CSE 575 - ASSIGNMENT 1

TEJA KUNDERU

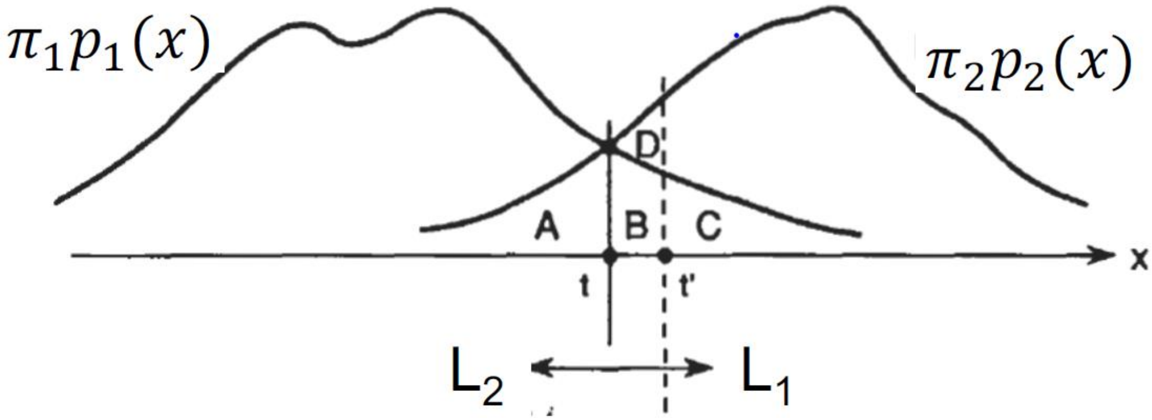
1211384158

Question 1:

If P(Y = 1 | X) = s and the label assigned is 1, the risk in this classification or the probability that the actual labelling is 2 is (1 - s).

Bayes risk is always the minimum of (P(Y = 1 | X), P(Y = 2 | X)) = risk r(X)

Bayes error is the expected risk i.e. E(r(X))



The Bayes error = E[r(X)] = $\int r(x)p(x)dx$

$= \int \min(\pi_1 p_1(x), \pi_2 p_2(x))\, dx$

$= \pi_1 \int_{L_1} p_1(x)dx + \pi_2 \int_{L_2} p_2(x)dx$

$= \pi_1 \varepsilon_1 + \pi_2 \varepsilon_2$

Everything from the left of the intersection point of the curves is L_2 and everything to the right is L_1.

The minimum bayes error when the classification is 1 is given by the area under the curve $\pi_2 p_2$ in the region L2.

If the point B is a little bit to the right, the area under the curve in the L_2 region increases. This also increases the error. But such a state is not possible as bayes classifier selects the class with minimum risk.

Thus bayes classifier is the optimal classifier.

Question 2:

1) MLE estimator of μ_{MLE} is the μ for which $x_1, x_2, \ldots.. x_N$ is most likely.

$$\mu_{MLE} = \max\left(P(x_1, x_2, \ldots\ldots x_N \mid \mu, \sigma^2)\right)$$

$$= \max\left(\prod_{i=1}^{N} P(x_i \mid \mu,\ \sigma^2)\right)$$

Now, to find out the max we apply logarithmic function to the equation. We then differentiate with respect to μ and equate it to 0.

After solving we get,

$$\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

2) The values of x used in the calculation of the MLE estimator of μ are samples drawn randomly from the continuous Gaussian distribution $N(\mu, \sigma^2)$. Hence the MLE estimator is also a random variable.

3) The bias of μ_{MLE} is the difference between the expected value of μ_{MLE} and μ

Therefore the estimator is unbiased if the difference between them is 0, which implies that the expected value of the estimator of μ must be equal to μ

$$E(\mu_{MLE}) = E\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \frac{1}{N}\sum_{i=1}^{N} E(x_i) = \frac{1}{N}\sum_{i=1}^{N} \mu = \frac{1}{N}\ x\ N\mu$$

Therefore, $E(\mu_{MLE}) = \mu$

Thus the estimator of μ is unbiased.

