COMP10200: Machine Learning Test 1

Mohawk College, Fall 2019

Name: _____

Instructions

- You have 110 minutes to complete this test.
- There are 15 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 83 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…"

_____ / 83

COMP10200: Supervised Classification Test

Sam Scott, Mohawk College, Fall 2018

Knowledge and Understanding ____ / 27

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

1. Give an example of one useful thing you can do with numpy arrays that you can't do with Python lists. Show the syntax, and also explain what the syntax does that Python lists don't or can't do. [2]

2. Why do we split data into testing and training sets when we're trying out a machine learning algorithm? Why don't we just train and test the algorithm on all the data? [2]

3. I tried kNN and Decision Trees on my data. For kNN, k=5 gave me 10% higher accuracy on the test set than k=1. For Decision Trees, forcing it to stop when a leaf node has 10 items in it gave me 10% higher accuracy on the test set than using the default SKLearn parameters. Do these results make sense? Why or why not? What, if anything, do these results tell me about my data? [4]

4. What is one big drawback of kNN? Explain why this drawback is a feature of kNN. [2]

5. List 2 ways to break tie votes in the kNN algorithm. [2]

6. One of the big points in favor of Decision trees is their interpretability. Explain what this means and give an example of why it might be important. [2]

7. What is a stopping criterion for a Decision Tree algorithm? What are two different stopping criteria implemented in SKLearn? (If you don't remember the exact names, just describe them.) [3]

8. State Bayes' Rule. Then briefly explain how the rule is used to make classifications in a Naïve Bayes classification system. You should state, briefly, how each of the probabilities is computed as part of your answer. [4]

9. What is the Naïve Assumption in Naïve Bayes? Why is the Naïve Assumption considered Naïve? [2]

10. List one similarity and one difference between a “Bag of Words” representation and a “Bag of Stems” representation? [2]

11. What are “stop words”? What are text classification systems most likely to do with stop words? [2]

Application _____ / 56

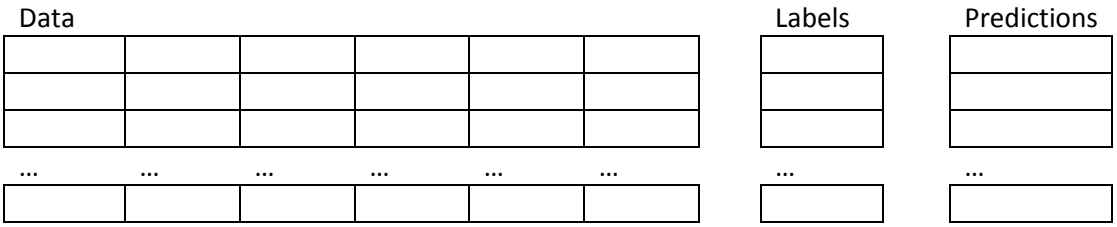
The questions in this section ask you to apply your knowledge of machine learning techniques. Answer all parts of the questions and show all of your work.

12. Numpy Questions (assume numpy is imported as np)

a. Suppose you have two numpy arrays named “scores” and “names”. They are related (parallel) arrays storing the scores and names of golf players. Write a single numpy expression that will return an array containing the name of every player who scored 0 or less. [2]

b. Suppose you have three related (parallel) numpy arrays holding information from a machine learning system (diagram shown below). The arrays are called “data” (holds the testing data), “labels” (holds the class name for each example), and “predictions” (these came from your ML system). Write a single numpy expression that will return an array containing the first feature value for each incorrect classification. [3]

This picture shows the structure, but not the contents, of the 3 arrays.



c. Suppose you have a 2D numpy array called “a”. Write a single numpy expression that will produce an array of all the mean values of each column. [2]

13. Consider the training data below for a categorization task. The features are F_1 through F_4.

F_1	F_2	F_3	F_4	CLASS
1	20	6	15	A
4	18	10	15	B
5	17	6	11	A
3	16	7	13	A
1	16	6	15	B
4	19	7	12	B

Simulate the kNN algorithm using Manhattan distance and k=5 to categorize the new example below. Make sure you show all your work. For this question, the process you use to arrive at the answer is worth a lot more than the answer itself. [7]

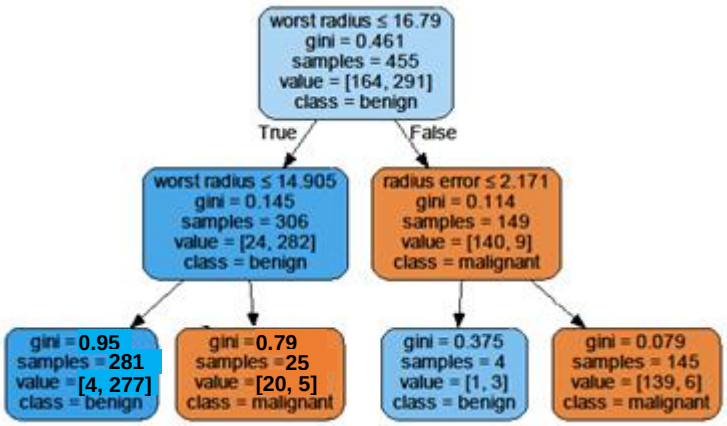
F_1	F_2	F_3	F_4	CLASS
1	19	10	12	?

