Machine Learning - Quiz 2

Name: _____

Directions: Write complete solutions with enough detail so that your reasoning is clear to Prof. Chakraborty.

Question 1 [1 + 2 + 4 = 7 points]

Given below is a confusion matrix for a classifier evaluated on a test dataset. The class of interest is `A`, that is `A` is the positive class.

		Predicted	
		A	B
Reference	A	118	28
	B	17	37

For the following questions, **show your work to receive full credit.**

(a) How many observations were there in the test dataset?

(b) Report the accuracy of the classifier.

(c) Report the recall and precision of the classifier.

Question 2 [2 + 3 + 2 = 7 points]

Consider the toy dataset below.

Obs.	Y	X
1	A	1.0
2	A	1.8
3	B	3.2
4	A	4
5	B	5
6	B	5.8

The Euclidean distance between two p-dimensional vectors $\mathbf{a} = (a_1, a_2, \ldots, a_p)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_p)$ is

$$||\mathbf{a} - \mathbf{b}||_2 = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_p - b_p)^2}$$

(a) Consider $K = 1$. Predict the class for a test data point with $X = 4.5$. **Explain how you arrived at your prediction.**

(b) Consider $K = 3$. Predict the class for a test data point with $X = 3.7$. **Explain how you arrived at your prediction.**

(c) Explain why standardizing the predictor X is not necessary for the dataset above.

Question 3 [3 points]

For the KNN approach, is a large or small K more prone to overfitting? Is a large or small K more prone to underfitting? **Explain your answer in terms of the bias-variance trade-off.**

Question 4 [3 points]

Prof. Chakraborty is working on a classification problem. He takes a data set, divides it into equally-sized training and test sets, and then tries out two different classification procedures. First, he uses logistic regression and gets an error rate of 23% on the training data and 27% on the test data. Next, he uses 1-nearest neighbors (i.e. $K = 1$) and gets an average error rate (averaged over both test and training data sets) of 18%. Based on these results, which method should he prefer to use for classifying new (unseen) observations? **Explain your answer.**