4) The estimator of σ^2 is

$$\sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2,$$ Where the true value of μ is known.

Using a similar approach to the last problem, we calculate the expectation of the estimator first.

$$E(\sigma^2_{MLE}) = E(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2)$$

$$= E((\frac{1}{N} * \sum(x_i^2 - 2x_i\mu + \mu^2))$$

$$= \frac{1}{N} * \sum(E(x_i^2) - 2E(x_i)\mu + \mu^2)$$

$$= \frac{1}{N} * \sum\left(E(x_i^2) - E^2(x_i)\right) + \frac{1}{N} * \sum E^2(x_i) - \frac{1}{N} * \sum 2E(x_i)\mu + \frac{1}{N} * \sum \mu^2)$$

$$= \frac{1}{N} * N * \sigma^2 + \frac{1}{N} * N * \mu^2 - \frac{1}{N} * 2\mu^2 * N + \frac{1}{N} * N * \mu^2$$

$$= \sigma^2$$

Therefore, the estimator of parameter σ^2 is unbiased. The expectation is equal to σ^2 itself and the bias is zero.

Question 3:

1) There are 13 independent parameters:
    - P(X1 = Sunny | Y = Yes)
    - P(X1 = Sunny | Y = No)
    - P(X2 = Warm | Y = Yes)
    - P(X2 = Warm | Y = No)
    - P(X3 = High | Y = Yes)
    - P(X3 = High | Y = No)
    - P(X4 = Strong | Y = Yes)
    - P(X4 = Strong | Y = No)
    - P(X5 = Cool | Y = Yes)
    - P(X5 = Cool | Y = No)
    - P(X6 = Change | Y = Yes)
    - P(X6 = Change | Y = No)
    - P(Y = Yes)

There are 6 features to consider in the data set. Each feature of X is binary. We only need to find one of the two parameters in a feature and the other can be determined from the first as the parameters sum up to 1.

We are trying to determine the class label so we need to estimate the parameters in each feature while taking Y into consideration.

Each parameter in the feature then has two possibilities because of Y being a binary class label. This gives a total of 12 independent parameters. The parameters are independent because of the naïve bayes assumption that the features are independent of each other given a class.

2) The estimations for the above parameters using MLE:

- P(Y = Yes) = $\frac{3}{4}$
- P(Y = No) = $\frac{1}{4}$

- P(Sky = Sunny) = P(Sky = Sunny | EnjoySpt = Yes) x P(EnjoySpt = Yes)

  +

  P(Sky = Sunny | EnjoySpt = No) x P(EnjoySpt = No)

  P(Sky = Sunny) = 1 x $(\frac{3}{4})$ + 0 x $(\frac{1}{4})$ = $\frac{3}{4}$

- P(Sky = Rainy) = 1 – P(Sky = Sunny) = 1 - $\frac{3}{4} = \frac{1}{4}$

Similarly,

- P(Temp = Warm) = 1 x $(\frac{3}{4})$ + 0 x $(\frac{1}{4})$ = $\frac{3}{4}$
- P(Temp = Cold) = 1 – $(\frac{3}{4}) = \frac{1}{4}$
- P(Humid = High) = $(\frac{2}{3})$ x $(\frac{3}{4})$ + 1 x $(\frac{1}{4})$ = $\frac{3}{4}$
- P(Humid = Normal) = 1 – $(\frac{3}{4}) = \frac{1}{4}$
- P(Wind = Strong) = 1 x $(\frac{3}{4})$ + 1 x $(\frac{1}{4})$ = 1
- P(Wind = Normal) = 1 – 1 = 0
- P(Water = Warm) = $(\frac{2}{3})$ x $(\frac{3}{4})$ + 1 x $(\frac{1}{4})$ = $\frac{3}{4}$
- P(Water = Cool) = 1 – $(\frac{3}{4}) = \frac{1}{4}$
- P(Forecast = Change) = $(\frac{2}{3})$ x $(\frac{3}{4})$ + 0 x $(\frac{1}{4})$ = $\frac{1}{2}$
- P(Forecast = Same) = 1 – $(\frac{1}{2}) = \frac{1}{2}$

