COMP7990 Quiz

Seat No
LT:
Seat:

Name:	SAMPLE ANSWER	Student ID:	

- There are 9 questions and 13 pages. **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: 90 minutes.

Question 1. You are given the following data about the electricity bills of an apartment. **(18 marks)**

Month	Average Monthly Temperature (in ºC)	Electricity Bill (in $; assume the bill date has been adjusted to that month already)
Jan	15	500
Feb	22	400
Mar	36	720
Apr	27	560
May	33	700
Jun	35	900

a) Can you suggest a relationship between the average monthly temperature and electricity bill? Please quantify such a relationship. (Precision of your numbers should be in 2 decimal places). **(5 marks)**

b) Please create a linear regression using the average monthly temperature to predict the electricity bill without normalizing the data. **(7 marks)**

c) If the above is normalized by decimal scaling, do you think the answer obtained in part a) and part b) will be different? Please give reasons to support your argument. No calculation is required. **(6 marks)**

a) They are (positively) correlated / cause-and-effect. Correlation = **0.82**1

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

$$\overline{X} = 28, \overline{Y} = 630$$

r_{xy} =(-13 * -130 + -6*-230 + 8*90+-1*-70+5*70 + 7*270) /
sqrt (169+36+64+1+25+49)sqrt(16900+52900+8100+4900+4900+72900)
= **0.82**1

b)

θ1 =(-13 * -130 + -6*-230 + 8*90+-1*-70+5*70 + 7*270) / (169+36+64+1+25+49) = **17.73**256
θ2=630-17.73256*28 = 133.48832 = **133.49**

c)

Will the answer in part a) be different?

Reason:

No change. Because the numerator and denominator of correlation will be multiplied by the same decimal scale factor and be canceled out. (Explanation derived from the formula will also be accepted. From lecture notes we say correlation is independent of units. If they quote that statement and somehow be able to relate from "unit" to "scaling", it is also considered as a correct answer).

Will the answer in part b) be different?

Reason:

Yes, it will be different. The regression line represent the linear relationship of the points on the graph. The normalized graph will be plotted in a different scale so as the regression line.

Question 2. You are given the following data about the interview score sheet of some job applicants. **(10 marks)**

Name	Academic Qualification (0-5, 5 is best)	Entry Test (0-5, 5 is the best)	Result
Alex	5	5	Hire
Bob	2	4	Not Hire
Cathy	4	3	Hire
Dave	5	1	Not Hire
Eva	4	4	?

a) To predict the interview result for Eva, is it better to apply K-NN algorithm or K-mean algorithm? Explain your reason. **(4 marks)**

b) With the answer you have provided in part a), run that algorithm with K = 3. Show your steps **(6 marks)**

a)

This is a classification problem. K-nn is a classfication algoirthm while K-mean is a clustering algorithm. Therefore, K-nn is suitable

b)

a) Using Euclidean distatnce Distance from Eva to each person is
Alex: sqrt(1+1) = sqrt (2)
Bob: sqrt(4 + 0) = sqrt (4)
Cathy: sqrt (0 + 1) = sqrt (1)
Dave: sqrt(1+9) = sqr (10)
2 Hire, 1 not hire. Decision Hire.

Question 3. Answer the following questions. **(20 marks)**

a) Describe the motivation of data normalization. (**4 marks**)
b) Suppose that you have the following data: [100, 200, 300, 400], normalize the data by min-max normalization by setting new min to 5 and new max to 8. (**8 marks**)
c) Suppose we have the following values: [4, 34, 31, 30, 21, 31, 21, 25, 22, 28, 10, 15]. Use smoothing by bin means to smooth the above data using **equal-depth** binning with 3 bins. Show your steps. (**8 marks**)

a) *Sample solution: Because the features have different scales and normalization could make them comparable. Especially, the features with wider scale may dominate the Euclidean distance.*

b)

Sample solution:

$$x_i' = \frac{(x_i - min)}{max - min}(max_{new} - min_{new}) + min_{new}$$

min:100, max:400, max-min=300
max(new)=8; min(new)=5
x_1=100, $x_1' = \frac{100-100}{300} * (8-5) + 5 = 5$
x_2=200, $x_2' = \frac{200-100}{300} * (8-5) + 5 = 6$
x_3=300, $x_3' = \frac{300-100}{300} * (8-5) + 5 = 7$
x_4=400, $x_4' = \frac{400-100}{300} * (8-5) + 5 = 8$

c)

