Instructions: This exam contains 11 pages (including this cover page) and 5 questions. The maximum number of points you can earn on the exam is 20. Read each problem carefully. Show your work. You may consult the textbook and your notes to solve the problems.

<table>
<thead>
<tr>
<th>Question</th>
<th>Points</th>
<th>Bonus Points</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Total:</td>
<td>20</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

1. (4 points) True or False
   
   (a) (0.5 points) In the Generative approach to solving classification problems, we model the conditional probability of the labels given the observations.

   A. True  
   B. False

   (b) (0.5 points) The maximum likelihood model parameters ($\theta_1, \theta_2$) can be learned using linear regression for the model $y_i = \log(x_i^{\theta_1} e^{\theta_2}) + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$ is iid noise and $x_i > 0$.

   A. True  
   B. False

   (c) (0.5 points) A classifier that attains 100% accuracy on the training set and 70% accuracy on test set is better than a classifier that attains 70% accuracy on the training set and 75% accuracy on test set.

   A. True  
   B. False
(d) (0.5 points) Logistic regression binary classifier is able to correctly separate the training data (circles vs. triangles) given in Figure below.

A. True  
B. False 

(e) (0.5 points) For small training sets, Naive Bayes generally is more accurate than logistic regression.

A. True  
B. False 

(f) (0.5 points) Leave-one out cross validation (LOOCV) generally gives less accurate estimates of true test error than 10-fold cross validation.

A. True  
B. False 

(g) (0.5 points) The following two Bayes nets encode the same set of conditional independence relations.

A. True  
B. False
(h) (0.5 points) Does this conditional independence relation hold?

\[ X_2 \perp H_4 | X_3 \]

A. True B. False

2. (4 points) Choose ALL CORRECT answers. Every question have at least one correct answer. NO PARTIAL CREDIT: set of all correct answers (only) must be selected.

(a) (0.5 points) A and B are two events. If \( P(A,B) \) decreases while \( P(A) \) increases, what must be true

A. \( P(A|B) \) decreases. B. \( P(B|A) \) decreases. C. \( P(B) \) decreases. D. All of above.

(b) (0.5 points) Which of the following best describes what discriminative approaches try to model? (\( w \) are the parameters in the model) (Select one.)

A. \( p(y|x, w) \) B. \( p(y, x) \) C. \( p(x|y, w) \) D. \( p(w|x, y) \)

(c) (0.5 points) The error function most suited for gradient descent using logistic regression is


(d) (0.5 points) How does the bias-variance decomposition of a ridge regression estimator compare with that of ordinary least squares regression? (Select one.)

A. Ridge has larger bias, larger variance. B. Ridge has smaller bias, larger variance. C. Ridge has larger bias, smaller variance. D. Ridge has smaller bias, smaller variance.
(e) (0.5 points) Consider the following dataset: $A = (0, 2), B = (0, 1)$ and $C = (1, 0)$. The k-means algorithm is initialized with centers at A and B. Upon convergence, the two centers will be at

A. A and C.  
B. C and the midpoint of AB.  
C. A and the midpoint of BC.  
D. A and B.

(f) (0.5 points) We have a stream of A’s, B’s, and C’s generated from the probability distribution $(p_A = 1/2, p_B = 1/2, p_C = 0)$. However, we instead use a coding that would be optimal if the distribution was $(p_A = 1/2, p_B = 1/4, p_C = 1/4)$. Our coding will (in expectation)

A. Take more bits than the optimal coding.  
B. Take fewer bits than the optimal coding.  
C. Take same bits as the optimal coding.  
D. Not enough information is provided to tell.

(g) (0.5 points) Suppose your model is overfitting. Which of the following is NOT a valid way to try and reduce the overfitting? (Select one.)

A. Increase the amount of training data.  
B. Improve the optimization algorithm being used for error minimization.  
C. Decrease the model complexity.  
D. Reduce the noise in the training data.

(h) (0.5 points) You are reviewing papers for the World’s Fanciest Machine Learning Conference, and you see submissions with the following claims. Which ones would you consider accepting?

A. My method achieves a training error lower than all previous methods!  
B. My method achieves a test error lower than all previous methods! (When regularization parameter $\lambda$ is chosen so as to minimize test error.)  
C. My method achieves a test error lower than all previous methods! (When regularization parameter $\lambda$ is chosen so as to minimize cross-validation error.)  
D. My method achieves a cross-validation error lower than all previous methods! (When regularization parameter $\lambda$ is chosen so as to minimize cross-validation error.)
3. (6 points) We are given samples $x = \{1, 1, 1\}$ and samples $y = \{-2, -2, -2\}$. You can assume that the number of samples in $x$ and $y$ is the same, and denoted by $N$. We start by specifying two models

$$p(x|\mu_0) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} (x - \mu_0)^2 \right\}$$

$$p(y|\mu_1) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} (y - \mu_1)^2 \right\}$$

(a) (1 point) Write log-likelihood for parameter $\mu_0$, $\mathcal{LL}(\mu_0; x)$. Hint: this log-likelihood involves samples $x_i$, where $i = 1, \ldots, N$. You can drop the term $\frac{1}{2} \log(2\pi)$.