3) P(Y = Yes | x) where x = (Sunny, Warm, High, Strong, Cool, Change)

P(Y = Yes | x) = $\frac{P(x \mid Y = Yes)\ x\ P(Y = Yes)}{P(x)}$

P(x | Y = yes) = P(X1 = Sunny, X2 = Warm, X3 = High, X4 = Strong, X5 = Cool, X6 = Change | Y = Yes)

= $\prod_{i=1}^{6} P(Xi \mid Y = Yes)$ because of the naïve bayes assumption and conditional independence.

= P(X1 = Sunny | Y = Yes) x P(X2 = Warm | Y = Yes) x P(X3 = High | Y = Yes) x P(X4 = Strong | Y = Yes) x P(X5 = Cool | Y = Yes) x P(X6 = Change | Y = Yes)

= 1 x 1 x ($\frac{2}{3}$) x 1 x ($\frac{1}{3}$) x ($\frac{1}{3}$)

= ($\frac{2}{27}$)

P(x | Y = Yes) x P(Y = Yes) = ($\frac{2}{27}$) x ($\frac{3}{4}$) = ($\frac{1}{18}$)

P(x) = $\prod_{i=1}^{6} P(Xi \mid Y = Yes)\ x\ P(Y = Yes) + \prod_{i=1}^{6} P(Xi \mid Y = No)\, x\, P(Y = No)$

= P(x | Y = Yes) x P(Y = Yes) + P(x | Y = No) x P(Y = No)

= ($\frac{2}{27}$) x ($\frac{3}{4}$) + (0 x 0 x 1 x 1 x 1 x 1) x ($\frac{1}{4}$)

= ($\frac{1}{18}$)

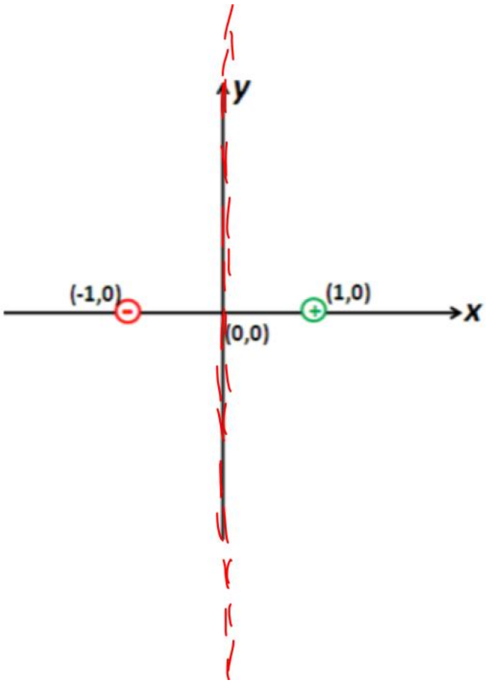
Therefore, P(Y = 1 | x) = ($\frac{1}{18}$) / ($\frac{1}{18}$)

= 1

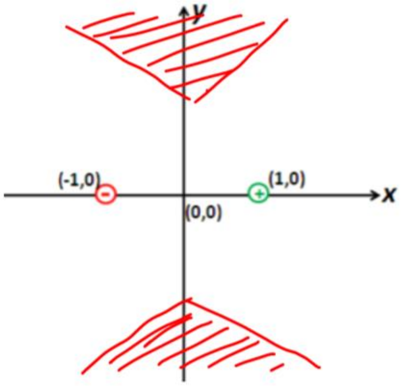
The classifier will always assign a class label of 1 (i.e., EnjoySpt - Yes) since the probability of Y = 1 for the given example is 1.

Question 4:

If we use the L_2 distance, the decision boundary is along the Y-axis.



