MCIS6273 Data Mining (Prof. Maull) / Spring 2025 / HW1

Points Possible	Due Date	Time Commitment (estimated)
40	Sunday May 4 @ Midnight	*up to* 20 hours

- **GRADING:** Grading will be aligned with the completeness of the objectives.
- **INDEPENDENT WORK:** Copying, cheating, plagiarism and academic dishonesty *are not tolerated* by University or course policy. Please see the syllabus for the full departmental and University statement on the academic code of honor.

OBJECTIVES

- Perform basic unsupervised learning with K-Means.
- Perform supervised Learning with Naive Bayes Classifier
- BONUS: Learn about machine learning ethics.

WHAT TO TURN IN

You are being encouraged to turn the assignment in using the provided Jupyter Notebook. To do so, make a directory in your Lab environment called `homework/hw0`. Put all of your files in that directory. Then zip or tar that directory, rename it with your name as the first part of the filename (e.g. `maull_hw0_files.zip`, `maull_hw0_files.tar.gz`), then download it to your local machine, then upload the `.zip` to Blackboard.

If you do not know how to do this, please ask, or visit one of the many tutorials out there on the basics of using zip in Linux.

If you choose not to use the provided notebook, you will still need to turn in a `.ipynb` Jupyter Notebook and corresponding files according to the instructions in this homework.

ASSIGNMENT TASKS

(50%) Perform basic unsupervised learning with K-Means.

In slides and the book you learned about unsupervised learning, where the goal of this technique is to take data for which there are no classes *a prior* and have the algorithms learn patterns for the data.

These techniques are popular and widely used in *exploratory data analysis* (EDA). One of the most widely used and sucessful techniques (and often the first "go to" algorithm to try) is K-Means.

K-Means works by placing all data into k clusters, usually provide at the beginning of the process. Such clustering can be very successful and provide a basis for labeling data in the future, should the clusters prove to be valid.

We will use K-Means on the chocolate data to determine just what clusters are in the data, and to explore how to characterize these clusters.

You will find the starter notebook:

- Example notebook to complete

invaluable and if you use it, your work will be greatly reduced.

You will also find the following resources necessary to study:

- SciKit Learn K-means

You may use an AI-copilot for this part, but you must *declare that such a tool was used in producing your work*. I will otherwise grade as if your work was plagiarize – that is, if you do not declare use of AI *and* I find that the work is **not** original, it will be penalized as such.

§ Task: Use the chocolate data to cluster it into three clusters ($K = 5$).

Show your work in the notebook provided.

§ Task: Use the same data to cluster it into 8 and 11 clusters ($K = 8$, $K = 11$).

Show your work in the notebook provided.

§ Task: Characterize the clusters using the indices produced from K-means (K=5).

Use the functions provided `generate_cluster_report()` and `get_cluster_characteristics()` to answer this question.

You will need to run these functions, play with them then crea

(30%) Perform supervised Learning with Naive Bayes Classifier

In the previous, you learned how to do unsupervised learning with K-Means, which does not need labeled data to work.

In this part, you will get your feet wet with unsupervised learning using Naive Bayes classification.

In some of the lecture notes, you learned that Naive Bayes can be used to use prior probabilities of know (and unknown) data to determine how well a hypothesis fit data. We thus characterize Bayes like this:

$$\Pr(C|D) = \frac{\Pr(D|C)\Pr(C)}{\Pr(D)}$$

We know that $\Pr(D)$ is a constant and can be dropped in our calculations without loss of generality.

Given this, there are several ways to perform Naive Bayes, and we will be using it to do *classification tasks*. A classification task requires that the target classes of training data are known *a priori*, and we are given *test data* without classes and allow the classifier to assign the class of test data once training is completed.

In this assignment we will use the Bernoulli classifier in ScikitLearn. This classifier is well suited for multivariate binary-valued data such as that which we prepared in the first assignment.

The decision rule for Bernoulli NB is:

$$\Pr(d_i|C) = \Pr(d_1 = 1|C)d_1 + (1 - \Pr(d_i = 1))(1 - d_i)$$

I have provide a notebook which you will complete. The notebook is on Github here:

- Example notebook to complete

§ Task: Implement test-training set so that the test set is 1000 samples and the test set 1789 samples.

You must include this implementation in your notebook.

§ Task: Interpret the output of the classifier.

After you have run the cell `clf.predict(X[:])` you will see the list of numbers which represent the class labels from the training data.

Explain the output and answer the following questions:

- What is classifier accuracy? Show you answer in the notebook.
- What recommendation might you give to improve the classifier?

(0%) BONUS: Learn about machine learning ethics.

With the increasing rise of machines and AI in human decision making and human activities, we are increasingly in need of critical conversations about the ethics of AI – arguably the conversation is incomplete around the ethics of data mining and data science writ large, so this conversation will act as a proxy and extension of that broader conversation.

You will listen the podcast *The Machine Ethics* podcast which covers wide ranging conversations about AI and Ethics and brings to the fore relevant conversations about machine driven decision making and the intersection with human beings. You will learn about “digital sociology” and the relevant impact this field has on how we might develop human-machine boundaries, especially in critical decision making spaces that intersect with human society.

Listen to **Episode #93** (October 3, 2024):

- “Socio-technical systems with Lisa Talia Moretti” / duration: 61m49s

You will need to absorb as much as you can and take notes. There will be an open note assessment on Blackboard, which will ask approximately 10 questions relevant to the talk, so it is best you actively listen to this fascinating conversation.

§ Task: Listen to the podcast and do the companion assessment.

Once you are done, there will be an online assessment about the podcast, which will be available on or after May 1, 2025 through the last day of the final (which will be posted).