CS 273P: Machine Learning and Data Mining

Homework 1

Due date: **11:59 PM, Jan 20, 2023**

Instructor: Xiaohui Xie

This homework (and many subsequent ones) will involve data analysis and reporting on methods and results using Python code. You have to submit 1) **`a single PDF file`** that contains everything to Gradescope, and associated each page of the PDF to each problem, 2) **`a single zip file`** containing source codes to Canvas. The PDF file should include any text you wish to include to describe your results, the complete code snippets of how you attempted each problem, any figures that were generated, and scans of any work on paper that you wish to include. It is important that you include enough detail that we know how you solved the problem, since otherwise we will be unable to grade it.

A Jupyter notebook is provided, and we recommend that you use this notebook to write your report. It will help you not only ensure all of the code for the solutions is included, but also provide an easy way to export your results to a PDF file [1]. I recommend liberal use of Markdown cells to create headers for each problem and sub-problem, explaining your implementation/answers, and including any mathematical equations. For parts of the homework you do on paper, scan it in such that it is legible (there are a number of free Android/iOS scanning apps, if you do not have access to a scanner), and include it as an image in the iPython notebook[2]. If you have any questions/concerns about using iPython, ask us on Ed Discussion. If you decide not to use iPython notebooks, but go with Microsoft Word or Latex to create your PDF file, you have to make sure all of the answers can be generated from the code snippets included in the document.

Summary so far: (1) submit a single, standalone PDF report, with all code snippets to Gradescope; (2) submit a single zip file including source codes to Canvas. (3) I recommed Jupyter notebooks.

Points: This homework adds up to a total of **110 points**, as follows:

Problem 0: Get Connected	5 points
Problem 1: Python & Data Exploration	20 points
Problem 2: kNN Predictions	35 points
Problem 3: Naïve Bayes Classifiers	45 points
Statement of Collaboration	5 points

Problem 0: Get Connected (5 points)

Please visit our class forum on Ed Discussion: `https://edstem.org/us/courses/31466/discussion/`. Ed Discussion will be the place to post your questions and discussions, rather than by email to me or the TAs, since chances are that other students have the same or similar questions, and will be helped by seeing the discussion. Remember, your Ed Discussion participation will be taken into account for the participation grade as well. You do not need to mention anything regarding this in the report, we will be able to check whether you have visited Ed Discussion or not.

Problem 1: Python & Data Exploration (20 points)

In this problem, we will explore some basic statistics and visualizations of an example data set. First, download the zip file for Homework 1, which contains some course code (the `mltools` directory) and the "Fisher iris" data set, and load the latter into Python:

[1] For example, by doing a `Print Preview` in Chrome and `printing` it to a PDF.

[2] Tips from Gradescope: `http://gradescope-static-assets.s3-us-west-2.amazonaws.com/help/submitting_hw_guide.pdf`

```python
import numpy as np
import matplotlib.pyplot as plt

iris = np.genfromtxt("data/iris.txt",delimiter=None)  # load the text file
Y = iris[:,-1]              # target value is the last column
X = iris[:,0:-1]            # features are the other columns
```

The Iris data consist of four real-valued features used to predict which of three types of iris flower was measured (a three-class classification problem).

1. Use `X.shape` to get the number of features and the data points. Report both numbers, mentioning which number is which. `(5 points)`
2. For each feature, plot a histogram (`plt.hist`) of the data values `(5 points)`
3. Compute the mean & standard deviation of the data points for each feature (`np.mean`, `np.std`) `(5 points)`
4. For each pair of features (1,2), (1,3), and (1,4), plot a scatterplot (see `plt.plot` or `plt.scatter`) of the feature values, colored according to their target value (class). (For example, plot all data points with $y = 0$ as blue, $y = 1$ as green, etc.) `(5 points)`

Problem 2: KNN predictions (35 points)

In this problem, we will investigate the image classification problem using the CIFAR-10 dataset.

You can either download the dataset following the instructions inside the provided notebook, or directly from the website `https://www.cs.toronto.edu/~kriz/cifar.html`.

Fill in the corresponding parts in the provided jupyter notebook.

Problem 3: Naïve Bayes Classifiers (45 points)

In order to reduce my email load, I decide to implement a machine learning algorithm to decide whether or not I should read an email, or simply file it away instead. To train my model, I obtain the following data set of binary-valued features about each email, including whether I know the author or not, whether the email is long or short, and whether it has any of several key words, along with my final decision about whether to read it ($y = +1$ for "read", $y = -1$ for "discard").

x_1 know author?	x_2 is long?	x_3 has 'research'	x_4 has 'grade'	x_5 has 'lottery'	y $\Rightarrow$ read?
0	0	1	1	0	-1
1	1	0	1	0	-1
0	1	1	1	1	-1
1	1	1	1	0	-1
0	1	0	0	0	-1
1	0	1	1	1	1
0	0	1	0	0	1
1	0	0	0	0	1
1	0	1	1	0	1
1	1	1	1	1	-1

In the case of any ties, we will prefer to predict class +1.

I decide to try a naïve Bayes classifier to make my decisions and compute my uncertainty.

1. Compute all the probabilities necessary for a naïve Bayes classifier, i.e., the class probability $p(y)$ and all the individual feature probabilities $p(x_i|y)$, for each class y and feature x_i `(10 points)`
2. Which class would be predicted for $\underline{x} = (0\ 0\ 0\ 0\ 0)$? What about for $\underline{x} = (1\ 1\ 0\ 1\ 0)$? `(10 points)`

3. Compute the posterior probability that $y = +1$ given the observation $\underline{x} = (1\ 1\ 0\ 1\ 0)$. `(5 points)`

4. Why should we probably not use a "joint" Bayes classifier (using the joint probability of the features x, as opposed to a naïve Bayes classifier) for these data? `(10 points)`

5. Suppose that, before we make our predictions, we lose access to my address book, so that we cannot tell whether the email author is known. Should we re-train the model, and if so, how? (e.g.: how does the model, and its parameters, change in this new situation?) Hint: what will the naïve Bayes model over only features $x_2 \ldots x_5$ look like, and what will its parameters be? `(10 points)`

Statement of Collaboration (5 points)

It is **mandatory** to include a `Statement of Collaboration` in each submission, with respect to the guidelines below. Include the names of everyone involved in the discussions (especially in-person ones), and what was discussed.

All students are required to follow the academic honesty guidelines posted on the course website. For programming assignments, in particular, I encourage the students to organize (perhaps using Ed Discussion) to discuss the task descriptions, requirements, bugs in my code, and the relevant technical content `before` they start working on it. However, you should not discuss the specific solutions, and, as a guiding principle, you are not allowed to take anything written or drawn away from these discussions (i.e. no photographs of the blackboard, written notes, referring to Ed Discussion, etc.). Especially `after` you have started working on the assignment, try to restrict the discussion to Ed Discussion as much as possible, so that there is no doubt as to the extent of your collaboration.