EEL5840 Fundamentals of Machine Learning
Spring 2024
Midterm Exam
March 1, 2024
Time Limit: 2 hours Name: _____

- Write legibly
- There are a total of 9 questions for a total of 100 points
  - Some questions are worth more than other questions.
- **Closed-book, no computer, one-page formulas, calculator**
  - **Write your name in the formula sheet.**

Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page.

Helpful Formulas

- The inverse of an isotropic matrix $\Sigma = a\mathbf{I} = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$ is $\Sigma^{-1} = \frac{1}{a}\mathbf{I} = \begin{bmatrix} 1/a & 0 \\ 0 & 1/a \end{bmatrix}$
- Exponential distribution:

$$P(x|\lambda) = \lambda e^{-\lambda x}, \quad \lambda > 0, x \geq 0$$

- Beta distribution:

$$P(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad \alpha > 0, \beta > 0, x \in [0,1]$$

where $\Gamma(x) = (x-1)!$.

Closed-book, no computer, one-page formulas, calculator

Grade Table (for the teaching team use only)

Question:	1	2	3	4	5	6	7	8	9	Total
Points:	10	10	15	10	10	15	10	10	10	100
Score:										

1. (10 points) If a face image is a 100×100 image, written in row-major, this is a $10,000$-dimensional vector. If we shift the image one pixel to the right, this will be a very different vector in the $10,000$-dimensional space. How can we build face recognizers robust to such distortions?

2. (10 points) Suppose we have some binary data, $x_i \in \{0, 1\}$. The training data is as follows:

x	t
0	$[-1, 1]^T$
0	$[-1, -2]^T$
0	$[-1, -1]^T$
1	$[1, 1]^T$
1	$[1, 2]^T$
1	$[2, 1]^T$

Let us embed each x_i into a 2D feature space using the basis function:

$$\phi(0) = [1, 0]^T, \qquad \phi(1) = [0, 1]^T$$

Consider the multiple[1] linear regression model $y = \mathbf{W}\phi(x)$, where $\mathbf{W}$ is a 2×2 matrix. Compute the Maximum Likelihood Estimation (MLE) for $\mathbf{W}$ from the above data.

[1] *Multiple* linear regression refers to situations in which the model predicts more than one variable.

3. (15 points) Consider a training dataset $\{(x_i, t_i)\}_{i=1}^{N}$ where $x_i \in \mathbb{R}^D$, $t_i \in \mathbb{R}$ and D is the feature space dimensionality. Suppose you are training a linear regression model, $y(x) = \mathbf{w}^T\mathbf{x} + w_0$. Answer the following questions:

   (a) (4 points) Consider the objective function

$$J(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}(t_i - y(x_i))^2 + \lambda R(\mathbf{w})$$

   What is the role of the regularizer $R(\mathbf{w})$? What is the effect of changing the value of λ?

   (b) (4 points) How would you train a linear regression model that is robust to outliers?

(c) (4 points) In scenarios where at least two dimensions exhibit collinearity, meaning they are linearly dependent, which regularization technique would you employ in your linear regression model while preserving all dimensions?

(d) (3 points) How does regularization help address the challenges posed by the curse of dimensionality?

4. (10 points) In the field of financial transactions, such as online retail, when predicting the total cost of an online purchase, errors in the prediction are likely to be strictly positive because the actual cost cannot be less than zero.

   Let $\epsilon_i = t_i - y_i$ represent the (positive) residual error for input sample x_i and consider the following objective function:

$$J(\mathbf{w}) = \sum_{i=1}^{N} \epsilon_i$$

   What is the Bayesian interpretation of this objective function? Show and justify your work.

5. (10 points) Suppose we are modeling coin tosses (Bernoulli trials),

$$P(x|\mu) = \mu^x(1-\mu)^{1-x} \quad \text{where } x \in \{0,1\}$$

and you think the coin (modeled by parameter μ) is either fair, or is biased towards heads. To model this kind of prior belief, you consider a mixture of two beta distributions:

$$P(\mu) = 0.5\text{Beta}(\mu|\alpha_1,\beta_1) + 0.5\text{Beta}(\mu|\alpha_2,\beta_2)$$

Answer the following questions:

(a) (5 points) Write down the observed data likelihood for the Maximum A Posteriori (MAP) approach.

(b) (5 points) Is this mixture model a conjugate prior? Why or why not?

6. (15 points) Consider a mixture model with 2 components ($K = 2$), where one component is Gaussian-distributed, $\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, and the other component is Exponential-distributed, $E(x|\lambda) = \lambda e^{-\lambda x}$. For a dataset with N i.i.d. samples $\{x_i\}_{i=1}^N$, answer the following questions:

   (a) (8 points) Introduce the hidden latent variables for this mixture model and write down the complete data likelihood.

   (b) (7 points) Derive the update equation for the parameter λ.

7. (10 points) Consider a binary classification task, let the positive class be C_1 and negative class C_0. A positive sample occurs with probability $p_1 = P(C_1)$, and a negative sample with probability $p_0 = P(C_0)$.

   Suppose that you have determined that each class is modeled according to a bivariate Gaussian distribution of the form:

$$P(\mathbf{x}|C_1) \sim \mathcal{N}\left(\mu_1 = \begin{bmatrix}1\\0\end{bmatrix}, \Sigma_1 = \mathbf{I}\right) \quad \text{and} \quad P(\mathbf{x}|C_0) \sim \mathcal{N}\left(\mu_2 = \begin{bmatrix}0\\-1\end{bmatrix}, \Sigma_2 = \mathbf{I}\right)$$

   where $\mathbf{x} = \begin{bmatrix}x_1\\x_2\end{bmatrix}$ and $\mathbf{I}$ is the 2×2 identity matrix.

   (a) (6 points) Let $g_i(\mathbf{x}) = \ln(P(\mathbf{x}|C_i)P(C_i))$ be the discriminant function for class C_i. Then, let $g(\mathbf{x}) = g_1(\mathbf{x}) - g_0(\mathbf{x})$ be the Naïve Bayes decision boundary.

   Find $g(\mathbf{x})$ for this example. Show your work.

(b) (4 points) Let $P(C_1) = \frac{1}{5}$. In the table below, include the label prediction (0 or 1) for each sample (x_1, x_2) based on your solution for part (a). Show your work.

Input $x_i = (x_i^{(1)}, x_i^{(2)})$	Classifier Prediction y_i
$(0, 0)$	
$(1, 0)$	
$(2, 2)$	
$(-1, 2)$	

(Useful information: $\ln(1/4) \approx -1.3862$.)

8. (10 points) Within clustering algorithms, we discussed Gaussian Mixture Models (GMMs) and the K-Means algorithm. What are the advantages of GMMs over K-Means? Describe at least 4 unique advantages.

9. (10 points) One source of noise is error in the labels. Using tools introduced in lecture, can you propose a method to find data points that are highly likely to be misclassified?

HONOR STATEMENT

I understand that I am bound to uphold the honor code of the University of Florida. I have neither given nor received assistance on this examination. In addition, I did not use any outside materials on this exam other than the one page of formulas that was allowed.

Sign Your Name: _____

Write the Date: _____

Print Your Name: _____

Turn in your formula sheet with your exam!!!