**EEL5840 Fund. of Machine Learning** **Name:** _____
**Summer 2022**
**Midterm Exam**
**June 27, 2022**
**Time Limit: 2 hours** UFID _____

- Write legibly
- There are a total of 8 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.**

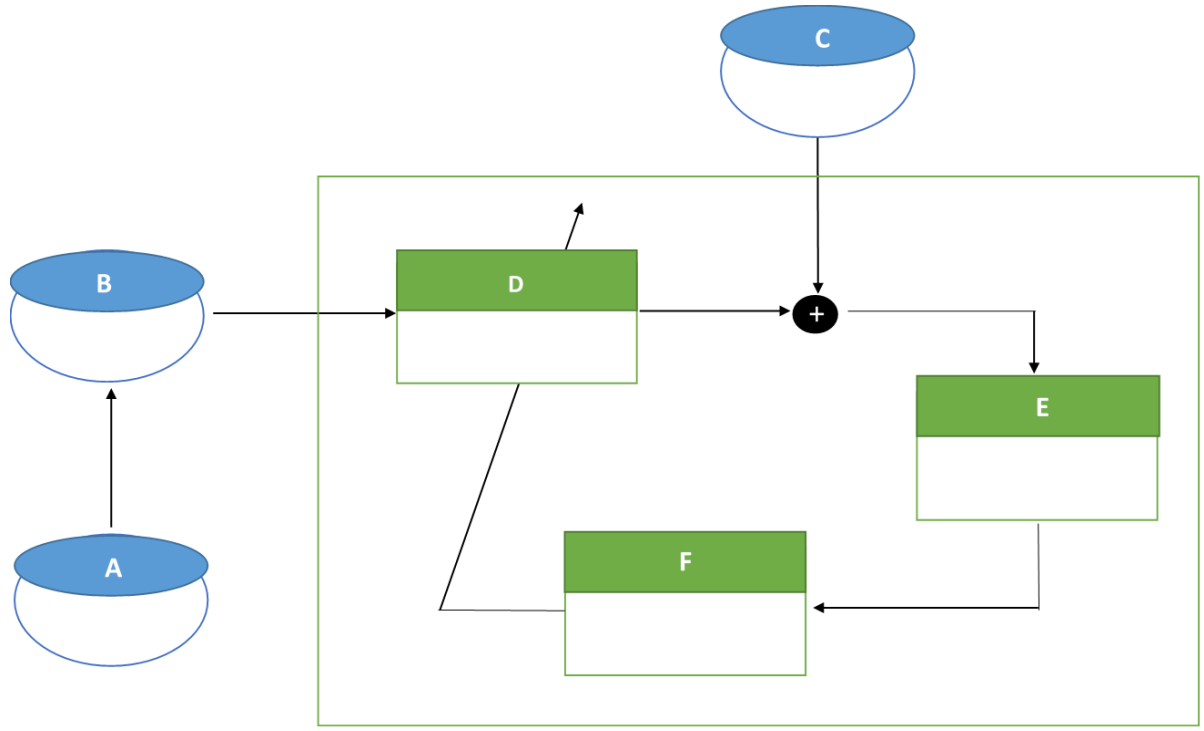
| Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page. |
|---|

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

Grade Table (for teacher use only)

| Question: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Total |
|---|---|---|---|---|---|---|---|---|---|
| Points: | 10 | 5 | 10 | 10 | 10 | 20 | 10 | 25 | 100 |
| Score: | | | | | | | | | |

1. (10 points) Consider the block diagram for a supervised learning system depicted below. Name each block and explain in words the function of each block.



2. (5 points) In class, we saw that the Bayesian interpretation of an objective function with regularization is equivalent to maximizing a data likelihood distribution times a prior probability. Depending on the objective function we choose, we may end up with specific probabilistic models for the data likelihood and the prior.

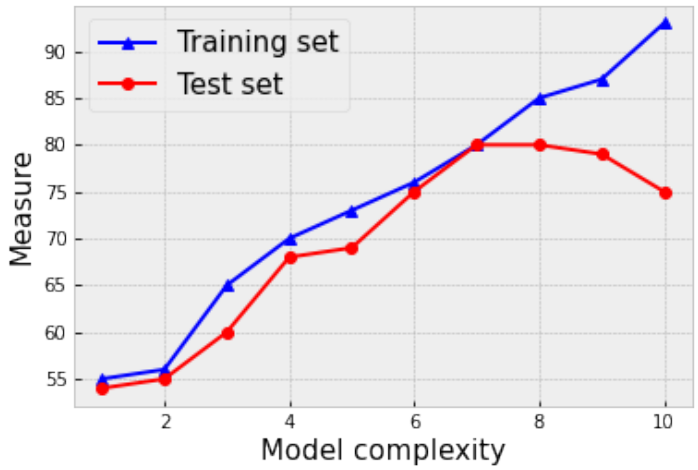
   Why is the Bayesian interpretation useful in machine learning? What advantages does this Bayesian interpretation bring when performing point/parameter estimation?

