MIDTERM FOR CAP4770 - PART 2

P2 (5 points). Suppose you sampled the height of four friends, and got the measurements $x_1 = 175$ cm, $x_2 = 172$ cm, $x_3 = 180$ cm and $x_4 = 185$ cm. What is the best single number summary of height of the four friends that minimizes the mean squared error? That is, what is the number $\hat{c}$ that minimizes

$$\frac{1}{4}\left((x_1 - c)^2 + (x_2 - c)^2 + (x_3 - c)^2 + (x_4 - c)^2\right),$$

over all possible numbers c?

P3 (5 points). Suppose the input X and the output Y are related by $Y = \log X + \varepsilon$. Here, X and Y are random, and ε is a random noise independent of X and with mean 10. What is the function $\hat{f}$ that gives you the best prediction of the output with respect to the square error loss? That is, $\hat{f}$ minimizes $E[(Y - f(X))^2]$ over all functions f.

P4 (5 points). Discuss in a few sentences why K-Nearest-Neighbor may not work well when the data has a large number of features.

P5 (5 points). Is stratified sampling always needed? Discuss in a few sentences why or why not.

P6 (5 points). Suppose you are not sure how long you will rent your apartment: either one year or two years with equal probabilities. Suppose the rate of price inflation in the next year is uncertain: It could be anywhere between 5% and 10% with a uniform distribution. The cost of the apartment is \$10000 this year. What is the expected total cost of renting the apartment?

P7 (5 points). You are given the following training data (input-output (x, y) pairs), sorted according to the input: $(1.2, 14.6)$, $(2.3, 36.3)$, $(3.1, 43.0)$, $(4.2, 19.6)$, $(8.3, 80.6)$, $(8.4, 89.0)$, $(9.4, 81.0)$, $(9.9, 93.1)$, $(11.2, 116.3)$, $(13.0, 134.9)$.

Now, what is the K-Nearest-Neighbor prediction for an input $x_0 = 6$ for $K = 2$? How about for $K = 4$?

P8 (5 points). You are given the following input-output (x, y) pairs, sorted according to the input: $(8.4, 89.0)$, $(9.4, 81.0)$, $(9.9, 93.1)$, $(11.2, 116.3)$.

For the above data, please give an example of regression that has a coefficient of determination, i.e., R^2, less than 0. Give the predictions from your regression and show your calculation steps and the R^2 value. You may use Python to do the actual calculations.

P9 (5 points). Suppose you are given an array of values for a scalar-valued feature: $[89.0, 81.0, 93.1, 116.3]$. You will do feature scaling on it. What does the min-max scaling give you? How about standardization? Show you calculation steps and the results. You can use Python to do the actual calculation.

P10 (5 points). Consider a categorical variable representing the modes of transportation with four categories: CAR, PLANE, TRAIN, OTHER. Suppose in the one-hot encoding, the order of new categorical variables (aka dummy variables) is the same as listed. What will be the result of one-hot encoding of the following instances?
instance 1: TRAIN
instance 2: CAR
instance 3: OTHER.

P11 (5 points). Discuss the pros and cons of 30-fold cross-validation versus 3-fold cross-validation with respect to: computation time, how prone it is to overfitting, and how confident you are about the validation score on each split.

P12 (20 points). Consider binary classification. The table below shows 20 instances and their scores assigned by the classifier. We assume 1 corresponds to the positive class, 0 to the negative class. We take the convention that the classifier declares an instance positive if and only if the instance's score is greater than or equal to the threshold value.

Actual Class	Scores
0	0.05
0	0.09
0	0.12
0	0.18
0	0.23
0	0.25
0	0.31
1	0.35
0	0.39
0	0.41
0	0.45
0	0.47
1	0.48
1	0.75
1	0.83
1	0.88
1	0.92
0	0.95
1	0.97
1	0.99

(a) When the threshold value is equal to 0.5, fill the confusion matrix below.

	Predicted Negative	Predicted Positive
Actual Negative		
Actual Positive		

(b) What are the numbers of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN)?

(c) What are the values of precision, recall, sensitivity, specificity, false positive rate, false negative rate, true positive rate and the F_1 score?

(d) What is the largest threshold value to achieve 100% recall? What is the precision at that threshold value?

(e) Show by example that the precision is not always an increasing function of the threshold.