Step 1: Sort the list as [4, 10, 15, 21, 21, 22, 25, 28, 30, 31, 31, 34]
Step 2: Compute the number of elements in each bin: 12/3=4
Step 3: Partition data into bins
Bin 1: [4, 10, 15, 21]. The mean is 12.5
Bin 2: [21, 22, 25, 28]. The mean is 24
Bin 3: [30, 31, 31, 34]. The mean is 31.5
Step 4: Smooth each bin by mean
[12.5, 12.5, 12.5, 12.5, 24, 24, 24, 24, 31.5, 31.5, 31.5, 31.5]

Question 4. Answer the following questions. **(15 marks)**

a) Consider the following dataset (**Figure 1**), can perceptron separate the points of y=1 and y=-1? Justify your answer. (**5 marks**)

b) Consider the feed-forward neural network (**Figure 2**). The model weights are shown on the edges. Suppose we use the sigmoid function (i.e., $f(x) = \frac{2}{1+\exp(-2x)}$) as the activation function in this network. Given a sample with $x_1 = 0$ and $x_2 = 0$, what are the output values of each node. (**10 marks**)

x_1	x_2	y
1	-1	1
1	1	-1
-1	1	1
-1	-1	-1

Figure 1. Dataset for part a)

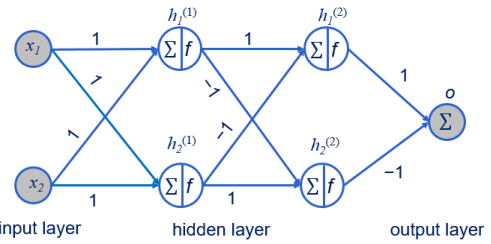

Figure 2. Feed-forward NN for part b)

a)

Sample solution: No. Perceptron is a linear classifier and it can not solve non-linear problem.

b)

Node	Output Value
$h_1^{(1)}$	2/(1+exp(-2*(0*1+0*1))) =1
$h_2^{(1)}$	2/(1+exp(-2*(0*1+0*1))) =1
$h_1^{(2)}$	2/(1+exp(-2*(1*1+1*-1))) =1
$h_2^{(2)}$	2/(1+exp(-2*(1*1+1*-1))) =1
o	1*1 + 1*(-1) = 0

Question 5. Categorize the following datasets into qualitative data, quantitative data (discrete), and quantitative data (continuous). Circle the correct answer. **(5 marks)**

Datasets	**Answer**
Brand of car	Qualitative / Quantitative Discrete / Quantitative Continuous
Floor level	Qualitative / Quantitative Discrete / Quantitative Continuous
Price	Qualitative / Quantitative Discrete / Quantitative Continuous
Race	Qualitative / Quantitative Discrete / Quantitative Continuous
Body weight	Qualitative / Quantitative Discrete / Quantitative Continuous
Number of students	Qualitative / Quantitative Discrete / Quantitative Continuous
Car speed	Qualitative / Quantitative Discrete / Quantitative Continuous
Religious	Qualitative / Quantitative Discrete / Quantitative Continuous
Room number	Qualitative / Quantitative Discrete / Quantitative Continuous

Question 6. Interprete the following terms. **(4 marks)**

	Terms	Interpretation
a)	Statistical Inference	Statistical inference is the process of using data from a sample to gain information about the population.
b)	Bias Sampling	Sampling bias occurs when the method of selecting a sample causes the sample to differ from the population in some relevant way.
c)	Outlier	a part of sample data that significantly deviates from the majority of the sample.
d)	Kernel	to transform data points to another dimension so that non-linear separable data become linearly separable.

Question 7. Given a set of sample data, calculate the following statistics. Write your answer up to 2 decimal places. **(4 marks)**

5, 10, 28, 29, 37, 39, 41, 44, 47, 56, 61, 99

Median	40
Range	94
IQR	23
Outlier(s)	99

Question 8. The table below shows the scores of master students in the mid-exam and final exams. We would like to know **if students' final-exam scores are significantly better than their mid-exam scores**. Assume Mid-exam and Final-exam scores follow **normal distributions**. Choose a proper hypothesis test to answer this question. **(12 marks)**

Student ID	**Mid-exam**	**Final-exam**
1	92	86
2	72	77
3	74	76
4	73	85
5	52	83
6	60	71
7	57	72
8	85	78
9	77	72
10	64	73
11	80	74
12	53	79

Show your calculation with 3 decimal places. You are advised to work on a rough paper and tidy your answer below.

