COMP7990 Quiz

Seat No **LT: WLB103** **Seat:**

Name:	**SAMPLE ANSWER**	**Student ID:**	

- There are 6 questions. **The full mark on the quiz is 100**.
- Write your answer directly in the designated space provided.
- This is a closed-book quiz. You are allowed to use a calculator to attempt the quiz.
- Quiz time: 60 minutes.

Question 1: Data Normalization and Binning

1. Explain why data normalization is necessary in machine learning. (4 marks)

   Solution:

   Data normalization ensures that all features contribute equally to distance-based models, like KNN or K-means. It prevents variables with larger ranges from dominating distance calculations.

2. Given the dataset: [50, 200, 300, 100, 500], apply Z-score (sample mean and sample standard deviation) and min-max ([0, 1]) normalization for the third number 300. Show your steps and calculations. (6 marks)

   Solution:

   1. Z-score Normalization for 300

Step 1: Calculate the mean ($\bar{X}$):

$$\bar{X} = \frac{50 + 200 + 300 + 100 + 500}{5} = \frac{1150}{5} = 230$$

Step 2: Calculate the sample variance (s^2):

Sample variance formula:

$$s^2 = \frac{\sum(X_i - \bar{X})^2}{n - 1}$$

$$s^2 = \frac{(50 - 230)^2 + (200 - 230)^2 + (300 - 230)^2 + (100 - 230)^2 + (500 - 230)^2}{5 - 1}$$

$$s^2 = \frac{(-180)^2 + (-30)^2 + (70)^2 + (-130)^2 + (270)^2}{4}$$

$$s^2 = \frac{32400 + 900 + 4900 + 16900 + 72900}{4} = \frac{128000}{4} = 32000$$

Step 3: Calculate the sample standard deviation (s):

$$s = \sqrt{32000} \approx 178.89$$

Step 4: Calculate the Z-score for 300:

$$Z = \frac{300 - 230}{178.89} = \frac{70}{178.89} \approx 0.39$$

2. Min-Max Normalization for 300

The Min-Max normalization formula is:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \times (new\_max - new\_min) + new\_min$$

Assume we want to normalize the values between 0 and 1:

- $X = 300$
- $X_{min} = 50$
- $X_{max} = 500$
- $new\_min = 0, new\_max = 1$

Step 1: Apply the min-max formula:

$$X' = \frac{300 - 50}{500 - 50} \times (1 - 0) + 0$$

$$X' = \frac{250}{450} \times 1 = 0.5556$$

3. Perform equal-width binning with 3 bins on the following dataset: [8, 12, 20, 5, 15, 25, 30, 10]. Smooth the data using bin means. (5 marks)

Solution:

Step 1: Sort the dataset

Sort the data in ascending order:

$[5, 8, 10, 12, 15, 20, 25, 30]$

Step 2: Determine the bin width

To perform equal-width binning, the bin width is calculated as:

$$\text{Bin width} = \frac{\text{Max value} - \text{Min value}}{\text{Number of bins}} = \frac{30 - 5}{3} = \frac{25}{3} \approx 8.33$$

So, each bin will cover a range of approximately 8.33 units.

Step 3: Divide the data into 3 equal-width bins

- **Bin 1:** Range = [5, 5 + 8.33) = [5, 13.33) → Values: [5, 8, 10, 12]
- **Bin 2:** Range = [13.33, 13.33 + 8.33) = [13.33, 21.66) → Values: [15, 20]
- **Bin 3:** Range = [21.66, 30] → Values: [25, 30]

Step 4: Calculate the mean for each bin

- **Bin 1 mean:** $(5 + 8 + 10 + 12)/4 = 8.75$
- **Bin 2 mean:** $(15 + 20)/2 = 17.5$
- **Bin 3 mean:** $(25 + 30)/2 = 27.5$

Step 5: Smooth the data using bin means

- For **Bin 1**, replace all values with the mean 8.75: $[8.75, 8.75, 8.75, 8.75]$
- For **Bin 2**, replace all values with the mean 17.5: $[17.5, 17.5]$
- For **Bin 3**, replace all values with the mean 27.5: $[27.5, 27.5]$

Final Smoothed Dataset:

$[8.75, 8.75, 8.75, 8.75, 17.5, 17.5, 27.5, 27.5]$

Question 2: Linear Regression and Correlation

You are given the following data about house prices and the size of houses:

