j)^2$$

2. Is SSE based on intra-cluster distance or inter-cluster distance?

Intra-cluster distance.

3. Consider the following strategy for selecting k, the number of clusters. Run k- means clustering with different k values and choose k that minimizes the SSE. What is the potential problem of this strategy?

Overfitting. Every point becomes its own cluster.

4. Suggest a way in which you can eliminate the problem you found above.

Use BIC criterion.

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). Yes. The cluster assignment changes in the second iteration and becomes (A, B, C) and (D, E, F).

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

No.

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

There is always an improvement and SSE is bounded from below. Therefore, it will converge in a finite number of steps.