Data Mining HW4 Solutions

Question 1

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

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%.

Solution:

We will use the paired t-test.
Number of cases = n = 10
Degree of freedom = n - 1 = **9**
Significance level (α) = 0.01
Significance level for two tailed test = 1 - (α/2) = **0.995**

Check T-table for (9, 0.995) -> we get a critical value of T = **3.250**

Calculate difference in error rates of M1 and M2, and the mean:
d = (error rates of M1) - (error rates of M2)

The differences for each fold are as follows:

d = (30.4 - 22.7), (32.1 - 14.2), (20.7 - 22.9), (22.6 - 20.3), (31.5 - 21.7), (41.0 - 22.4), (27.5 - 20.1), (25.4 - 19.1), (21.5 - 16.2), (26.1 - 32.0)

d = 7.7, 17.9, -2.2, 2.3, 9.8, 18.6, 7.4, 6.3, 5.3, -5.9

Mean (d) = (7.7 + 17.9 + -2.2 + 2.3 + 9.8 + 18.6 + 7.4 + 6.3 + 5.3 + -5.9) / 10 = **6.72**

Now calculate the variance:

$$var(M_1 - M_2) = \frac{1}{k}\sum_{i=1}^{k}[err(M_1)_i - err(M_2)_i - (\overline{err}(M_1) - \overline{err}(M_2))]^2.$$

var(M1 - M2) = 1/10 * [(30.4-22.7-6.72)^2 + (32.1-14.2-6.72)^2 + (20.7-22.9-6.72)^2 + (22.6-20.3-6.72)^2 + (31.5-21.7-6.72)^2 + (41-22.4-6.72)^2 + (27.5 - 20.1-6.72)^2 + (25.4-19.1-6.72)^2 + (21.5-16.2-6.72)^2 + (26.1-32-6.72)^2]

= 1/10 * [537.596]
= **53.76**

Next we calculate the t-statistic:

$$t = \frac{\overline{err}(M_1) - \overline{err}(M_2)}{\sqrt{var(M_1 - M_2)/k}},$$

t = 6.72 / $\sqrt{53.76/10}$
T = **2.89**

Our t-value of 2.89 < 3.250 and so the null hypothesis can't be rejected. We conclude that there is no significant difference between error rates of the two models and any difference between the error rates of M1 and M2 can be attributed to chance.

Alternatively - Calculation illustrated as table here:

	M1	M2	Diff	Mean Error Rate Diff	E1-D1	(E1-D1)^2
	30.4	22.7	7.7	6.72	0.98	0.9604
	32.1	14.2	17.9	6.72	11.18	124.9924
	20.7	22.9	-2.2	6.72	-8.92	79.5664
	22.6	20.3	2.3	6.72	-4.42	19.5364
	31.5	21.7	9.8	6.72	3.08	9.4864
	41	22.4	18.6	6.72	11.88	141.1344
	27.5	20.1	7.4	6.72	0.68	0.4624
	25.4	19.1	6.3	6.72	-0.42	0.1764
	21.5	16.2	5.3	6.72	-1.42	2.0164
	26.1	32	-5.9	6.72	-12.62	159.2644
Sum	278.8	211.6	67.2			537.596
Mean	27.88	21.16	6.72			
Variance(M1-M2)						53.7596
Variance(M1-M2)/10						5.37596
sqrt(Variance(M1-M2)/10)						2.318611654
t						2.898286132

Alternatively: Depending on if you use the formula for the sample SD or the population SD, you could find a different T-value of **2.74.** This is because the denominator would contain (10-1) instead of (10). This should also be considered acceptable.

I believe the difference is that they used a slightly different formula to calculate the variance / SD using N-1 in the denominator instead of N. Basically, this aligns with using the formula for a sample SD instead of the population SD:

Standard Deviation Formula	
Population	**Sample**
$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X-\bar{x})^2}{n-1}}$
X – The Value in the data distribution *μ – The population Mean* *N – Total Number of Observations*	*X – The Value in the data distribution* *$\bar{x}$ – The Sample Mean* *n - Total Number of Observations*

We will accept both answers for full credit in this case.

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.

