**MMA 869**

**Machine Learning and AI**

**Dr. Stephen W. Thomas**

**Individual Assignment 1**

Version 2: Updated July 7, 2020

**Click Here and Input Due Date**

**Click Here and Enter Your Name**

# Assignment Instructions

This assignment contains seven (7) questions. Each question has one or more tasks. There are three types of tasks: tasks that require you to write code; tasks that require you to write text responses; and tasks require you to write math and/or diagrams. A detailed grading rubric is provided on D2L for each question.

For tasks that require **code**:

* Use Python (preferred) or R to complete the task.
* Use the template file provided (Assignment1\_869.py for Python or Assignment1\_869.R for R] and rename it to contain the question number and your full name without spaces (e.g., Assignment1\_869\_Q1\_StephenThomas.py). Do not submit Jupyter Notebooks (.ipynb) or RMarkdown (.Rmd) files.
* Submit code that runs without errors.
* Submit code that is reproducible. E.g., set random number seeds as appropriate. You should be able to run you code again and again and again, from the top of the file to the bottom of the file, and get the exact same results each time. I should be able to run your code, from scratch, on my machine, again and again, and get the exact same results that you get.
* Submit code that is organized. Make your code readable. Provide comments to describe what the code is doing and why. Don’t leave “old” code laying around. Overall, if your code is clear and easy to read, then we will be happy. When we are happy, we give better marks.

For tasks that require **text responses**:

* Type your response in this Word document.
* Be clear about which task you are responding to.
* Use English. Use proper grammar, spelling, and punctuation. Be professional and clear. Be complete, but not overly-verbose.
* You may refer to your code. Please do so very clearly. E.g., “As can be seen in on line X of file Y…”

For tasks that require **math**or **diagrams**:

* Type or insert your response in this Word document.
* Be clear about which task you are responding to.
* Where appropriate, please use Word’s features such as Equations, Symbols, or Diagrams. You may also import diagrams created in other programs, as long as they are clear and easy to read.
* Please do not submit pictures/scans of hand-written diagrams, equations, or work of any kind.

Your assignment submission should contain exactly three files:

* Assignment1\_869\_Q1\_FirstLast.py or Assignment1\_869\_Q1\_FirstLast.R
* Assignment1\_869\_Q7\_FirstLast.py or Assignment1\_869\_Q7\_FirstLast.R
* Assignment1\_869\_FirstLast.docx

# How Lovely!

## Preamble

Download the “customer” dataset: *jewelry\_customers.csv.*

## Scenario

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 (i.e., how much savings they have in their bank account). Use these features to segment your customers and create customer *personas*.

## Tasks

1. [Code] Perform a clustering analysis of the dataset.
   1. Load, clean, and preprocess the data as you find necessary.
   2. Cluster the data using any clustering algorithm discussed in class. Measure goodness-of-fit. Try different values of hyper parameters to see how they affect goodness-of-fit.
   3. Print summary statistics for each cluster.
2. [Text] What do you think the best hyper parameter values are? Why?
3. [Text] Describe and interpret the clusters with words. That is, create personas.
4. [Text] How good are your results? What could you do to make them better?

# Catching Recidivists Before They Strike

## Scenario

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.

## Tasks

1. [Math and Diagrams] Given the training data below, use the ID3 algorithm and entropy-based information gain to construct a decision tree by hand. Show all the steps and follow the algorithm closely. Show the resulting decision tree. Show the “summary statistics” at each leaf node.
2. [Text] Use the resulting decision tree to predict the class of the following prisoner: Good Behavior = false, Age < 30 = false, Drug dependent = true. How reliable is the prediction?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

# More Classification Measures

## Scenario

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.

## Tasks

1. [Text] 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 business scenarios it is best used. Show some example datasets/confusion matrices to illustrate your points.

# The Intern

## Scenario

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.

## Tasks

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.

1. [Text] 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.
2. [Text] A hospital wanted to predict whether a MRI scan contained cancer.
3. [Text] An IT team wanted to filter spam from email inboxes.
4. [Text] A sports analytics department wants to predict which team will win the match.
5. [Text] 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.
6. [Text] *Describe one more project, whereby the best measure is one that you have not yet listed in Tasks 1-6 above.*

# Uncle Steve’s Grocery Store

## Preamble

*Note: there is no actual data for this question. This question is just a thought exercise. We are using our intuition and understanding of the real world; we are* not *using actual data.*

## Scenario

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.

## Tasks

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, and why.

1. [Text] A rule that has high support and high confidence.
2. [Text] A rule that has reasonably high support but low confidence.
3. [Text] A rule that has low support and low confidence.
4. [Text] A rule that has low support and high confidence.

# Viva la Vino

## Scenario

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.

Recall the following distance and similarity formulas for the Euclidean and Cosine metrics:

|  |  |
| --- | --- |
|  |  |
|  |  |

## Tasks

Use your skills to predict what Steve would rate Zin.

1. [Math] Use user-based (k=2) collaborative filtering with cosine distance.
2. [Math] Use item-based (k=3) collaborative filtering with Euclidean distance.

# Yum, Orange Juice!

## Preamble

Download the file *OJ.csv*. The target feature is *Purchase*. The rest of the features are self-explanatory, hopefully.

## Scenario

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, your task is to build a model to predict whether a grocery store customer will Purchase Citrus Hill (CH) or Minute Maid (MM) orange juice.

## Tasks

1. [Text] Choose an appropriate metric to analyze a model’s performance. Justify.
2. [Code] Build a prediction model as follows:
   1. Preprocess the data however you see fit. In code comments, describe what you did and why.
   2. Split the data into training and testing sets. In code comments, describe what you did and why.
   3. Build three different models, using three different machine learning algorithms. (Any three will do.) Tune each model. Print out the best hyperparameter values for each model. Print out performance of each fine-tuned model.
3. [Text] Using business language (not technical language), describe and compare the performance of each fine-tuned model.
4. [Text] Overall, which model is best suited to this business problem? Justify.
5. [Text] Is this model good enough to deploy today? Justify. If you had more time, how could you make the performance of your model better?