Indiana University Bloomington

Spring-2024
CSCI-B455
PRINCIPLES OF MACHINE LEARNING
Midterm Examination

March 04, 2024, Monday, 3:00 p.m. – 4:15 p.m.

Name & Surname	
University ID	
Signature	

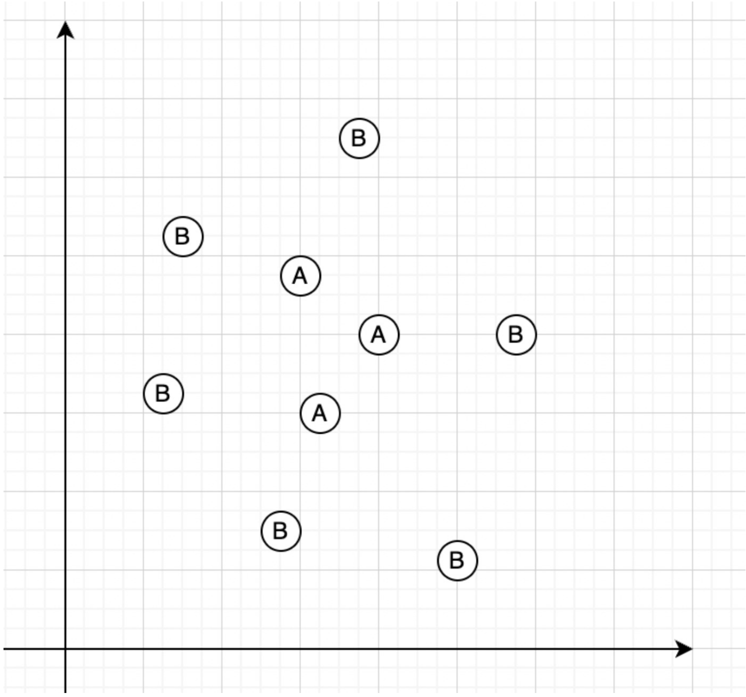
Rules:

1. There are 10 questions in this examination, each question carries 10 marks.
2. Duration of the exam is 75 minutes.
3. Write your name and surname on every page at the designated positions.
4. Put your ID card on your desk so that the proctors can check your identity.
5. The use of lecture notes, books, and any other resources, computers, mobile phones, and any digital equipment is prohibited.
6. Every student taking this examination is subject to the university discipline code. Any act or attempt of cheating, including helping others, will be considered a violation of the code.

Name & Surname:

1. Below is the 2D binary classification graph where we have 2 types of classes as the positives, labeled by A, and negatives, labeled by B. As a data scientist, we want to draw a RECTANGULAR hypothesis to categorize the data.

   a) Please draw the **most specific** and the **most general** RECTANGULAR hypothesis which will separate data points of class A from class B.



   b) Discuss the risks when you choose the most specific or the most general hypothesis.

   Hint: Consider ***false positives*** *and* ***false negatives****.*

2. Answer the following questions.

   a) Elaborate on the concepts of **overfitting** and **underfitting** in machine learning models.

   b) Elaborate on how can you detect overfitting and underfitting using **training**, **testing**, and **validation** datasets.

   c) Define **bias** and **variance** and explain what high and low values of bias and variance indicate about the concepts in part a.

   d) How does increasing the complexity of **regression** models impact the bias and variance?

3. Imagine you are a wildlife biologist studying a certain species of bird. Your research focuses on a disease that affects these bird species. You discover that the disease has an occurrence probability of 0.3. This disease can be detected through a specialized diagnostic test. The diagnostic test is highly accurate, correctly identifying the presence of the disease in 85% of birds that are actually infected. However, there is a small chance of false positives, where the test incorrectly reports a healthy bird as being infected. This false positive rate is 5 in 100 birds.

   Given this scenario, your task is to determine the **probability that a bird is infected with the disease when the diagnostic test results come back positive**.

4. In the context of decision-making with varying risks, consider a scenario where the costs of misclassification differ for two classes, C1 and C2.

   The loss matrix is defined as follows: $\lambda_{1,1} = 0, \lambda_{2,2} = 0, \lambda_{1,2} = 8, \lambda_{2,1} = 6$.

   a) Given the loss matrix, calculate the risks associated with deciding on Class C1 and Class C2.

   b) What would be the optimal decision rule to choose Class C1 (α_1) over Class C2 (α_2) in terms of $P(C_1|X)$?