3. (10 points) True (T) or False (F)?

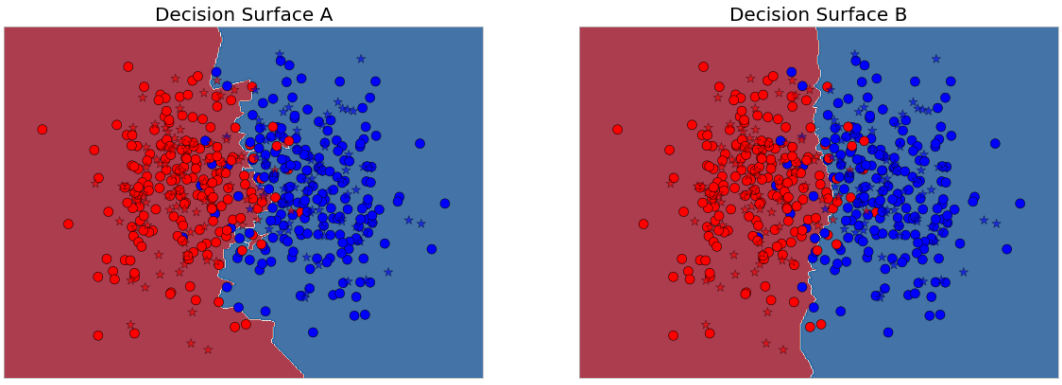
___ Put simply, the curse of dimensionality states that we should never work with high-dimensional data.

___ When a model is overfitting, it has high variance but low bias. When a model is underfitting, it has low variance but high bias.

___ The figure below illustrates the experimental design carried over the training/test set in order to select the model complexity. Based on this figure, the performance measure in the y-axis must be a measure of the error.




___ The figure below shows two decision surfaces obtained from training a K-Nearest Neighbors classifier with the training set (circles) and evaluate prediction on test set (stars). One was obtained for $k = 20$ and the other for $k = 3$. Decision surface A used $k = 3$.




___ When performing clustering with the K-Means algorithm using Mahalanobis distance or with Gaussian Mixture Models with full covariance matrices, we should be able to learn elliptical-shaped clusters.

___ The Naïve Bayes classifier is a type of a discriminative classifier.

___ A data likelihood described by the Bernoulli distribution with parameter $\rho$, $P(x|\rho) = \rho^x(1-\rho)^{1-x}$, and the Geometric prior distribution on the parameter $\rho$, $P(\rho|\lambda) = \lambda(1-\lambda)^\rho$, form a conjugate prior relationship.

___ A ROC (Receiver Operating Characteristic) curve is a binary (2-class) performance metric.

___ Consider the following confusion matrix:

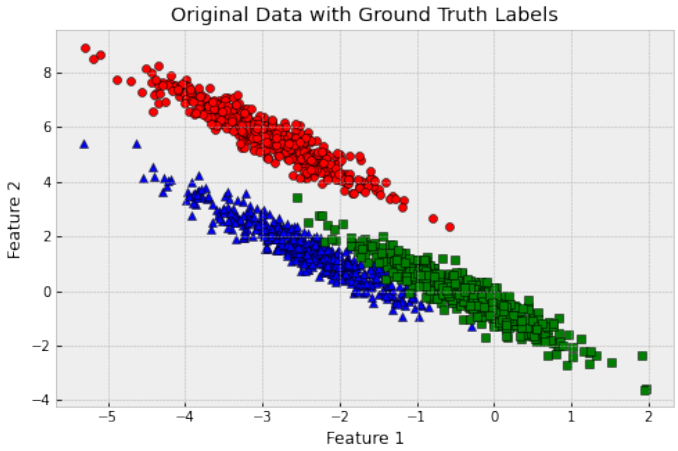
| | | Predicted labels | | |
|---|---|---|---|---|
| | | $C_1$ | $C_2$ | $C_3$ |
| True labels | $C_1$ | 96 | 3 | 1 |
| | $C_2$ | 3 | 42 | 5 |
| | $C_3$ | 6 | 2 | 42 |

The False Positive Rate (FPR) for class $C_1$ is $\frac{5}{150}$, for class $C_2$ is $\frac{6}{150}$ and for class $C_3$ is $\frac{9}{100}$.