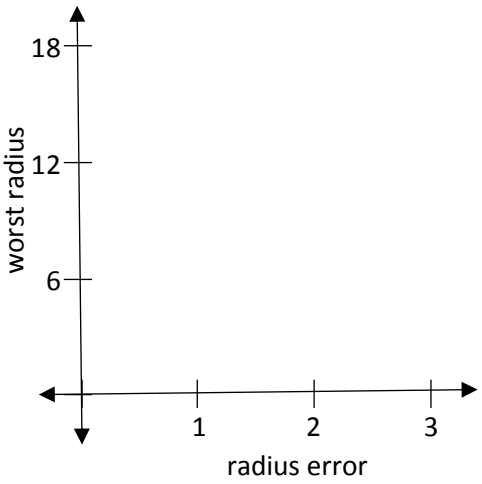
b. Using the procedure shown in class, compute the normalized version of the first training example from the previous question. Make sure you show your work. [5]

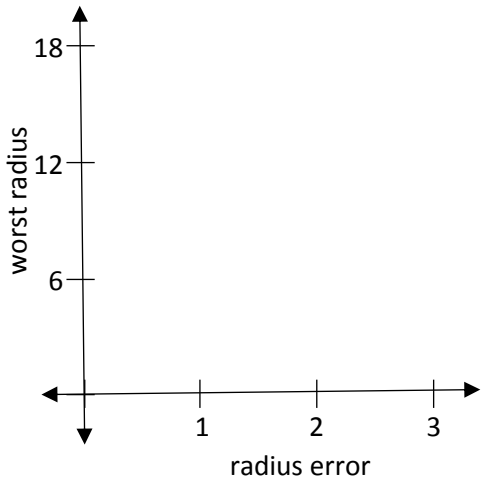
c. Do you think normalizing the data will make a difference in this task? Explain why or why not. [3]

14. Decision Trees

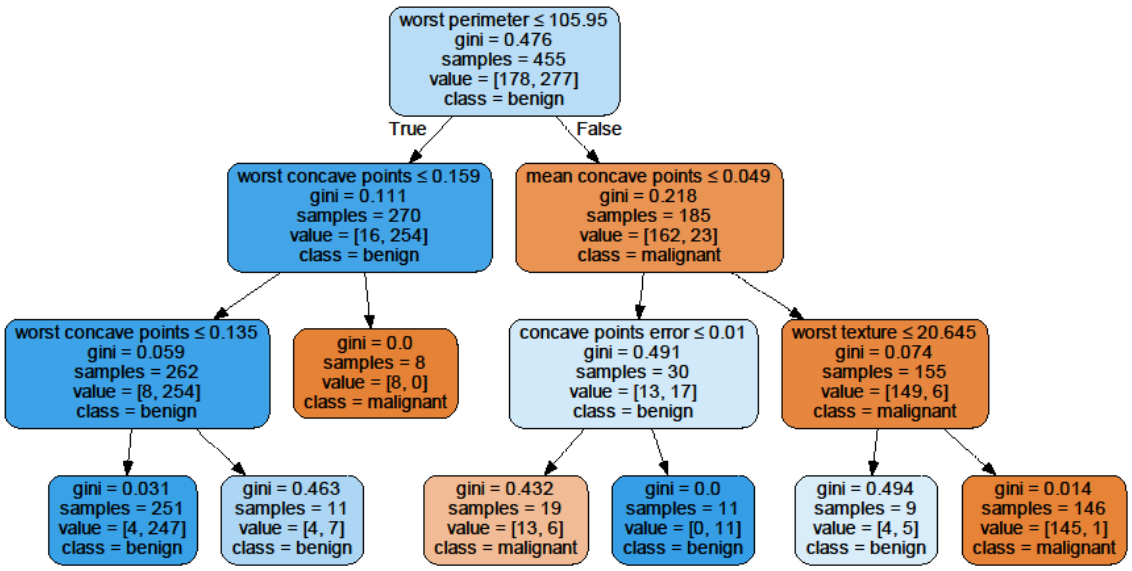
a. The decision tree below is a modified version of one that was produced using only two features of the Wisconsin Breast Cancer data set. On the axes below, draw the decision boundaries created by this tree and label each region with its categorization. You have been given two sets of axes in case you make a mistake and want to begin again. [7]