   Formulas:

   α_i: classifying input in class C_i ,

   $\lambda_{i,j}$: incurred loss when input is classified as C_i, while it is C_j,

   Risk when we decide on C_i:

$$R(\alpha_i|X) = \sum_{j=1}^{K} \lambda_{i,j} P(C_j|X)$$

   Choose C_i when $R(\alpha_i|X) = min_k R(\alpha_k|X)$

5. The relative square error is defined as

$$E_{RSE} = \frac{\sum_t [r^t - g(x^t|\theta)]^2}{\sum_t (r_t - \bar{r})^2}$$

,where r^t is the given output of the observation x^t, $\bar{r}$ is the avergae of all r^t on the training data, and $g(x^t|\theta)$ is the value returned by our regression function. What **insights** can you provide about the model's performance when the Relative Squared Error (RSE) is close to 0, is close to 1, and is greater than 1? Please **explain** the rationale behind these observations.

6. a) What are the steps that can be taken when dealing with missing data in the training set?

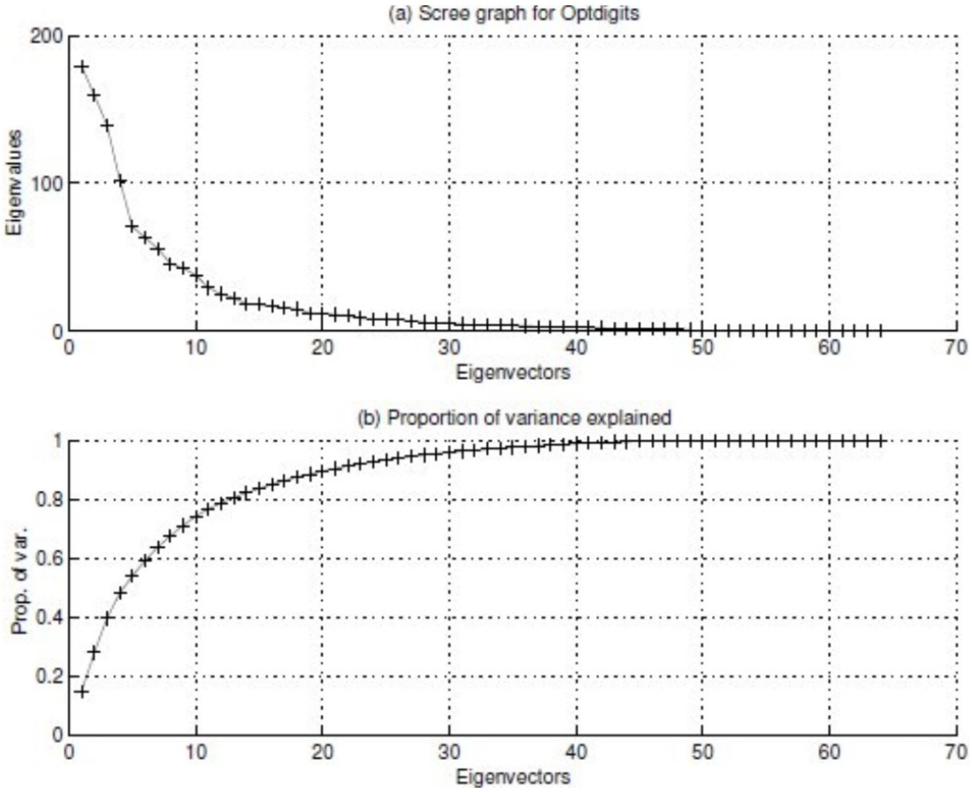
   b) How will you deal with numerical and categorical missing values?

   c) How would the presence of missing data impact the bias and variance of predictive models?

7. You have a dataset X with N samples and d features. You want to apply PCA (Principal component analysis) to this dataset for dimensionality reduction and data visualization.

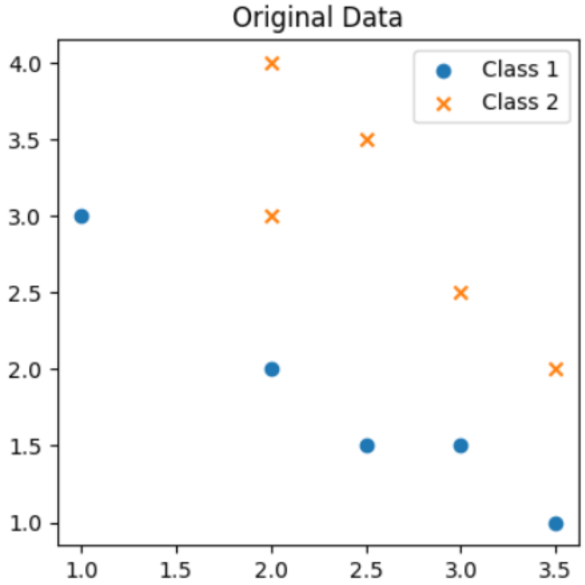
   a) Describe the steps to compute the principal components of the dataset X.