___ The Expectation-Maximization (EM) algorithm is implemented in three steps: initial guess for the hidden latent variables $Z$, expectation step and maximization step.

4. (10 points) Write down the pseudo-code to implement the K-Nearest Neighbors classifier with a weighted distance voting scheme. Make sure to include all necessary steps.

5. (10 points) Consider the following dataset with 3 classes:




Suppose that you will run three different clustering algorithms and hope to arrive at the ground truth labels shown in the figure above. Given the choices (1) K-Means with Euclidean distance, (2) Gaussian Mixture Models with full covariance or (3) Gaussian Mixture Models with isotropic covariance matrix, which one would you choose to perform clustering with this dataset? Justify your selection and elaborate why the other 2 would not be a good choice for this dataset.

6. (20 points) Suppose you have a training set with $N$ data points $\{x_i\}_{i=1}^{N}$, where $x_i \in \mathbb{Z}_0^+$ (set of nonnegative integers - $0, 1, 2, 3, \dots$). Assume the samples are independent and identically distributed (i.i.d.), and each sample is drawn from a Geometric random variable with probability mass function:

$$P(x|\rho) = \rho(1-\rho)^x$$

Moreover, consider the Beta density function as the prior probability on the success probability, $\rho$,

$$P(\rho|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\rho^{\alpha-1}(1-\rho)^{\beta-1}$$

Answer the following questions:

(a) (5 points) Derive the maximum likelihood estimate (MLE) for the rate parameter $\rho$. Show your work.

(b) (5 points) Derive the maximum a posteriori (MAP) estimate for the rate parameter $\rho$. show your work.

(c) (5 points) Is the Beta distribution a conjugate prior for the success probability, $\rho$, of the Geometric distribution? Why or why not?

(d) (5 points) Suppose you would like to update the Beta prior distribution and the MAP point estimation in an online fashion, as you obtain more data. Write the pseudo-code for the online update of the prior parameters. In your answer, specify the new values for the parameters of the prior.

7. (10 points) Consider a training set containing nonnegative integer samples ($x \in \mathbb{Z}_0^+$, i.e., $x \in \{0, 1, 2, 3, \dots\}$) for 3 classes, $C_1$, $C_2$ and $C_3$. The training set has 50 samples for class $C_1$, 100 for $C_2$ and 50 for $C_3$. Your goal is to train a Naïve Bayes Classifier with Geometric-distributed data likelihoods:

$$P(x|C_1) \sim \text{Geometric}\left(\rho_1 = \frac{1}{2}\right)$$
$$P(x|C_2) \sim \text{Geometric}\left(\rho_2 = \frac{1}{7}\right)$$
$$P(x|C_3) \sim \text{Geometric}\left(\rho_3 = \frac{3}{4}\right)$$

This means that you can write each data likelihood using the following equation:

$$P(x|C_i) = \rho_i(1 - \rho_i)^x$$

Answer the following questions:

(a) (3 points) Compute the prior probability for each class.

(b) (7 points) Consider the test point $x = 2$. Which class will it be assigned to? Show your work.

8. (25 points) Use the Expectation-Maximization (EM) algorithm to solve for the parameters of an Exponential Mixture Model given a set of training data $\mathbf{X} = \{x_i\}_{i=1}^{N}$, where $x_i \geq 0, \forall i$. Recall the form of the Exponential probability density function is $P(x|\lambda) = \lambda e^{-\lambda x}$ for $x \geq 0$. Answer the following questions:

   (a) (3 points) Assuming your data is i.i.d., write down the observed data likelihood, $\mathcal{L}^0$.

   (b) (3 points) Can you introduce hidden latent variables $Z$ to this problem? Describe precisely what they are.

(c) (4 points) For the hidden variables you defined above, write down the complete data likelihood, $\mathcal{L}^c$.

(d) (5 points) Write down the EM optimization function, $Q(\Theta, \Theta^t)$, where $\Theta = \{\pi_k, \lambda_k\}_{k=1}^K$. Your final solution should contain the sum of simple (and simplified) log-terms.

(e) (5 points) Derive the update equations for the parameters $\lambda_k$.

(f) (5 points) Derive the update equations for the parameters $\pi_k$.

**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!!!**