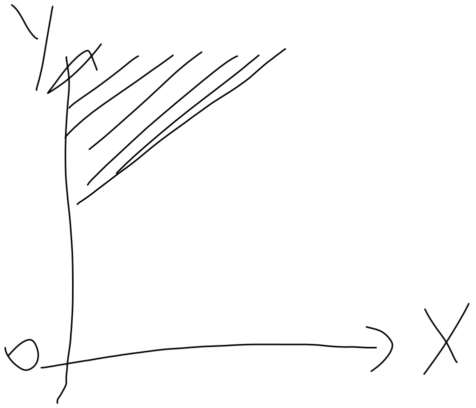
If we use the L_∞ distance instead, the decision boundary looks like –



This is because for L_∞

$D_1 = \max(|x - 1|, |y|)$

$D_2 = \max(|x - 1|, |y|)$



Decision boundary is where D1 = D2

Consider points in the first quadrant with $x > 0$ and $y > 0$

a) If $y > x + 1$

Then y is also greater than $x - 1$

$Y > x + 1 > x - 1$ and $x > 0, y > 0$

This would look like this –

This will be the same for all four quadrants because of the mod operator being used in the calculation.

b) If $y < x + 1$

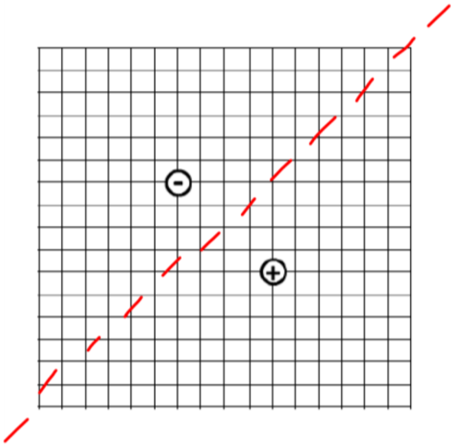
This means $x + 1 > y$ and $x + 1 > x - 1$

Therefore the max value will always be D_2

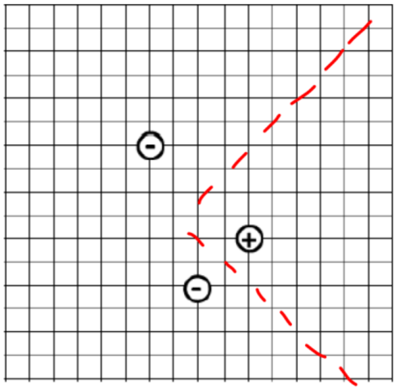
This is the reason for the particular appearance of the Decision boundary.

Question 5:

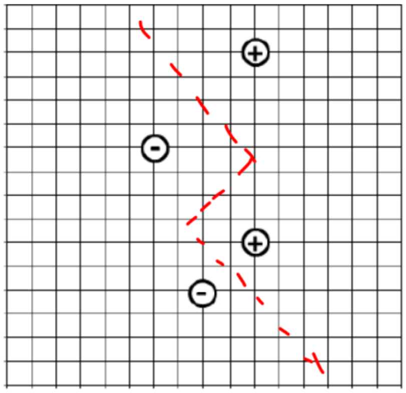
a)



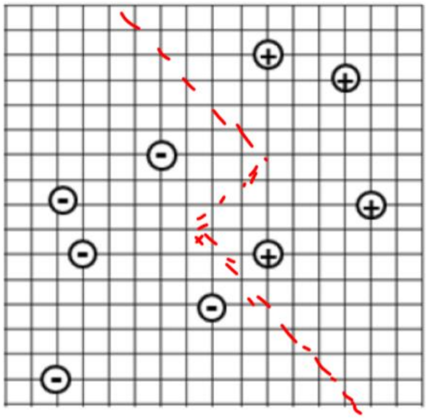
b)



c)



d)