   b) Suppose the eigenvalues obtained from the covariance matrix of X are $\lambda_1, \lambda_2, ..., \lambda_d$, sorted in decreasing order, express the formula to calculate the proportion of variance explained by the first k principal components. Additionally, examine the given scree graph and provide a good number of dimensions that can be considered and state why?


(a) Scree graph for Optdigits

(b) Proportion of variance explained

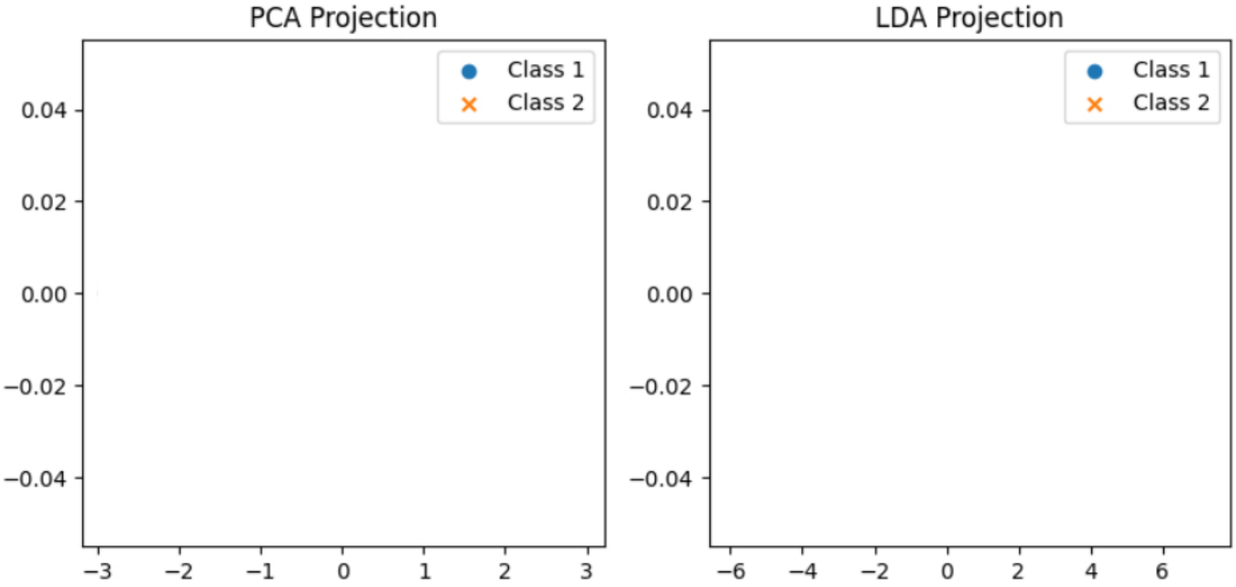
8. Given below is the plot of a two-dimensional synthetic data.



The empty plots represent the PCA and LDA projections along these directions.

a) Draw an estimate projection of the respective plots and explain the reasoning behind it.

b) Which projection would facilitate the classification of the two classes present in this synthetic data more effectively?



9. **Customer Segmentation in E-commerce with Hidden Group Information**

   Suppose you are an e-commerce company with a diverse customer base, and you want to understand the purchasing behavior of your customers. There are two distinct groups of customers, but their segmentation information is hidden. You suspect that there are two underlying segments, each characterized by different shopping preferences and spending patterns.

   Even though you initially didn't have information about which segment each customer belongs to, please provide the steps of the strategy that you can implement to accommodate the missing information for computing means and variances to gain insights into the shopping preferences and spending patterns of customers in each segment.

10. In multivariate classification involving k classes and d dimensions (columns), the number of parameters required is computed using the formula $k \times \frac{d(d+1)}{2}$. However, we know that some assumptions can be applied to reduce the number of parameters to be computed.

    Explain **two specific assumptions** that justify the reduction of parameters needed for multivariate classification, considering the reduced parameter count as $\frac{d(d+1)}{2}$, d, or 1.

    Each class in the provided multivariate data follows the normal distribution. The probability that x belongs to class C_i is given as

$$P(x|C_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} exp[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)]$$