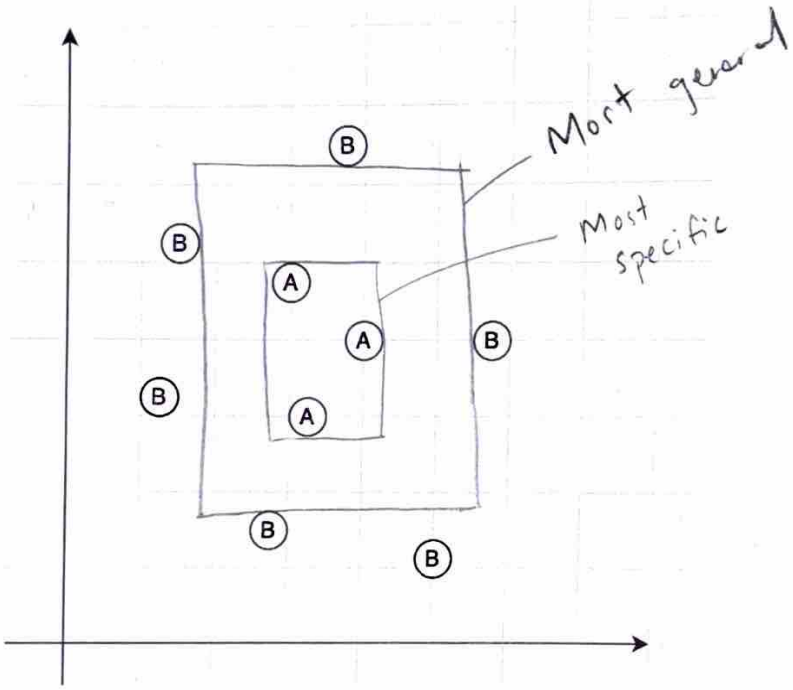
Name & Surname:

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**.*

Chosing a General Hypothesis puts us in a situation where we are more likely to accept values that are false in our rectangle. These are false positive, a value that is actually false but according to the hypothesis is a true value.

Conversely, if we pick a specific hypothesis we may end up classifying values that are true as false because the hypothesis (rectangle) was much smaller.
These are false negatives, values that are actually true but are shown as false by our classifies.

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?

a) Overfitting is when our model is too complex for our data to a point where predictions are incorrect as our model just spits out memorized data. Underfitting is when our model is not complex enough and it fails to capture the relationship between our features and our target.

b) Splitting the data into the training, testing and validation is very useful in catching overfitting or underfitting. Once the model is trained on our training data set, we can subject it to our validation data set and see how it performs on unseen data. If the model performs well on the training data set and performs poorly on the validation set, it might be overfitting. If it performs poorly on both it might be underfitting. The validation set allows us to fine tune our model before we put it through testing data. The testing data is imperitive as it is a way of checking how the model performs in the real world and acts as a final check.

c) bias: The error in a model's capability to make actual predictions due to being too simple

variance: The sensitivity of the model to cause incorrect predictions from learning the noisy data.

High and low bias indicates high or low levels of underfitting
High and low variance indicates high or low levels of overfitting.

d) As complexity of a regression model increases:

Bias decreases
Variance increases.

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.**

$P(d=1) = 0.3$, $P(d=0) = 0.7$

$P(t=1 \mid d=1) = 0.85$, $P(t=1 \mid d=0) = 0.05$

$P(d=1 \mid t=1) = ?$

$$P(d=1 \mid t=1) = \frac{P(t=1 \mid d=1) \cdot P(d=1)}{P(t=1)}$$

$$= \frac{P(t=1 \mid d=1) \cdot P(d=1)}{P(t=1 \mid d=1) \cdot P(d=1) + P(t=1 \mid d=0) \cdot P(d=0)}$$

$$= \frac{0.85 \times 0.3}{0.85 \times 0.3 + 0.05 \cdot 0.7}$$

$$= \frac{0.255}{0.295} \approx \boxed{0.87}$$

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 $R(\alpha_i|X) = min_k R(\alpha_k|X)$

a) C1 :

$$R(\alpha_1|x) = \lambda_{1,1} \cdot P(C_1|x) + \lambda_{1,2} \cdot P(C_2|x)$$
$$= 8 \cdot (1 - P(C_1|x))$$

C2 :

$$R(\alpha_2|x) = \lambda_{2,1} \cdot P(C_1|x) + \lambda_{2,2} \cdot P(C_2|x)$$
$$= 6 \cdot P(C_1|x)$$

b) Chosing Class C1 (α_1) over Class C2 (α_2)

$$\Rightarrow R(\alpha_1|x) < R(\alpha_2|x)$$
$$\Rightarrow 8 \cdot (1 - P(C_1|x)) < 6 \, P(C_1|x)$$
$$\Rightarrow \frac{4}{3} - \frac{4}{3} P(C_1|x) < P(C_1|x)$$
$$\Rightarrow \frac{4}{3} < \frac{7}{3} P(C_1|x)$$
$$\Rightarrow \frac{4}{7} < P(C_1|x)$$

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.