Size (sq ft)	Price (USD 100s)
800	300
900	330
1000	350
1200	400
1500	480

1. Calculate the Pearson correlation coefficient between the house size and price. (12 marks)

Solution:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Step 1: Calculate the means

- Mean of house sizes ($\bar{X}$):

$$\bar{X} = \frac{800 + 900 + 1000 + 1200 + 1500}{5} = 1080$$

- Mean of house prices ($\bar{Y}$):

$$\bar{Y} = \frac{300 + 330 + 350 + 400 + 480}{5} = 372$$

Step 2: Calculate the terms $(X_i - \bar{X})(Y_i - \bar{Y})$

For each pair of data points, we calculate the product of the differences from the mean:

- $(800 - 1080)(300 - 372) = (-280)(-72) = 20160$
- $(900 - 1080)(330 - 372) = (-180)(-42) = 7560$
- $(1000 - 1080)(350 - 372) = (-80)(-22) = 1760$
- $(1200 - 1080)(400 - 372) = (120)(28) = 3360$
- $(1500 - 1080)(480 - 372) = (420)(108) = 45360$

Sum of products:

$$20160 + 7560 + 1760 + 3360 + 45360 = 78200$$

Step 3: Calculate the sum of squares for each dataset

For house sizes $(X_i - \bar{X})^2$:

- $(800 - 1080)^2 = (-280)^2 = 78400$
- $(900 - 1080)^2 = (-180)^2 = 32400$
- $(1000 - 1080)^2 = (-80)^2 = 6400$
- $(1200 - 1080)^2 = (120)^2 = 14400$
- $(1500 - 1080)^2 = (420)^2 = 176400$

Sum of squared differences for house sizes:

$$78400 + 32400 + 6400 + 14400 + 176400 = 308000$$

For house prices $(Y_i - \bar{Y})^2$:

- $(300 - 372)^2 = (-72)^2 = 5184$
- $(330 - 372)^2 = (-42)^2 = 1764$
- $(350 - 372)^2 = (-22)^2 = 484$
- $(400 - 372)^2 = (28)^2 = 784$
- $(480 - 372)^2 = (108)^2 = 11664$

Sum of squared differences for house prices:

$$5184 + 1764 + 484 + 784 + 11664 = 19880$$

Step 4: Apply the Pearson correlation formula

Now we can substitute the calculated values into the formula:

$$r_{XY} = \frac{78200}{\sqrt{308000} \cdot \sqrt{19880}}$$

Calculate the square roots:

$$\sqrt{308000} \approx 554.98, \quad \sqrt{19880} \approx 140.99$$

Now substitute:

$$r_{XY} = \frac{78200}{554.98 \times 140.99} \approx \frac{78200}{78250.212} \approx 0.99936$$

2. What are the differences between Pearson Correlation and Linear Regression. (5 marks)

   Solution: Pearson correlation and linear regression are both methods for analyzing relationships between two variables, but they have distinct purposes and interpretations.

   Pearson Correlation: It measures the strength and direction of a linear relationship between two variables. It describes only the association (correlation) between two variables. It doesn't imply causation and doesn't provide a predictive model.

   Linear Regression: Models the relationship between a dependent (response) variable and an independent (predictor) variable by fitting a line to the data. The goal is to predict or explain the dependent variable using the independent variable(s). Linear regression provides a fitted line with a slope and intercept that best represent the data.

   In Summary

   Pearson Correlation quantifies the strength of association between two variables, while linear regression provides a predictive relationship, modeling how one variable changes with respect to another.

3. Please specify the pros and cons of gradient descent and normal equation to estimate parameters of linear regression. (6 marks)

   Solution:

   Gradient Descent: 1. Need to choose a learning rate. 2. Needs many iterations. 3. Works well even if the number of dimensions is large.

   Normal Equation: 1. No need to choose a learning rate. 2. Don't need to iterate. 3. Need to compute $(X^T X)^{-1}$ 4. Slow if the number of dimensions is very large.

Question 3: Classification Algorithms

1. Explain the differences between (1) KNN and K-means; (2) Linear regression and SVM. (6 marks)

   Solution:

   KNN and K-means:

- **K-Means** is used for **clustering** (unsupervised learning) and involves finding k clusters by iteratively updating cluster centroids.
- **K-NN** is used for **classification** (supervised learning), and it makes predictions by looking at the k-nearest neighbors of a test point in the dataset.

