UNIVERSITY OF CANTERBURY

PRACTICE EXAM

Prescription Number(s):	**STAT462-24S1 (C)**
Paper Title:	**Data Mining**

Time allowed:	2 HOURS
Number of Pages:	7

Read these instructions carefully.

- Answer all FIVE questions
- All questions carry equal marks.
- Calculators are permitted.
- Use black or blue ink only.
- Show all working.
- Write your answers in a format that you can sort into one document to hand in.

Questions Start on Page 3

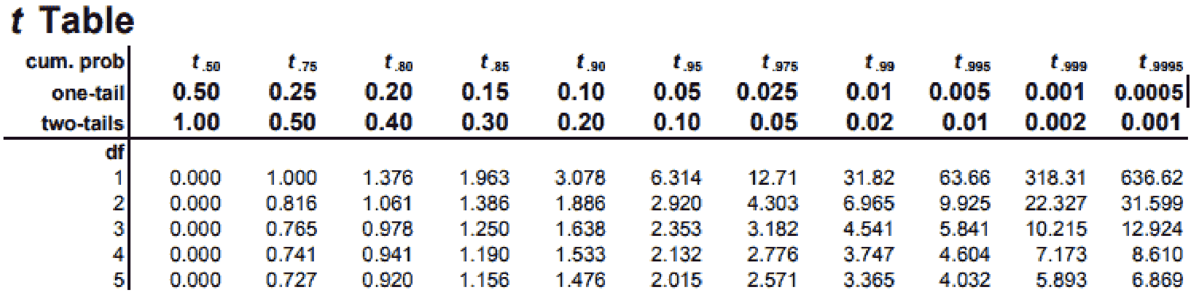
1. **Simple Linear Regression**

We consider the following dataset: $\underline{x} = (0, 1, 3, 4)$ and $\underline{y} = (1, -2, 2, -1)$.

The following table might be useful to keep track of various computations.

n	mean($\underline{\mathrm{x}}$)	mean($\underline{\mathrm{y}}$)	var ($\underline{\mathrm{x}}$)	var ($\underline{\mathrm{y}}$)	cov ($\underline{\mathrm{x}}, \underline{\mathrm{y}}$)	$\widehat{b}_0$	$\widehat{b}_1$	TSS	ESS	RSS	R^2

x_i	0	1	3	4
y_i	1	-2	2	1
$x_i - \text{mean}(\underline{\mathrm{x}})$				
$y_i - \text{mean}(\underline{\mathrm{y}})$				
$\widehat{y}_i$				
$\widehat{y}_i - \text{mean}(\underline{\mathrm{y}})$				



***t* Table**

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869

(a) Compute the coefficients $\hat{b}_0, \hat{b}_1$ which minimise the least square error of the corresponding linear regression model $x \mapsto b_0 + b_1 x$.

(b) Which percentage of variation in the data can be explained by your regression model from (a), as measured by the R-squared?

(c) What is the predicted response for a new data point $x^\star = 5$?

(d) Compute the 60% confidence interval for b_1. Is b_1 significant at this confidence level? *Note: t-distribution quantiles can be found in the table provided.*

2. **Classification**

In this scenario, New Zealand is troubled by a recent outbreak of the invasive Kangaroo pox affecting livestock. The government is mandating all livestock to be tested in order to minimise the risk of further spread of the disease.

The Ministry of Health is looking at ordering large quantities of tests from one of two competing health care technology companies, *Quack Enterprises* and *ProfitFirst Medical.*

The following tables describe sensitivity and specificity of the two competing test variants.

A *positive test result* means that the test judges the tested subject to be infected, and a *negative test result* means that the test judges the tested subject to be healthy.

Quack Enterprises	test result: positive	test result: negative
actually infected	2	1
actually healthy	7	90

ProfitFirst Medical	test result: positive	test result: negative
actually infected	1	2
actually healthy	2	95

(a) Assuming that the used sample is representative of the population, what percentage of livestock carry the disease?

(b) What is the specificity and sensitivity of each test?

(c) For each of these tests, what is the probability that a given positively tested subject is actually inflicted with the disease?

(d) The goal of the Ministry of Health is to find a test that is very sensitive to the disease. Which test is more suitable for this purpose?

3. **Linear Regression and k-NN Regression**. We consider a training set $\mathrm{Tr} = \{(x_i, y_i)_i\}$ with $(x_1, y_1) = (0, 0)$, $(x_2, y_2) = (1, 1)$ and $(x_3, y_3) = (3, 9)$.

(a) Write down explicitly the 2-Nearest-Neighbor regression function $f_{2-NN}(x)$ for this dataset. Your definition needs to specify an return value for every possible input x.

(b) Compute the training-MSE-optimal coefficients $\beta_0, \beta_1, \beta_2$ for a quadratic regression model $f_2(x) = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2$. Hint: If you look closely at the numbers you should be able to guess the coefficients directly.