b. The decision tree below was produced using the Wisconsin Breast Cancer data set. Use the tree to categorize the 5 numbered examples shown below. Circle the leaf node that you end up at in each case and number the circles from 1 to 5. [5]



	worst perimeter	worst concave points	mean concave points	concave points error	worst radius	radius error	worst texture
1	104.3	0.17	0.05	0.005	15.4	2.1	21.5
2	107.2	0.12	0.05	0.015	22.3	1.5	22.1
3	106.1	0.19	0.04	0.002	16.7	1.9	20.1
4	103.3	0.13	0.03	0.009	18.9	2.3	20.2
5	106.4	0.16	0.06	0.011	19.1	1.6	20.5

Categorizations:

1:_____ 2:_____ 3:_____ 4:_____ 5:_____

Don't forget to circle and label the leaf node in the tree for each categorization. You must get both the categorization and the leaf node correct.

15. Naïve Bayes Classification

a. Below are all the relevant probabilities that a Gaussian Naïve Bayes classifier would need to be able to predict whether a mortgage customer is a good risk or a bad risk. Perform all the necessary calculations on the customer shown, and then state whether the system would judge the customer to be a good risk or a bad risk. Show all your work! [7]

Customer: missed payments **m = 2**, income **i = 70000**, credit score **c = 500**, debts **d=14000**

P(BAD) = 0.3 P(GOOD) = 0.7 P(m=2) = 0.5 P(i=70000) = 0.6 P(c=500) = 0.4 P(d=14000) = 0.4

P(m=2|GOOD)= 0.5 P(i=70000|GOOD) = 0.8 P(c=500|GOOD) = 0.4 P(d=14000|GOOD) = 0.3

P(m=2| BAD)= 0.5 P(i=70000| BAD) = 0.25 P(c=500| BAD) = 0.76 P(d=14000| BAD) = 0.8

b. After the calculations in the last section, would you say that the classifier is making its prediction with a high confidence or a low confidence? Explain your reasoning. If you were unable to complete the last part of the question, instead you should give examples of what might be considered high or low confidence predictions and explain why they are high or low confidence. [2]

c. We are using a Multinomial Naïve Bayes classifier to classify Instagram comments as positive or not. We're doing it based on the frequencies of just the two word stems "lov" and "whatev". (Obviously this is not a very realistic example.) We have a corpus of 1000 twitter posts (400 positive, 600 negative) and we've counted the total number of occurrences of those stems for each class.

	"lov"	*"whatev"*
POSITIVE	100	50
NOT POSITIVE	25	75

We are categorizing a new post which contains both "lov" and "whatev" twice. Do the necessary calculations to figure out whether the system will categorize this as POSITIVE or NOT POSITIVE. You do not have to compute the full probabilities, just do the calculations that are necessary. Make sure you show your work.[5]

16. Consider the two confusion matrices below for a Naïve Bayes classifier and a Decision Tree classifier. Both represent performance categorizing Instagram comments as POSITIVE (+) or NEGATIVE (-). The purpose of this classification is to push ads to a user that are related to things they have expressed positive views about on Instagram.

    Compute the accuracy of each classifier, then compute other relevant statistics to help you evaluate their performance. Finally, state which algorithm is doing a better job. Say why you think it is doing a better job with reference to the statistics you calculated. [8]

Naïve Bayes

	PREDICT +	**PREDICT -**
TRUE +	300	100
TRUE -	100	500

Decision Tree

	PREDICT +	**PREDICT -**
TRUE +	400	0
TRUE -	200	400

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.

17. You are designing a system that will use text classification to suggest diagnoses from the written notes of doctors, nurses, and other health professionals. The finished system will suggest diagnoses in real time as the health professionals are typing up their notes. You have a database of one million written notes (average size 250 words each). These notes have been categorized by placing labels on them indicating which diseases the patient was eventually diagnosed with. You are considering k-Nearest Neighbour and Decision Trees, using a simple bag of words representation. Make an initial recommendation and give two good reasons to prefer the recommended system over the other. [5]

18. Consider the quote below from Elon Musk. Do you think that what he is saying could be seen as a valid criticism or warning relating to Machine Learning systems? Explain what it is that you think Elon might be worried about, and then say whether or not you think he is right to worry. Justify your response using specific references to how Machine Learning systems actually work. [5]

    "With artificial intelligence, we are summoning the demon. You know all those stories where there's the guy with the pentagram and the holy water and he's like, yeah, he's sure he can control the demon? Doesn't work out."