Linear regression and SVM:

- **Linear Regression** is best suited for **regression** tasks with a focus on minimizing prediction error and is straightforward but sensitive to outliers.
- **SVM** is designed primarily for **classification**, focusing on maximizing the margin between classes and is highly versatile, particularly with kernel methods for non-linear separations.

Consider the following dataset for a K-NN classification problem:

Name	Feature 1	Feature 2	Label
Alice	1	2	Yes
Bob	2	3	No
Carol	3	1	Yes
Dave	3	4	No

Using **K=3**, classify the following new data point: (Feature 1 = 2, Feature 2 = 2). Use Euclidean distance and show your steps. (8 marks)

Solution:

K-NN with K=3:
New point: (2, 2)

- Distances:
  - Alice: $\sqrt{(2-1)^2 + (2-2)^2} = 1$
  - Bob: $\sqrt{(2-2)^2 + (2-3)^2} = 1$
  - Carol: $\sqrt{(2-3)^2 + (2-1)^2} = 1.41$
  - Dave: $\sqrt{(2-3)^2 + (2-4)^2} = 2.24$
- Nearest neighbors: Alice, Bob, Carol → 2 Yes, 1 No → Predicted class: Yes

Question 4: Hypothesis Testing (16 marks)

Given the following two datasets representing scores of students from two different classes in a test,

Class A scores: [80, 85, 78, 92, 88, 75, 85]
Class B scores: [75, 80, 72, 89, 85, 70, 78]

1. If these two datasets do not follow a normal distribution, would you prefer a parametric or non-parametric test? Specify the exact test you would choose. (6 marks)
2. If these two datasets follow a normal distribution, perform the most approporiate test we learned to check if there is a significant difference in their performance. Use a significance level of 0.05. Note: you do not need to perform the actual calculations; just write down the essential steps to solve this problem. The steps should be detailed, not just general words. (10 marks)

Solution:

1. Non-parametric test, Mann-Whitney U test
2.

- **State the Hypotheses**:

Null Hypothesis (H_0): There is no difference between the means of the two groups.

Alternative Hypothesis (H_1): There is a significant difference between the means of the two groups.

- **Choose the Significance Level**:

Select the critical value (t=2.179 at α=0.05) to determine the threshold for rejecting the null hypothesis.

- **Calculate the Test Statistic**:

For comparing two means, the test statistic is typically the **t-statistic** for a independent **t-test**:

$$t = \frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{s_{Difference}} = \frac{\bar{x} - \bar{y}}{\sqrt{s^2_{Pooled}(\frac{1}{m} + \frac{1}{n})}}$$

$$s^2_{Pooled} = \frac{df_x s_x^2 + df_y s_y^2}{df_{Total}} = (\frac{df_x}{df_{Total}})s_x^2 + (\frac{df_y}{df_{Total}})s_y^2$$

Where $\bar{x}$ (S_x) is the mean (sample standard deviation) of Class A scores and $\bar{y}$ (S_y) is the mean (sample standard deviation) of Class B scores.

$df_x = df_y = 6, df_{Total} = 12$

- **Determine the significance**:

Compare the calculated t-statistic to the critical value from the **t-distribution.**

- **Make a Decision**:

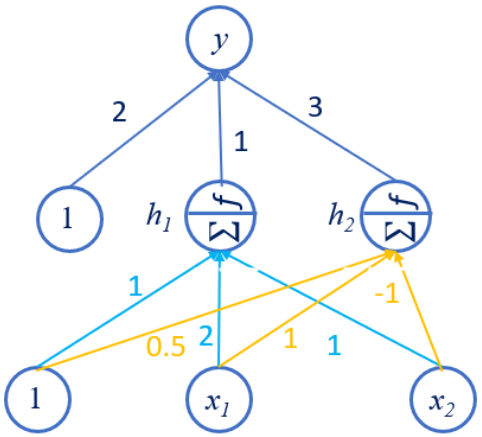
  - ✓ If the t-statistic < -2.179 or t-statistic >2.179 , reject the null hypothesis. This suggests there is a statistically significant difference between the means.
  - ✓ If the -2.179≤ t-statistic≤2.179, fail to reject the null hypothesis, meaning no significant difference between the groups.

Question 5: Neural Networks

1. **Explain the concept of a perceptron and its limitations.** (6 marks)

   Solution: A perceptron is a linear classifier that adjusts weights based on misclassified points. Its limitation is the inability to 1. solve non-linearly separable problems; 2. deal with misclassified points.