Null hypothesis: There is no significant improvement/difference between the two exams' scores

Test method: Paired/Dependent Samples t

Calculation process:

$$t = \frac{\bar{x}_{differences} - \mu_o}{S_{differences} / \sqrt{2}}$$

$$S.E = Sdifferences / \sqrt{n} = 12.6212/\sqrt{12} = 3.6434$$

$$t = \frac{7.25 - 0}{3.6434} = 1.9899$$

Look up the T table: critical value is 2.201

t=1.9899 < 2.201 (critical value)

Conclusion: it cannot reject the null hypothesis at α= .05, so there is no significant improvement.

Question 9. The table below shows the accuracy of three recommendation algorithms on 8 tasks. We would like to know **if these three algorithms yield significantly different accuracy**. Assume the scores of Algorithm 1, Algorithm 2 and Algorithm 3 follow normal distributions. Choose a proper hypothesis test to answer this question. **(12 marks)**

Task	Algorithm1	Algorithm2	Algorithm3
Task1	0.51	0.69	0.85
Task2	0.46	0.92	0.82
Task3	0.34	0.61	0.74
Task4	0.58	0.63	0.75
Task5	0.66	0.89	0.81
Task6	0.45	0.89	0.84
Task7	0.44	0.84	0.71
Task8	0.58	0.79	0.82

Show your calculation with 3 decimal places. You are advised to work on a rough paper and tidy your answer below.

Null hypothesis: There is no significant difference among the three algorithms regarding the accuracy

Test method: ANOVA

Calculation process:

Step-1 : Between group variation

$$SST = n_1(\bar{Y}_1 - \bar{Y})^2 + n_2(\bar{Y}_2 - \bar{Y})^2 + \ldots + n_p(\bar{Y}_p - \bar{Y})^2$$

$=8\times(0.5025-0.6925)^2+8\times(0.7825-0.6925)^2+8\times(0.7925-0.6925)^2$

$=0.4336$

Step-2 : In group variation

$$SSE = (Y_{11} - \bar{Y}_1)^2 + (Y_{21} - \bar{Y}_1)^2 + \ldots + (Y_{51} - \bar{Y}_1)^2 + (Y_{12} - \bar{Y}_2)^2 + (Y_{22} - \bar{Y}_2)^2 + \ldots + (Y_{52} - \bar{Y}_2)^2 + (Y_{13} - \bar{Y}_3)^2 + (Y_{23} - \bar{Y}_3)^2 + \ldots + (Y_{53} - \bar{Y}_3)^2$$

$= (0.51-0.5025)^2 + (0.46-0.5025)^2 + (0.34-0.5025)^2 + (0.58-0.5025)^2 + (0.66-0.5025)^2 + (0.45-0.5025)^2 + (0.44-0.5025)^2 + (0.58-0.5025)^2$
$+ (0.69-0.7825)^2 + (0.92-0.7825)^2 + (0.61-0.7825)^2 + (0.63-0.7825)^2 + (0.89-0.7825)^2 + (0.89-0.7825)^2 + (0.84-0.7825)^2 + (0.79-0.7825)^2$
$+ (0.85-0.7925)^2 + (0.82-0.7925)^2 + (0.74-0.7925)^2 + (0.75-0.7925)^2 + (0.81-0.7925)^2 + (0.84-0.7925)^2 + (0.71-0.7925)^2 + (0.82-0.7925)^2$

$=0.1974$

Step-3 : variance between samples

MST=SST/p-1

=0.4336/2

=0.2168

Step-4 : variance within samples

MSE=SSE/(*n*-p*)*

=0.1974/(24-3)
=0.0094

Step-5 : test statistic F for one way ANOVA test

F=MST/MSE

=0.2168/0.0094
=23.064

Look up the F table: critical value is 5.79

F=23.064 > 5.79 (critical value)

Conclusion: it can reject the null hypothesis at α= .05, so there is a significant difference among the three algorithms regarding the accuracy.

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

Rough paper

Rough paper