Solution:

Listed below is a reference solution which chooses common answers to this problem. However, note that since the question is pretty open ended there could be many possible correct answers and when grading we will have to evaluate the validity of the chosen classification method and the justification/explanation.

a). Naive Bayes models, built on the principles of feature independence and Bayesian probability, excel at accommodating incremental data updates without requiring full retraining. These models can efficiently adjust class probabilities and feature distributions in response to fresh data without the need to revisit the entire dataset. Similarly, Bayesian classification, another approach supporting incremental data, leverages Bayes' theorem to estimate hypothesis probabilities and updates prior beliefs as new instances arrive, making it a valuable tool for evolving data scenarios.

b). Rule-based Classification and Decision Trees are renowned classification methods celebrated for their transparency and interpretability. Both approaches present clear and intuitive rules or tree structures, allowing users to easily grasp and articulate classification decisions. This simplicity and transparency are particularly valuable in fields such as law, medicine, and regulations, where the comprehensibility of decision-making processes is paramount. Decision trees, in particular, offer a flowchart-like structure with internal nodes representing attribute tests and branches indicating outcomes, making them an essential tool in domains that prioritize transparent and easily understandable models.

c). Deep Neural Networks (DNNs) are complex classification models that rely on extensive training data and exhibit opaque decision-making. Their intricate, multi-layered structures and large parameter spaces make it difficult to understand the influence of individual parameters on classification outcomes. Furthermore, DNNs demand significant computational resources for training and fine-tuning. These networks lack interpretability due to their black-box nature and complex decision processes.

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.

Solution:

Let us first calculate the Euclidean distance from one point to every other initial centroids, then we can find closest centroid to each points.

Euclidean distance = sqrt((x1-x2)^2 + (y1-y2)^2)

Distances for p_1(1,2):
Centroid A = 0
Centroid B = sqrt((1-3)^2 + (2-3)^2) = sqrt(5)
Centroid C = sqrt((1-5)^2 + (2-5)^2) = sqrt(20)
Closest centroid is A

Distances for p_2(2,1):
Centroid A = sqrt((2-1)^2 + (1-2)^2) = sqrt(4) = 2
Centroid B = sqrt((3-3)^2 + (2-3)^2) = sqrt(1) =1
Centroid C = sqrt((3-5)^2 + (2-5)^2) = sqrt(13)
Closest centroid is A

Distances for p_3(3,2):
Centroid A = sqrt((3-1)^2 + (2-2)^2) = sqrt(2)
Centroid B = sqrt((3-3)^2 + (2-3)^2) = sqrt(1)
Centroid C = sqrt((3-5)^2 + (2-5)^2) = sqrt(13)
Closest centroid is B

Distances for p_4(3,3):
Centroid A = sqrt((3-1)^2 + (3-2)^2) = sqrt(5)
Centroid B = 0

Centroid C = sqrt((3-5)^2 + (3-5)^2) = sqrt(8)
Closest centroid is B

Distances for p_5(4,1):
Centroid A = sqrt((4-1)^2 + (1-2)^2) = sqrt(10)
Centroid B = sqrt((4-3)^2 + (1-3)^2) = sqrt(5)
Centroid C = sqrt((4-5)^2 + (1-5)^2) = sqrt(17)
Closest centroid is B

Distances for p_6(5,2):
Centroid A = sqrt((5-1)^2 + (2-2)^2) = sqrt(16)
Centroid B = sqrt((5-3)^2 + (2-3)^2) = sqrt(5)
Centroid C = sqrt((5-5)^2 + (2-5)^2) = sqrt(9)
Closest centroid is B

Distances for p_7(5,5):
Centroid A = sqrt((5-1)^2 + (2-2)^2) = sqrt(16)
Centroid B = sqrt((5-3)^2 + (2-3)^2) = sqrt(5)
Centroid C = sqrt((5-5)^2 + (2-5)^2) = sqrt(9)
Closest centroid is C

Distances for p_8(6,4):
Centroid A = sqrt((6-1)^2 + (4-2)^2) = sqrt(29)
Centroid B = sqrt((6-3)^2 + (4-3)^2) = sqrt(10)
Centroid C = sqrt((6-5)^2 + (4-5)^2) = sqrt(2)
Closest centroid is C

So our initial Clusters are:
Cluster A = **{p_1, p_2}**
Cluster B = **{p_3, p_4, p_5, p_6}**
Cluster C = **{p_7, p_8}**

We can now calculate the new centroids, which is the average of all the points in the cluster:

New centroids:
Centroid A = (1+2/2, 2+1/2) = **(1.5, 1.5)**
Centroid B = (3+3+4+5)/4, (2+3+1+2)/4 = **(3.75, 2)**
Centroid C = (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?

Solution:

This is a very open ended question, and each individual response will have to be evaluated, so no solution is provided.