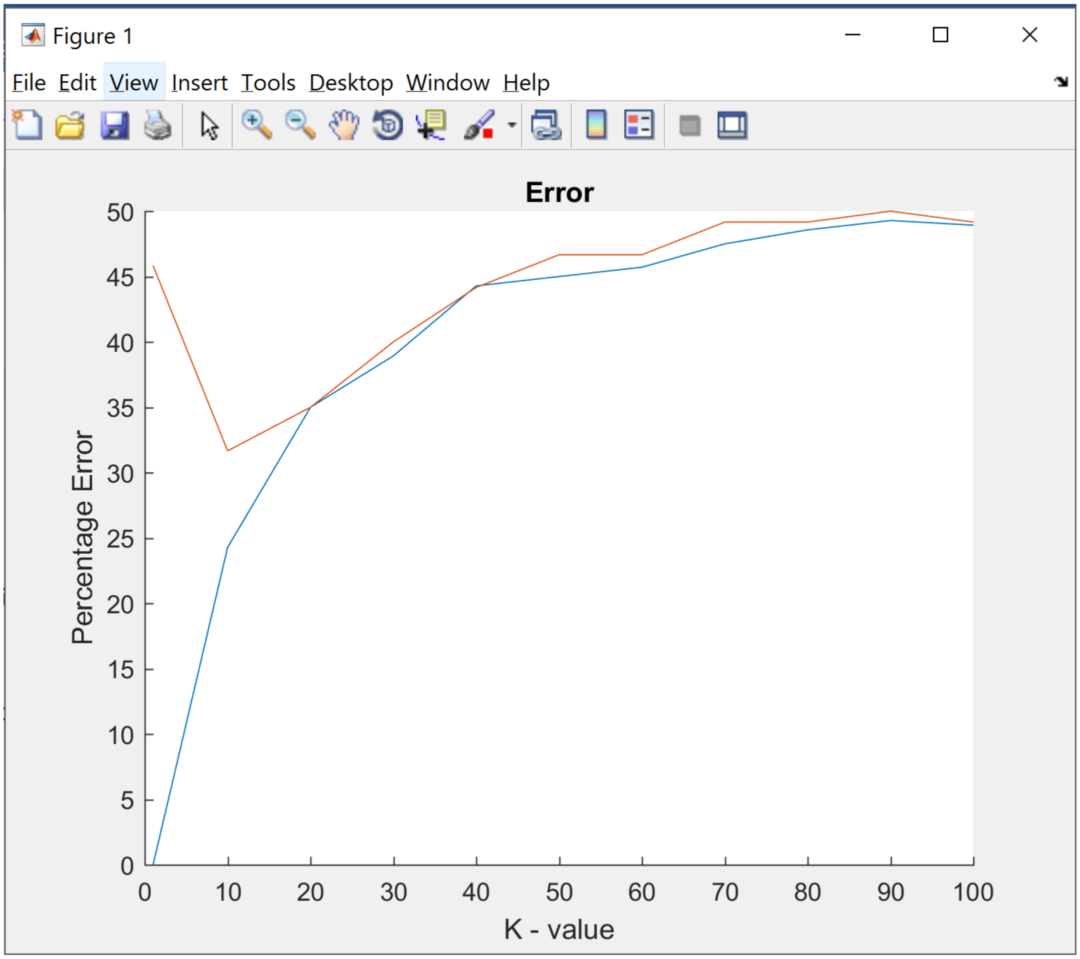


Question 5:

1) Plot:

Blue line indicates Training Error

Red line indicates Testing Error



2) Pseudo Code:

- The cosine distances are calculated for each row in the training data with every other row.
- This is done for each row in testing data as well and the distances are computed by comparing with the rows from the training data.
- The labels for the distances from the training labels are then appended to the distance matrix

- These distances are then sorted in the descending order.
- A specific number of distances are then chosen based on the value of K and their labels are compared to those from the training or testing data depending on the test running.
- The number of labels that match are kept track of and summed up over all the rows of the data.
- This is the number of labels that match with the dataset for a specific value of K.
- These are subtracted from the total number of rows to calculate the number of labels that were erroneous.
- The erroneous number gives us the percentage error for a value of K.
- The basic code is iterated 11 times as there are 11 values of K and plotted on a graph.

3) The value of K that minimizes the error in case of training does not minimize the error for testing.
A good value of K is when both types of errors increase with K.
Selecting the training error to choose the value of K can lead to underfitting and overfitting.
So the best way to choose K is to use the evaluation data.