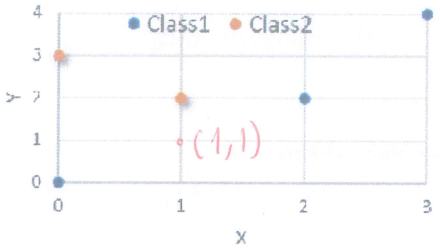
Name Surname: **CENG 463 Machine Learning - Midterm Exam** Signature:

Q1) A dataset of labeled 2-D points are given in the below figure. Classify a newcoming point with coordinates (X, Y) = (1,1) by using K-Nearest Neighbor algorithm with (a) K=1 and (b) K=3. Show your computation steps by using the Euclidean distance (20 points).



Distance $= D(0,0) = \sqrt{1^2 + 1^2} = \sqrt{2}$

$D(1,2) = \sqrt{(1-1)^2 + 1^2} = 1$

$D(2,2) = \sqrt{(2-1)^2 + (2-1)^2} = \sqrt{2}$

$D(0,3) = \sqrt{(1-0) + (3-1)^2} = \sqrt{1+4} = \sqrt{5}$

$D(3,4) = \sqrt{(3-1)^2 + (4-1)^2} = \sqrt{13}$

Order: D(1,2), D(0,0), D(2,2), D(0,3), D(3,4)
Class 2, Class 1, Class 1, Class 2, Class 1

a) K=1 ⇒ Class 2

b) K=3 ⇒ Class 1 by majority voting

Q2) Using above data fill values of a design matrix and write equations to solve $Y = F(X)$ linear regression ($Y = B0 + X * B1$) (15 points).

$B = (X^T X)^{-1} X^T Y$

For Class 2:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \quad B = \begin{bmatrix} B_0 \\ B_1 \end{bmatrix} = \left(\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

Q3) Compute the likelihood of selecting banana having the properties of yellow, long and sweet by using a naïve bayes (25 points).

	Yellow	Long	Sweet	Total
Apple	5	0	5	10
Banana	3	3	4	10
Other	2	2	6	10

P(Yellow | Banana) = 3/10, P(Banana) = Prior = 10/30
P(Long | Banana) = 3/10, P(Yellow) = 10/30
P(Sweet | Banana) = 4/10 P(Long) = 5/30
P(Sweet) = 15/30

$$P(\text{Banana} \mid \text{Long, sweet and Yellow}) = \frac{P(\text{Long} \mid \text{Banana}) * P(\text{sweet} \mid \text{Banana}) * P(\text{Yellow} \mid \text{Banana}) * P(\text{Banana})}{P(\text{Long}) \cdot P(\text{Sweet}) \cdot P(\text{Banana})}$$

$$= \frac{4/10 \cdot 3/10 \cdot 3/10 \cdot 10/30}{10/30 \cdot 5/30 \cdot 15/30} = \frac{\frac{4}{10} \cdot \frac{3}{10} \cdot \frac{3}{10} \cdot \frac{10}{30}}{\frac{10}{30} \cdot \frac{5}{30} \cdot \frac{15}{30}} = \frac{36}{750} = 0.048$$

Q4) Give short answers to the following questions with 1-2 sentences (16 Points).

a) Why Cross-validation is used?

Reduce the bias that may be caused random training selection

b) When do you use regression instead of classification?

We use regression for quantitative estimation. Classification estimates categorical data.

c) What are the major 3 metrics of regression to control quality of fit? In which order do you use them? R-square (R2), Mean Square Error, Mean Absolute Error, Residual Sum of Squares, Total Sum of Squares

d) Which metric is most useful to measure high variability in regression? Mean Square Error or Mean Absolute Error?

Mean Square Error is better suited to measure high variability.

e) Write down a second order polynomial of two variables Z = F(X,Y) with co-linearity term.

$Z = B_0 + B_1X + B_2Y + B_3XY + B_4X^2 + B_5Y^2$

f) What is the advantage of Lasso regularization over the Ridge Regularization?

Lasso use L1 regularization and converges coefficients faster & exactly to 0

g) What are the minimum number of samples to solve a 3rd order polynomial regression?

min number of samples = number of coefficients = (order + 1)

h) Give an example of normalizing input data to the same range.

Let X be data, Normalized data $= \frac{X - \min(x)}{\max(x) - \min(x)}$ or $\frac{X}{\max(x)}$

Q5) You have designed a cancer test and your findings with 120 test subjects are measured as a confusion matrix below (24 points).

a) Compute TP, FP, FN, TN, Precision, Recall, Accuracy, F-Score.

		Predicted	
		Cancer	Healthy
Actual	Cancer	15	5
	Healthy	10	90

TP = 15, FP = 10, FN = 5, TN = 90

Precision = $TP/(TP+FP) = 15/25$; Recall = $TP/(TP+FN) = 15/20$

Accuracy = $(TP+TN)/(TP+FP+TN+FN) = 105/120 = \frac{7}{8} = 0.875$

F-Score $= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{\frac{3}{5} \cdot \frac{3}{4}}{\frac{3}{5} + \frac{3}{4}} = \frac{2 \cdot 3 \cdot 3}{12 + 15} = \frac{18}{27} = \frac{2}{3}$

b) What are the possible consequences of making an error (FP and FN) in classification? How would you improve your classifier?

If the test cannot identify cancer patient, they have high risk while others falsely identified as cancer they may retake the test or take unnecessary treatment. Test should have near zero False Negatives. We may add more samples.

c) Based on this cancer test example, which metric(s) (precision, recall, accuracy or F-score) would you use to evaluate your classifier? Explain your reasoning.

Recall and F-Number are more suitable than other metrics such as accuracy which is 87.5%