TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI

VICTORIA
UNIVERSITY OF WELLINGTON

**EXAMINATIONS – 2020**

**TRIMESTER TWO**

**COMP 422**

**DATA MINING, NEURAL NETWORKS AND GENETIC PROGRAMMING**

**Time Allowed:**

**OPEN BOOK**

**Permitted materials:** This is an open-book test. You can use any material for doing the test.

**Instructions:** The expected workload of this test is 60 minutes (1 hour), but you have 24 hours to work on it.
There are a total of 60 marks on this test.
Attempt all questions.

## Questions

| | **Marks** |
|---|---|
| 1. Data Mining and Knowledge Discovery | [20] |
| 2. Computer Vision and Image Processing | [30] |
| 3. Performance Evaluation | [10] |

Student ID: ........................

1. **Data Mining and Knowledge Discovery** **(20 marks)**

(a) **(2 marks)** Data Mining (DM) is an important step in the Knowledge Discovery in Database (KDD) process. Briefly describe two steps before the DM step during the KDD process. For each step, state why it is necessary.

(b) **(3 marks)** Decision tree, naive Bayes and support vector machine are three commonly used data mining methods. However, they are not straightforward for some data types. For each of the following scenarios, briefly describe a strategy to handle the data type by the given method.

i) Classify data with continuous numeric features by decision tree.
ii) Classify data with continuous numeric features by naive Bayes.
iii) Classify data with categorical features by support vector machine.

(c) **(3 marks)** John is solving a classification task using Naive Bayes. The training data has 100 training instances, each with 10 binary features. There are two classes, A and B. He directly uses the following Bayes rule to calculate the conditional probabilities:

$$P(class|F_1,\ldots,F_{10}) = \frac{P(F_1,\ldots,F_{10}|class)|P(class)}{P(F_1,\ldots,F_{10})} \tag{1}$$

The class with the highest conditional probability will be the predicted class.

However, when he calculates the probabilities for a test instance $(f_1,\ldots,f_{10})$, he found out that $P(f_1,\ldots,f_{10}) = P(f_1,\ldots,f_{10}|A) = P(f_1,\ldots,f_{10}|B) = 0$ from the training data, that is, the test instance is never seen in the training data. Therefore, the above calculation cannot be used for classification.

Briefly describe three possible strategies that can help John overcome this issue.

(d) **(6 marks)** VC dimension can be used to measure the complexity of a group of classification models. Consider a dataset with a single feature and two classes (A and B), prove the following statements:

i) If $f$ is a threshold-based classifier with the parameter $\theta$, i.e. it returns A if the feature value is larger than the threshold $\theta$, and returns B otherwise, then its VC dimension is 1.
ii) If $f$ is an interval classifier with the parameter $\theta$, i.e. it returns A if the feature value is in the interval $[\theta, \theta+1]$, and returns B otherwise, then its VC dimension is 2.

(e) **(2 marks)** Support vector machine applies the principle of structural risk minimisation, which attempts to minimise both the empirical risk and VC dimension. Briefly describe how support vector machine minimises the empirical risk and VC dimension, respectively.

(f) **(4 marks)** Support vector machine is typically for binary classification (i.e. two classes). To use it for multi-class classification, one can use the one-against-rest and one-against-one strategies, to decompose the original multi-class classification task into a set of binary classification tasks. Given a 4-class movie classification task with the class labels {Action, Comedy, Drama, SciFi}, describe the set of binary classification tasks generated by

i) one-against-rest strategy.
ii) and ii) one-against-one.

For each binary classification task, describe the two class labels.

2. **Computer Vision and Image Processing** **(30 marks)**

(a) **(4 marks)** There are two errors in the following image file. Identify and briefly describe these two errors.

```
P2
# image file
24 8
15
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  3  3  3  3  0  0  7  7  7  7  0  0 11 11 11 11  0  0 25 25 25 25  0
0  3  0  0  0  0  0  7  0  0  0  0  0 11  0  0  0  0  0 25  0  0 25  0
0  3  3  3  0  0  0  7  7  7  0  0  0 11 11 11  0  0  0 25 25 25 25  0
0  3  0  0  0  0  0  7  0  0  0  0  0 11  0  0  0  0  0 25  0  0  0  0
0  3  0  0  0  0  0  7  7  7  7  0  0 11 11 11 11  0  0 25  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

(b) **(10 marks)** Assuming that the following first-order convolution mask is used,

$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

**enlarge** the following image array to a $5 \times 5$ array using the *convolution method*. Show the detailed steps of your working.

$$\begin{bmatrix} 8 & 6 & 4 \\ 6 & 6 & 6 \\ 4 & 6 & 8 \end{bmatrix}$$

(c) **(6 marks)** Briefly discuss the differences between Fourier image transform and Cosine image transform technologies. Identify two reasons for the Cosine image transform technology to be more popularly used for image compression.

(d) **(10 marks)** The SIFT algorithm is designed to retrieve robust local image features. For each of the following cases, identify and briefly explain one technique used by the SIFT algorithm.

i) To ensure that extracted local features are *rotation-invariant*;

ii) To ensure that extracted local features are *scale-invariant*;

iii) To ensure that extracted local features are *translation-invariant*;

iv) To ensure that extracted local features are *illumination-invariant*.

3. **Performance Evaluation** **(10 marks)**

(a) Assume that a classifier is applied to an image classification problem in the medical domain. There are two classes of images in the problem: images containing a *tumor* and images with *normal* characteristics. The learned classifier is applied to a test set with 100 *tumor* images and 400 *normal* images. At a threshold level of 0.80, the classifier reports 200 objects for class *tumor* of which 60 are correct. Assume that all the given images are classified as either class *tumor* or class *normal* by the classifier.

**Show your working.**

i) **(2 marks)** Calculate the overall accuracy of the classifier.

ii) **(2 marks)** Calculate the TPR and the FPR for class *tumor*.

iii) **(3 marks)** Draw an ROC curve (for class *tumor*) using the results obtained in part *ii)* and the additional results from the following table:

| **Threshold** | **0.60** | **0.70** | **0.85** | **0.90** |
|---|---|---|---|---|
| **TPR(%)** | 85 | 70 | 50 | 30 |
| **FPR(%)** | 70 | 50 | 25 | 10 |

(b) **(3 marks)** Sarah is evaluating the performance of a neural network classifier for a medical diagnosis task on a dataset, which consists of 9900 negative instances and 100 positive instances. She wants to use the standard ROC curve to measure the performance.

Give *one* possible issue of using the standard ROC curve in this case, and suggest *two* ways to address the issue.

* * * * * * * * * * * * * * *