2. Consider an artificial neural network with two input nodes, one hidden layer (with two nodes), and one output node. The activation function is ReLU ($f(x) = \max(0, x)$). Given the following input values (x_1=1, x_2=2), calculate the output value (y) of the network. Use the following weights (10 marks):



Solution:

- Hidden Layer Node 1: max(0,1*1+1*2+2*1)=5
- Hidden Layer Node 2: max(0,1*0.5+1*1+2*-1) = 0
- Output Node: 1*2+5*1+3*0=7

Question 6: SVM and Decision Boundary

1. **Describe how a SVM selects a decision boundary and why maximizing the margin is important.** (6 marks)

   Solution:

   A SVM selects the decision boundary that maximizes the margin between classes. Maximizing the margin helps improve generalization and reduces overfitting.

2. **Consider the following points:**

Point	X1	X2	Class
P1	2	3	+1
P2	3	4	+1
P3	1	1	-1
P4	2	1	-1

Draw the decision boundary that would be selected by an SVM, and explain your reasoning. (9 marks)

Solution:

- Points P1 and P2 are positive class, P3 and P4 are negative.
- The SVM will maximize the margin by selecting a boundary equidistant from the closest positive and negative points (support vectors).
- In this case, P1 and P4 are likely to be the support vectors, as they are the closest points from opposite classes.
- The boundary will be located **midway between the support vectors** P1(2,3) and P4(2,1), which means the decision boundary will be at X1=2.

Appendix

t-Table (for t-tests)

one-tail	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	0.2	0.1	0.05	0.02	0.01	0.002	0.001
DF							
1	3.078	6.314	12.706	31.821	63.656	318.289	636.578
2	1.886	2.92	4.303	6.965	9.925	22.328	31.6
3	1.638	2.353	3.182	4.541	5.841	10.214	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.61
5	1.476	2.015	2.571	3.365	4.032	5.894	6.869
6	1.44	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.86	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.25	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.93	4.318
13	1.35	1.771	2.16	2.65	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.14
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073

F-Table (for ANOVA)

Table F The *F* Distribution

df_D \ df_N	1	2	3	4	5
1	161.4	199.5	215.7	224.6	230.2
2	18.51	19.00	19.16	19.25	19.30
3	10.13	9.55	9.28	9.12	9.01
4	7.71	6.94	6.59	6.39	6.26
5	6.61	5.79	5.41	5.19	5.05
6	5.99	5.14	4.76	4.53	4.39
7	5.59	4.74	4.35	4.12	3.97
8	5.32	4.46	4.07	3.84	3.69
9	5.12	4.26	3.86	3.63	3.48
10	4.96	4.10	3.71	3.48	3.33
11	4.84	3.98	3.59	3.36	3.20
12	4.75	3.89	3.49	3.26	3.11
13	4.67	3.81	3.41	3.18	3.03
14	4.60	3.74	3.34	3.11	2.96
15	4.54	3.68	3.29	3.06	2.90
16	4.49	3.63	3.24	3.01	2.85
17	4.45	3.59	3.20	2.96	2.81
18	4.41	3.55	3.16	2.93	2.77
19	4.38	3.52	3.13	2.90	2.74
20	4.35	3.49	3.10	2.87	2.71
21	4.32	3.47	3.07	2.84	2.68
22	4.30	3.44	3.05	2.82	2.66
23	4.28	3.42	3.03	2.80	2.64
24	4.26	3.40	3.01	2.78	2.62

Table for U-test

Table 3 Critical values of *U* (5% significance).

n_1 \ n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2								0	0	0	0	1	1	1	1	1	2	2	2	2
3					0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4				0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5			0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6			1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7			1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8		0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9		0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10		0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11		0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12		1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13		1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14		1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15		1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16		1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17		2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18		2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19		2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20		2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

Table for W-test

Critical Values of the Wilcoxon Signed Ranks Test

n	Two-Tailed Test		One-Tailed Test	
	α = .05	α = .01	α = .05	α = .01
5	--	--	0	--
6	0	--	2	--
7	2	--	3	0
8	3	0	5	1
9	5	1	8	3
10	8	3	10	5
11	10	5	13	7
12	13	7	17	9
13	17	9	21	12
14	21	12	25	15
15	25	15	30	19
16	29	19	35	23

Rough paper