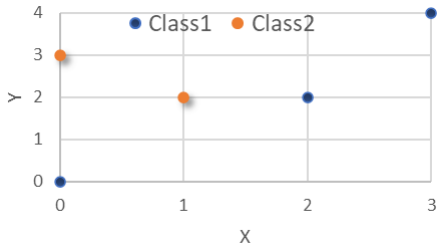
Name Surname: **CENG 463 Machine Learning - Midterm Exam** Signature:

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).



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).

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

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

a) Why Cross-validation is used?

b) When do you use regression instead of classification?

c) What are the major 3 metrics of regression to control quality of fit? In which order do you use them?

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

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

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

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

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

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

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

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.