When RSE is close to 0 :

This means the standard error is very small relative to the estimate.

This implies that prediction is quite precise.

When RSE is closer to 1 :

This implies the standard error is almost the same as the estimate.

This implies that the prediction is not as precise

When RSE is greater than 1 :

This implies that the standard error of the estimate is greater than the estimate itself.

Yeilding the most inaccurate estimation.

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?

a) The steps we can take are :
   1) drop all missing values
   2) Fill all missing values with imputation techniques :
      ie filling the missing values with the average of the given values.

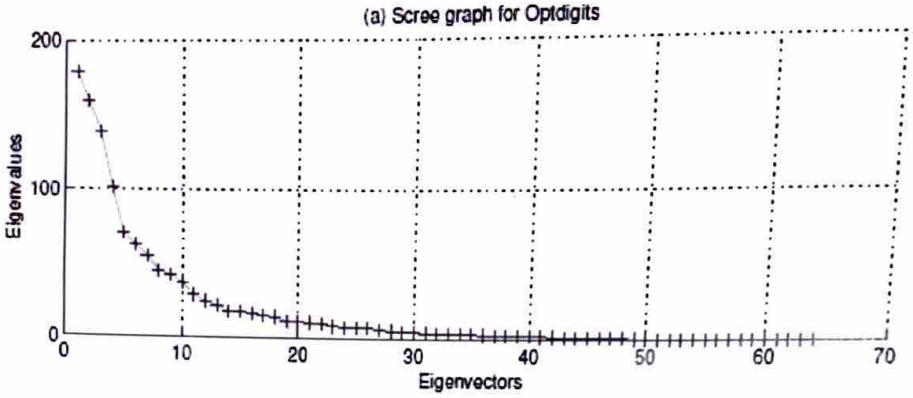
b) Coming to numerical, I would either drop it or impute the missing values with the average.
   If the category is missing then I would just delete them.

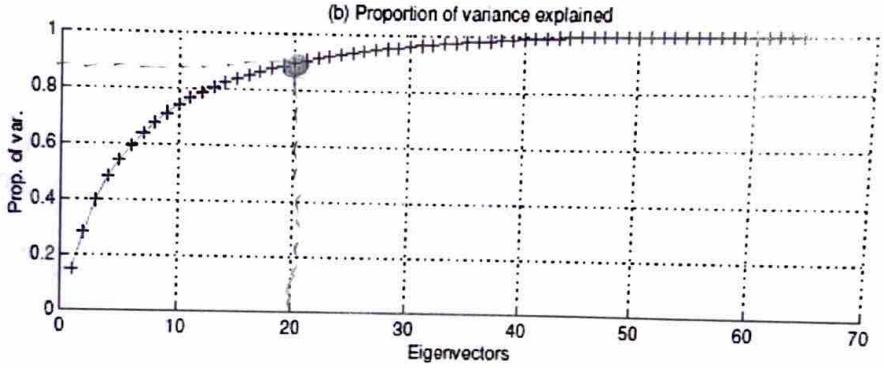
c) It is clear that having more data always makes our model better and more accurate. So having less data might make our model more likely to learn from noise. Thus making the variance higher. Or our model might cause more errors since there is not enough data to learn from. Thus increasing bias

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 $\lambda_1, \lambda_2, \ldots, \lambda_d$, 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?



