Queen's

Master of Management in Artificial Intelligence

MMAI 869

Machine Learning and AI

Dr. Stephen W. Thomas

Individual Assignment 1

Click Here and Input Due Date

Click Here and Enter Individual or Team Name

How I will Mark This Assignment

I will be using a relatively new marking system, called Specs Grading. Specs Grading is a research-backed system that has been shown to improve both the student learning experience as well as the instructor experience.

It is not important for you to understand Specs Grading – that's my job. All you need to know is the following. When I mark this assignment, I will give each question one of three possible marks:

- **Does not meet expectations (0 pts)**. The question is not completed, contains obvious errors, or is generally not satisfactory in some way.
- **Meets expectations (1 pts)**. The question is completed as specified, with correct (or reasonably close) answers and justifications, and it is obvious that you have learned what I intended you to learn.
- **Exceeds expectations (1.25 pts)**. The question is completed exceptionally well. You have gone above and beyond in some way. You have gone the extra mile to experiment, dive deeper, or in general, learn even more than I intended you to.

In Specs Grading, there is no partial credit. If a question has 4 parts, and you only answer 3 of them, then you will earn a "Does not meet expectations." Please answer all the parts of the question to earn a "Meets expectations" or above.

Your total mark will be a summation of each question, divided by 10. Therefore, if you earn a "Meets expectations" on all eight questions, you will earn a total mark of 80% = A-.

1. HOW LOVELY!

[For this problem, please answer the questions below in English, and also turn in your code. If possible, your code should be submitted as a notebook – either a Python Jupyter Notebook (.ipynb) or an R Markdown file (.Rmd), along with a PDF rendering of the notebook. Make your code completely reproducible. E.g., set random number seeds.]

You work at a local jewelry store. You've recently been promoted and the store owner asked you to better understand your customers. Using some sneaky magic (and the help of Environics!), you've managed to collect some useful features for a subset of your customers: age, income, spending score, and savings. Use these features to segment your customers and create customer *personas*.

1. Download the customer dataset: *jewelry_customers.csv.*
2. Perform a clustering analysis of the dataset.
    a. Try different values of parameters (e.g., K for K-means).
    b. What do you think the best parameter values are? Why?
3. Describe and interpret the clusters.
4. How good are the results?

2. CATCHING RECIDIVISTS BEFORE THEY STRIKE

A ***recidivist*** is a criminal that was released from prison, but commits another crime. You are a warden at a maximum-security prison in Kingston, and you want to determine which prisoners will likely become recidivists. Luckily, you have a Queen's degree, so you are going to take a data-driven approach. You have collected some historical training data that include some basic metadata, and whether the prisoner ended up becoming a recidivist or not.

Given the training data below, use the ID3 algorithm and entropy-based information gain to construct a decision tree by hand to predict which prisoners will become recidivists. Show all the steps and follow the algorithm closely. Use the resulting decision tree to predict the class of the following prisoner: Good Behavior = false, Age < 30 = false, Drug dependent = true.

id	Good Behavior	Age < 30	Drug Dependent	Recidivist
1	False	True	False	True
2	False	False	False	False
3	False	True	False	True
4	True	False	False	False
5	True	False	True	True
6	True	False	False	false

3. More Classification Measures

In class, we talked about several classification measures, such as accuracy, precision, recall, AUC, etc. There are many other measures out there, each with its own pros and cons.

Do a bit of research of your own to find at least two classification measures that we did not discuss in class. For each measure, describe what it is, how it is calculated, how it is different from accuracy, and in which scenarios it is best used. Show some example datasets/confusion matrices to illustrate your points.

4. THE INTERN

You are an in-demand, world-traveling, work-all-night consultant who specializes in designing supervised machine learning solutions for clients in a wide-range of industries. You have seen it all and you know what to do. To help you get more done in less time, you have hired an intern from Ivey, who, unfortunately, needs some handholding. Your intern does not understand when to use which classification measure. Your intern keeps getting it wrong. To help your intern learn from your experience, you have decided to look at some previous projects and describe which measure you used, and more importantly, why.

For each project below, describe which measure(s) are best, and why. Also, give an example of a measure which would be horrible to use, and why. List any assumptions you are making, about the dataset, problem, or business priorities that were involved in the project.

a) The fraud department at a bank wanted to predict which transactions were fraudulent. The training dataset had 100K credit card transactions, of which 97K are legit and 3K are fraud.
b) A hospital wanted to predict whether a MRI scan contained cancer.
c) An IT team wanted to filter spam from email inboxes.
d) A sports analytics department wants to predict which team will win the match.
e) A city government wanted to build a system to monitor Twitter to see if any local residents were tweeting about emergencies that needed quick response from the police department. They don't trust Twitter that much; they only want to send police in true emergencies.
f) *[Describe one more project, whereby the best measure is one that you have not yet listed in parts a-e above.]*

5. UNCLE STEVE'S GROCERY STORE

Uncle Steve runs a small, local grocery store. Looking for some customer insights, he has hired you to do some data science. He has given you a few years' worth of customer transactions, i.e., sets of items that customers have purchased. You have applied an association rules learning algorithm to the data, and the algorithm has generated a large set of association rules.

For each of the following scenarios, provide an example of one of the discovered association rules that satisfies the following conditions. (Just make up the rule, using your human experience and intuition!) Also, describe whether and why each rule would be considered subjectively interesting or uninteresting for Uncle Steve.

a) A rule that has high support and high confidence.

b) A rule that has reasonably high support but low confidence.

c) A rule that has low support and low confidence.

d) A rule that has low support and high confidence.

6. VIVA LA VINO

Some Smith faculty have started a wine club. At each meeting, members of the club perform blind taste tests of different wine varietals. Members indicate how much they enjoy each varietal, using an integer scale of 1 (worst) to 7 (best). After the most recent meeting, here are the ratings.

	Zin	Pinot Noir	Chard	Merlot	Cab	Pinot Gris
Yuri	7	6	7	4	5	4
Steve		7	6	4	3	4
Gary	3	3	3	1	1	5
Qurat	2	2	1	3	7	4
Brigid	5	6	7	2	3	3

Unfortunately, the club ran out of Zin before Steve had a chance to try it. Luckily, the club has you, a data-driven, clever, and charming Queen's student. Use your skills to predict what Steve would rate Zin. Use user-based collaborative filtering with cosine distance.

7. YUM, ORANGE JUICE!

[For this question, please answer the parts below in English, and also turn in your code. Your code should be submitted as a notebook – either a Python Jupyter Notebook (.ipynb) or an R Markdown file (.Rmd), along with a PDF rendering of the notebook.]

One cup of fresh orange juice has 124 mg of vitamin C, which is 200% of the recommended daily intake of vitamin C for an adult. With this as (completely unrelated) motivation, build a model to predict whether a grocery store customer will Purchase Citrus Hill (CH) or Minute Maid (MM) orange juice.

1. Download the file *OJ.csv*. The target feature is *Purchase*. The rest of the features are self-explanatory, hopefully.
2. Preprocess the data however you see fit. Describe what you did and why.
3. Split the data into training and testing sets. Describe what you did and why.
4. Choose an appropriate metric to analyze a model's performance. Justify.
5. Build five different models, using five different classifier algorithms. (Any five will do.)
    a. Tune each model. What were the best parameter values for each model?
6. Describe and compare the performance of each fine-tuned model.
7. Select the best model. Justify.
8. Is this model good enough to deploy today? Justify.