(b) (1 point) Write log-likelihood for parameter $\mu_1$, $\mathcal{LL}(\mu_1; y)$. Hint: this log-likelihood involves samples $y_i$, where $i = 1, \ldots, N$. You can drop the term $\frac{1}{2} \log(2\pi)$. 
(c) (1 point) Derive maximum likelihood estimates for $\mu_0$ and $\mu_1$ by taking derivative of corresponding log-likelihood functions and equating them to zero.

(d) (0.5 points) Plug-in the values in the dataset ($x = \{1, 1, 1\}$ and samples $y = \{-2, -2, -2\}$) in the formulas you derived above to compute solutions for $\mu_0$ and $\mu_1$.

- $\mu_0 = \underline{\text{ }}$
- $\mu_1 = \underline{\text{ }}$

(e) (1.5 points) Let new penalized loglikelihood function be equal to

$$P\mathcal{L}(\mu_0, \mu_1; x, y) = \mathcal{L}(\mu_0; x) + \mathcal{L}(\mu_1; y) - \frac{\lambda}{2}(\mu_0 - \mu_1)^2$$

Take derivative with respect to $\mu_0, \mu_1$, equate them to zero and solve the linear system to obtain closed form solutions for both parameters. Hint: You will get ratios that involve $\lambda$ and the number of samples $N$. Solutions for $\mu_0, \mu_1$ are very similar except that $x_i$s and $y_i$s swap their roles.
(f) (0.5 points) Use the closed form solutions above to evaluate $\mu_0, \mu_1$ for the data $x = \{1, 1, 1\}$ and samples $y = \{-2, -2, -2\}$, and for different values of $\lambda$.

- $\lambda = 0 : \mu_0 = \text{__________} \text{ and } \mu_1 = \text{__________}$
- $\lambda = N : \mu_0 = \text{__________} \text{ and } \mu_1 = \text{__________}$

(g) (0.5 points) Describe what happens to $\mu_0, \mu_1$ as $\lambda$ grows larger, $\lambda \to \text{inf}$.
4. (6 points) We are going to train a Naive Bayes classifier on a dataset consisting of four samples, each with two features. Each sample is labeled with either 1 or 2.

\[
X = \begin{bmatrix}
-1 & -1 \\
-1 & +1 \\
+1 & -1 \\
+1 & +1
\end{bmatrix},
\quad y = \begin{bmatrix}
1 \\
2 \\
2 \\
1
\end{bmatrix}
\]

Assume Naive Bayes model with Gaussian distribution for each of the features

\[
p(x_j|\mu) = \prod_c \left( \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} (x_j - \mu_{jc})^2 \right\} \right)_{[y=c]}
\]

\[
p(y = 1) = p(y = 2) = 0.5
\]

where \( \mu_{jc} \) denotes mean of the \( j^{th} \) feature in the class \( c \).

(a) (1 point) Write out log-likelihood for this model, \( \mathcal{L}(\mu; x, y) \). Use \( x_{ij} \) for \( j^{th} \) feature value for sample \( i \), and \( y_i \) for label of sample \( i \).

(b) (1 point) Take derivative of \( \mathcal{L}(\mu; x, y) \) with respect to \( \mu_{jc} \), equate it to zero, and solve.
(c) (0.5 points) Using the above formulas compute values of parameters $\mu_{jc}$ for $j \in \{1, 2\}$ and $c \in \{1, 2\}$

- $\mu_{11} = \underline{\text{__________}}$
- $\mu_{12} = \underline{\text{__________}}$
- $\mu_{21} = \underline{\text{__________}}$
- $\mu_{22} = \underline{\text{__________}}$

(d) (1 point) Use Bayes rule to obtain $p(y = c|x)$. This probability should be expressed in terms of $\mu$ parameters.

(e) (0.5 points) Assume that you predict label 1 for any vector $x$ for which $p(y = 1|x) \geq 0.5$. What is the set of feature vectors for which you would predict label 2? Hint: this is the set of feature vectors for which $p(y = 1|x) < 0.5$.

(f) (1 point) Introduce feature $x_3$ which is computed by $x_{i3} = \text{sign}(x_{i1} * x_{i2})$. Assuming Gaussian distribution of this feature in each class

$$p(x_3|y = c, \mu) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} (x_3 - \mu_{3c})^2 \right\}$$

derive maximum likelihood estimates for $\mu_{31}$ and $\mu_{32}$ and express them in terms of original features $x_{i1}$ and $x_{i2}$. 
(g) (0.5 points) The optimal parameters $\mu_{31}$ and $\mu_{32}$ on this dataset are:

- $\mu_{31} =$
- $\mu_{32} =$

(h) (0.5 points) In the plot below, shade the part of the set for which your Naive Bayes classifier on the three features $x_1, x_2, x_3$ would predict class 2.

![Plot](image)

5. (2 points (bonus)) Given a Bayes Net, which variables (minimal subset) should be given be to make the conditional independence statement True.

![Bayes Net](image)

- $\mathcal{Z} = \{ \}$
[scratch paper]