(c) Now we introduce a test set consisting of just one data point $\mathrm{Te} = \{(-1, 1)\}$. Compute the MSE for both the 2-NN and the quadratic model, for both the training set and the test set. This should give you four numbers.

Decide on the basis of these numbers which model you would pick and give a short and reasonable argument for your decision.

4. **Association Analysis**

Suppose that we have the following 100 market basket transactions.

	Transaction	Frequency
1	{apple}	9
2	{apple, carrot}	10
3	{apple, banana, carrot}	27
4	{apple, banana, carrot, orange}	21
5	{banana, orange}	3
6	{apple, banana, orange}	11
7	{carrot, orange}	5
8	{banana, carrot, orange}	14
		100

For example, there are 10 transactions of the form {apple, carrot}

(a) Compute the support of {orange}, {banana, carrot}, and {banana, carrot, orange}.

0.54 0.62 0.35

(b) Compute the confidence of the association rules:

$$\{\text{banana, carrot}\} \rightarrow \{\text{orange}\}; \text{ and } \quad C_1 = \frac{21+14}{27+21+14} = 0.56$$

$$\{\text{orange}\} \rightarrow \{\text{banana, carrot}\}. \quad C_2 = \frac{21+14}{54} = 0.65$$

Is confidence a symmetric measure? **Justify your answer.**

(c) Find the 3-itemset(s) with the largest support. {apple, banana, carrot}

(d) If $minsup = 0.2$, is {banana, carrot, orange} a maximal frequent itemset? **Justify your answer.**

(e) Lift is defined as e: X and Y are negatively correlated.

$$\text{Lift}(X \rightarrow Y) = \frac{s(X \cup Y)}{s(X)s(Y)}, = \frac{P(X \cap Y)}{P(X)P(Y)}$$

X decreases the chance of seeing Y and vice versa.

where $s(\ )$ denotes support. What does it mean if $\text{Lift}(X \rightarrow Y) < 1$.

(c) $F_{k-1} \times F_{k-1}$ method:

k=2 {apple, carrot} {banana, orange} {carrot, orange} => k=3 $s(\{a, c, o\}) = 21/100$

$s(\{a, b, c\}) = \frac{27+21+11}{100} = \frac{59}{100}$

$s(\{b, c, o\}) = \frac{21+14}{100} = \frac{35}{100}$

{apple, banana, carrot} has largest support 0.59.

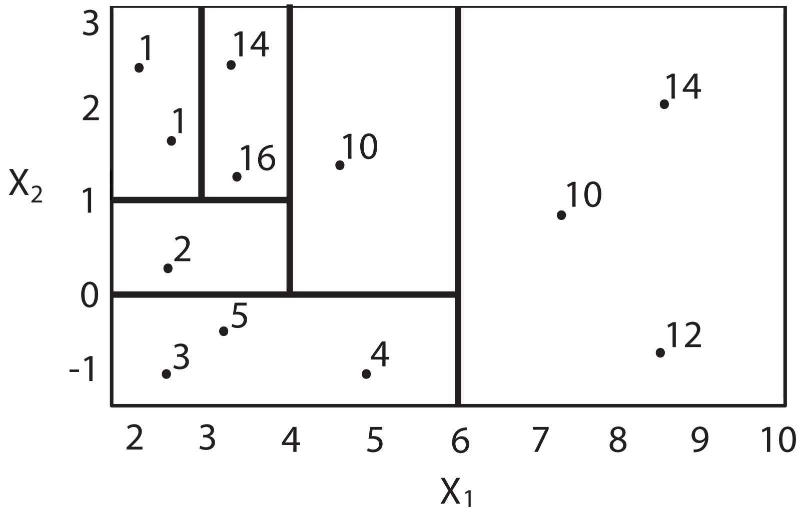
(d) No. Use $F_{k-1} \times F_{k-1}$ method, set k=4: we can merge {a, b, c} & {b, c, o} $\Rightarrow S(\{b, c, a, o\}) = 0.21 >$ minsup

So, {banana, carrot, orange} has a super frequent itemset {banana, carrot, orange, apple}.

5. **Classification and Regression Trees**

(a) Describe two potential advantages of regression trees over other statistical learning methods.

(b) This question uses the CART partition given below. The points in each box denote the training data and the numbers next to them are their response values.



i. Sketch the decision tree corresponding to the CART partition.

ii. Predict the response values for the following observations using your tree, where $\mathbf{x} = (x_1, x_2)$:

$$\mathbf{a} = (5, 1); \quad \mathbf{b} = (3.5, 2); \quad \text{and} \quad \mathbf{c} = (11, -2.)$$

iii. What is the training MSE for your tree?

(c) The predictive performance of a single regression tree can be substantially improved by aggregating many decision trees.

i. Briefly explain the method of bagging regression trees.

ii. Explain the difference between bagging and random forest.

iii. Briefly explain two differences between boosted regression trees and random forest.

End of Examination