COMP10200: Machine Learning Test 1

Mohawk College, Fall 2018

Name: _____

Instructions

- You have 110 minutes to complete this test.
- There are 11 pages on this test, including this cover page.
- You may use a scientific calculator.
- If you run out of room, you can answer on the backs of the pages.
- No other paper or electronic aids are allowed.
- Read the questions carefully and make sure you answer all parts of every question.
- The point value of each question is shown in [square brackets].
- The test is out of 100 marks and is worth 25% of the course grade.
- Good luck!

Use Abbreviations to Write Faster

Feel free to use abbreviations when answering these questions. Just make sure it's clear what you mean.

For example, if the words "precision" and "recall" are in the question, you could abbreviate them P and R in your answer. However, if those words were not in the question, you might need to write them out the first time and then put the abbreviations in brackets: "It might be a good idea to use Precision (P) here. P would be helpful because…"

_____ / 100

COMP10200: Supervised Classification Test

Sam Scott, Mohawk College, Fall 2018

Knowledge and Understanding ____ / 51

These questions test your basic knowledge of machine learning concepts and techniques. Be brief, answer all parts of the question, and use the correct terminology.

1. Describe one difference and one similarity between the `list` and `dict` types in Python. [3]

2. Describe two fundamental differences between Python `list` type and the Numpy `array` type. [3]

3. When preparing data for machine learning, we typically use *related arrays* for the training data and the training labels. What does it mean to say that those two arrays are *related*? [2]

4. When developing a machine learning system, why do we split the data into training and testing sets? Explain the role of each of these sets. [3]

5. What does it mean for a machine learning system to *overfit* the training data? Why is overfitting a problem? How can a decision tree learning algorithm be prevented from overfitting? [5]

   What it is:

   Why it's a problem:

   How to avoid overfitting for decision trees:

6. What does the k parameter control in the k-Nearest Neighbour algorithm? [2]

7. Why does it often help to normalize the data before running the k-Nearest Neighbour algorithm? Explain using an example. [3]

8. Describe one method of breaking a tie vote in the k-Nearest Neighbour algorithm. [2]

9. What does it mean to say that decision tree learning is a *divide-and-conquer* technique? Explain what *divide-and-conquer* means, and explain how decision tree learning uses this technique. [4]

10. Describe two different parameters in *sklearn* that can be used to alter the performance of a decision tree learning system and explain the effect that varying these parameters might have on the decision tree produced by the system. [5]

11. Explain the difference between $p(CLASS|DATA)$ and $p(DATA|CLASS)$. Which of these is the target value for a Naïve Bayes classifier? [3]

12. Explain the "Naïve" assumption that gives Naïve Bayes its name. What is the benefit of making this naïve assumption? [3]

13. What information is computed during the training phase (not the testing phase) for Gaussian Naïve Bayes? [3]

14. List and briefly explain 3 common techniques used in text representation for machine learning. [6]

Technique	Explanation

15. Explain, using an example, why precision and recall might be better than accuracy as a measure of performance when the data is very unbalanced. [4]

Application _____ / 39

The questions in this section ask you to apply your knowledge of machine learning techniques. Answer all parts of the questions and show all of your work.

16. Write a single line of code below that will print the maximum row sum of a two-dimensional *numpy* array called `mynums` (example shown at right). [4]

4	5	7
1	2	3
0	0	5

← Maximum row sum = 16

```
import numpy as np
```

_____

17. Suppose you have a set of machine learning data and labels stored in *numpy* arrays named `data` and `labels`. Write a single line of code below that will print the average of the first feature value for the examples with a class label of 3. [5]

```
import numpy as np
```

_____

18. Given the machine learning training data below, simulate the k-Nearest Neighbour algorithm using Euclidean distance and $k = 3$ for the previously unseen example. Show all of your work as well as the final classification result. [7]

Training Data

#	Feature1	Feature2	Class
1	10	100	A
2	9	90	A
3	3	30	B
4	4	40	B

Previously Unseen Example

Feature1	Feature2	Class
3	90	???

19. Do you think Normalizing the data will help with classification accuracy for the data in the previous question? Why or why not? [3]

20. From the data given below, use the Naïve Bayes method to decide whether the given document should be categorized as POS or NEG. Make sure you compute and state the relevant probabilities. Assume the Bag of Words (Multinomial) representation. To keep this question simple, the vocabulary contains only two words: "Terminator" and "liked". Show all your work. [5]