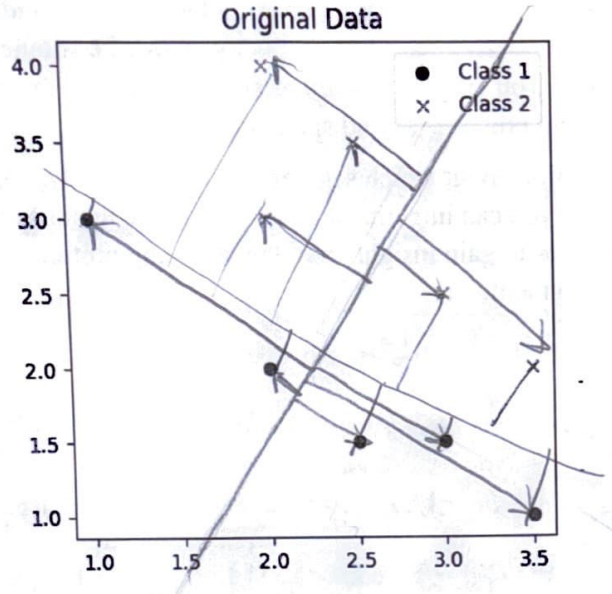


a)
- Compute the covariance Matrix of X
- Compute the eigenvalues and eigenvectors of the covariance Matrix
- eingenvalues must be sorted in decreasing order
- The principal components of X are the eigenvectors corresponding to the eigenvalues.

b) formula to calculate proportion of variance $= \dfrac{\lambda_1 + \lambda_2 + \lambda_3 \ldots \lambda_k}{\sum_{i=1}^{d} \lambda_i}$

20 dimensions can be considered as this is where the elbow corresponds to the eigenvectors. Thus, it takes 20 eigenvectors that build up the proportion of variance to about a 90%

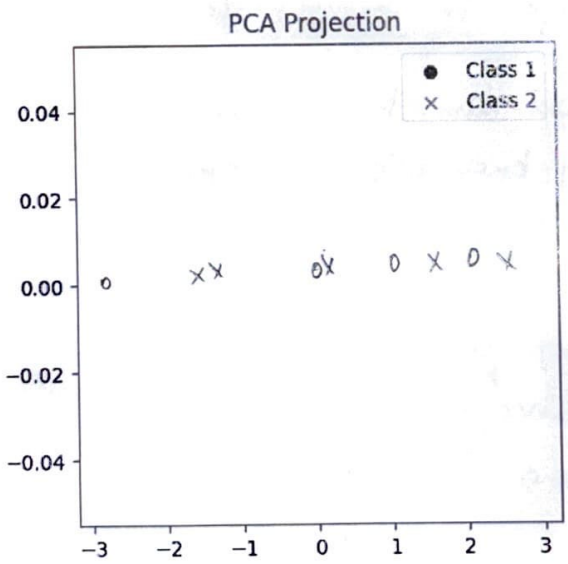
8. Given below is the plot of a two-dimensional synthetic data.

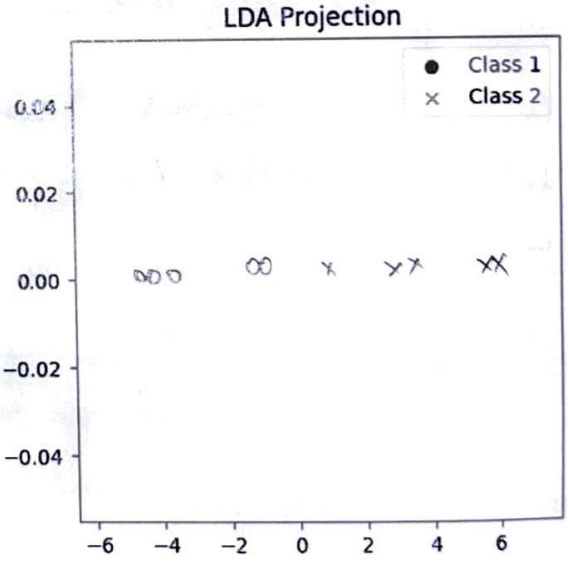


The empty plots represent the PCA and LDA projections along these directions.

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?





a) PCA Graph was made by chosing a line that yeilds the most vast range

b) Chosen the line that yeilds the biggest gap for the clearest classification.

b) LDA Projection

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.

I think that I would solve this problem using clustering. Since the segmentation information is hidden, I would start off by gathering data from the customer regarding frequency of spending, total amount spent to build my dataset.

I can now start clustering my data points in hopes to identify patterns and trends that will allow me to also start filling up more and missing data in hopes to form better clusters.

Once I have clusters, I would aim to finally create segment labels and bridge the gap of knowledge we started out with.

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

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

Assumption 1 : For the number of parameters to be equal to $\frac{d(d+1)}{2}$, the covariance matrix must be shared and the covariance matrix $S_1 = S$

Assumption 2 : For the number of parameters to be equal to d the assumption must be that the covariances are again shared and they are graphically depicted as hyper elipsoidal