Question 1

(Refer to the t-table here:
https://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm

Links to an external site.

)

Suppose that we want to compare two prediction models M1 and M2. We have performed 10 rounds of 10-fold cross-validation on each model, where the same data partitioning in round i is used for both M1 and M2. The error rates obtained for M1 are 30.4, 32.1, 20.7, 22.6, 31.5, 41.0, 27.5, 25.4, 21.5, 26.1. The error rates for M2 are 22.7, 14.2, 22.9, 20.3, 21.7, 22.4, 20.1, 19.1, 16.2, 32.0. Determine whether the two models' mean error rates are significantly different at the significance level of 1%.

Your Answer:

The number of degrees of freedom is the number of data points minus 1 which is 9. The critical value for a 1% significance level and 9 degrees of freedom is 2.821. The mean error rate for M1 is 27.88 and the mean error rate for M2 21.16 is so the mean of the difference is 6.72. The standard deviation of the error rates of M1 is 5.82 and the standard deviation of the error rates of M2 is 4.76 so the standard deviation of the difference is sqrt(5.82^2 + 4.76^2) = 7.52. The standard error of the differences is SD/sqrt(n) which is 2.38. The t statistic is the ratio of the mean of the difference and the

standard error of the difference which is 2.824 which is greater than the critical value of 2.821 so the difference is significant at the 1% level.

Question 2

Consider the different classification methods we have discussed in class. Name one classification method for each of the following scenarios and briefly explain why.

(a) A classification method that supports incremental data.

(b) A classification method whose classification decisions are easy to interpret.

(c) A classification method whose classification decisions are not easy to interpret and has many parameters to train.

Your Answer:

a. The Naive Bayes classifier supports incremental data because it efficiently handles new observations by updating the probability distributions to include the new data without requiring a full retraining process.

b. The classifications of a k nearest neighbors model are easy to interpret because they are simply based on the majority class closest k points to the point being classified.

c. A neural network will return classifications that are not easy to interpret and has many parameters to train because it has many nodes which all have inputs and outputs and it is impossible to trace through its training process by hand.

Question 3

Consider the following initialization of a K-Means Clustering problem (K=3 with labels [A, B, C]).

We have 8 points indexed from 1 to 8 with the following (X, Y) coordinate:

p_1 (1,2), p_2 (2,1), p_3 (3,2), p_4 (3,3), p_5 (4,1), p_6 (5,2), p_7 (5,5), p_8 (6,4)

In our initialization of K-means, we choose p_1 (1,2) as centroid A, p_4 (3,3) as centroid B, and p_7 (5,5) as centroid C.

After the first round of K-Means clustering, which cluster each point would be assigned to? what is the position of the new centroids? Show key steps of your computation.

Your Answer:

p_1 is centroid A thus is in cluster A.

p_2 is sqrt(1+1) = sqrt(2) away from centroid A, sqrt(1+4) = sqrt(5) away from centroid B, and sqrt(16+9) = 5 away from centroid C, so it's assigned to cluster A.

p_3 is sqrt(4) = 2 away from centroid A, sqrt(1) = 1 away from centroid B, and sqrt(4+9) = sqrt(13) away from centroid C, so it's assigned to cluster B.

p_4 is centroid B thus is in cluster B.

p_5 is sqrt(9+1) = sqrt(10) away from centroid A, sqrt(1+4) = sqrt(5) away from centroid B, and sqrt(1+16) = sqrt(17) away from centroid C, so it's assigned to cluster B.

p_6 is sqrt(16) = 4 away from centroid A, sqrt(1+4) = sqrt(5) away from centroid B, and sqrt(9) = 3 away from centroid C, so it's assigned to cluster B.

p_7 is centroid C thus is in cluster C.

p_8 is sqrt(25+4) = sqrt(29) away from centroid A, sqrt(1+9) = sqrt(10) away from centroid B, and sqrt(1+1) = sqrt(2) away from centroid C, so it's assigned to cluster C.

The new centroid for cluster A is ((1+2)/2,(2+1)/2) = (1.5,1.5)

The new centroid for cluster B is ((3+3+4+5)/4,(2+3+1+2)/4) = (3.75,2)

The new centroid for cluster C is ((5+6)/2,(5+4)/2) = (5.5,4.5)

Question 4

Briefly describe one data mining tool that you have used either in this course or in other settings. What did you use this tool for? What are the key strengths and possible limitations of this tool?

Your Answer:

I used logistic regression in a project for my Applied Regression class last semester to predict heart disease based on metrics like resting blood pressure and maximum heart rate. A strength of logistic regression are that it is an easily interpretable model, allowing for insight into what factors made the probability of heart disease go up or down. A limitation of logistic regression is that it's more prone to bias than a more complex model.