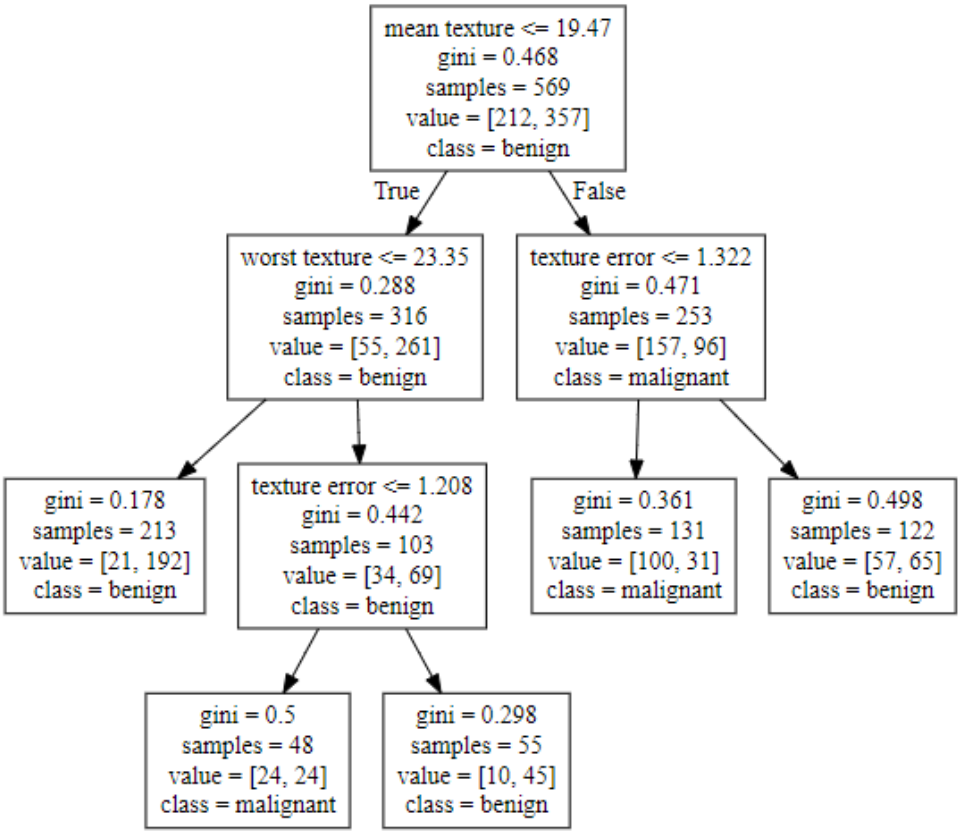
    document = "I liked Terminator 2 almost as much as Terminator 1."

    $p(\text{“}Terminator\text{”}|NEG) = 0.2$ $p(\text{“}Terminator\text{”}|POS) = 0.5$
    $p(\text{“}liked\text{”}|POS) = 0.9$ $p(\text{“}liked\text{”}|NEG) = 0.1$
    $p(\text{“}Terminator\text{”}) = 0.1$ $p(\text{“}liked\text{”}) = 0.5$
    $p(NEG) = 0.3$ $p(POS) = 0.7$

21. Fill in the predictions in the table below for the decision tree shown at the bottom of the page and compute the accuracy of the decision tree for this test data. [5]

Mean Texture	Worst Texture	Texture Error	Class	Prediction	Accuracy
14.3	34.4	1.1	Malignant		
20.0	23.1	1.5	Benign		
18.5	22.0	2.3	Benign		_____
21.1	25.3	1.1	Benign		

Do you think it's more likely that this decision tree overfits or underfits the data? Justify your answer using only the information contained in the diagram of the tree. [4]



22. A machine learning algorithm is trying to find faces in photographs. It categorizes photographs as F (contains a face) or NF (does not contain a face). From the performance summarized in the table below, create a confusion matrix and calculate precision, recall, and accuracy for the task of finding faces. Show all your work. [6]

Actual	F	NF	F	NF	F	NF	F	NF	F	F
Predicted	F	F	F	F	NF	NF	NF	NF	F	NF

Problem Solving _____ / 10

These questions ask you to reflect on and synthesize what you have learned about machine learning. Use these questions as an opportunity to show us how well you understand and can reason about what we have covered so far.

23. You are a consultant for a company that makes early warning systems to predict volcanic eruptions from seismic and geothermal data. They have training data consisting of thousands of numeric features and hundreds of thousands of examples. They are considering systems based on k-Nearest Neighbour, Naïve Bayes, and Decision Trees. Make an initial recommendation from among these three options and give two good reasons to prefer your recommended system over the others. [5]

24. Consider the quote below from Jeff Hawkins, inventor of the Palm Pilot in the 90's (one of the first handheld computing devices). Do you think that what he is saying could be a valid criticism of Machine Learning systems? Justify your response using specific references to how Machine Learning systems work. [5]

    "… scientists tried to program computers to act like humans without first understanding what intelligence is and what it means to understand. They left out the most important part of building intelligent machines, the intelligence …"