

CS 178: Machine Learning & Data Mining: Fall 2021

Homework 1

Due Date: **Monday, October 4, 2021**

This homework (and many subsequent ones) involves data analysis, and discussion of methods and results, using Python. You must submit [a single PDF file](#) that contains all answers, including any text needed to describe your results, the complete code snippets used to answer each problem, any figures that were generated, and scans of any (clearly readable) work on paper that you want the graders to consider. It is important that you include enough detail that we know how you solved each problem, since otherwise we will be unable to grade it.

We recommend that you use Jupyter/iPython notebooks 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 We 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. If you have any questions about using iPython, ask us on Piazza. If you decide not to use iPython notebooks, and instead create your PDF file with Word or LaTeX, make sure [all](#) of the answers can be generated from the code snippets included in the document.

TL;DR: (1) submit a single, standalone PDF report, with all code; (2) I recommend iPython notebooks.

Problem 0: Get Connected (0 points, but it will make the course easier!)

Please visit our class forum on Ed Discussion: <https://edstem.org/us/courses/14090>. EdDiscuss will be the place to post your questions and discussions, rather than by email to the instructor or TA. Often, other students have the same or similar questions, and will be helped by seeing the online discussion.

Problem 1: Python & Data Exploration (20 points)

In this problem, we will compute some basic statistics and create visualizations of an example data set. First, download the zip file for Homework 1, which contains some course code (the `mltools` directory) and a dataset of New York area real estate sales, “nyc_housing”. Load the data into Python:

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3
4 nych = np.genfromtxt("data/nyc_housing.txt", delimiter=None) # load the text file
5 Y = nych[:, -1] # target value (NYC borough) is the last column
6 X = nych[:, 0:-1] # features are the other columns
```

These data are from the “NYC Open Data” initiative, and consist of three real-valued features and a class value Y representing in which of three boroughs the house or apartment was located (Manhattan, the Bronx, or Staten Island).

1. Use `X.shape` to get the number of features and the number of 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 (2,3), 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, and $y = 2$ as red.) (5 points)

¹For example, by doing a [Print Preview](#) in Chrome and [printing](#) it to a PDF. Please also remember to check the resulting PDF before submitting.

Problem 2: k-nearest-neighbor predictions (25 points)

In this problem, you will continue to use the NYC Housing data and create a k-nearest-neighbor (kNN) classifier using the provided `knnClassify` python class. While completing this problem, please explore the implementation to become familiar with how it works.

First, we will shuffle and split the data into training and validation subsets:

```

1  nych = np.genfromtxt("data/nyc_housing.txt", delimiter=None) # load the data
2  Y = nych[:, -1]
3  X = nych[:, 0:-1]
4  # Note: indexing with ":" indicates all values (in this case, all rows);
5  # indexing with a value ("0", "1", "-1", etc.) extracts only that value (here, columns);
6  # indexing rows/columns with a range ("1:-1") extracts any row/column in that range.
7
8  import mltools as ml
9  # We'll use some data manipulation routines in the provided class code
10 # Make sure the "mltools" directory is in a directory on your Python path, e.g.,
11 # export PYTHONPATH=${PYTHONPATH}:/path/to/parent/dir
12 # or add it to your path inside Python:
13 # import sys
14 # sys.path.append('/path/to/parent/dir/');
15
16 np.random.seed(0) # set the random number seed
17 X,Y = ml.shuffleData(X,Y); # shuffle data randomly
18 # (This is a good idea in case your data are ordered in some systematic way.)
19
20 Xtr,Xva,Ytr,Yva = ml.splitData(X,Y, 0.75); # split data into 75/25 train/validation

```

Make sure to set the random number seed to 0 before calling `shuffleData` as in the example above (and in general, for every assignment). This ensures consistent behavior each time the code is run.

Learner Objects Our learners (the parameterized functions that do the prediction) will be defined as python objects, derived from either an abstract classifier or abstract regressor class. The abstract base classes have a few useful functions, such as computing error rates or other measures of quality. More importantly, the learners will all follow a generic behavioral pattern, allowing us to train the function on one data set (i.e., set the parameters of the model to perform well on those data), and then make predictions on another data set.

You can now build and [train](#) a kNN classifier on `Xtr,Ytr` and make predictions on some data `Xva` with it:

```

1  knn = ml.knn.knnClassify() # create the object and train it
2  knn.train(Xtr, Ytr, K)     # where K is an integer, e.g. 1 for nearest neighbor prediction
3  YvaHat = knn.predict(Xva)  # get estimates of y for each data point in Xva
4
5  # Alternatively, the constructor provides a shortcut to "train":
6  knn = ml.knn.knnClassify( Xtr, Ytr, K );
7  YvaHat = predict( knn, Xva );

```

If your data are 2D, you can visualize the data set and a classifier's decision regions using the function

```

1  ml.plotClassify2D( knn, Xtr, Ytr ); # make 2D classification plot with data (Xtr,Ytr)

```

This function plots the training data and colored points as per their labels, then calls `knn`'s `predict` function on a densely spaced grid of points in the 2D space, and uses this to produce the background color. Calling the function with `knn=None` will plot only the data.

1. Modify the code listed above to use only the first two features of X (e.g., let X be only the first two columns of `nych`, instead of the first three), and visualize (plot) the classification boundary for varying values of $K = [1, 5, 10, 50]$ using `plotClassify2D`. (10 points)
2. Again using only the first two features, compute the error rate (number of misclassifications) on both the training and validation data as a function of $K = [1, 2, 5, 10, 50, 100, 200]$. You can do this most easily with a for-loop:

```

1 K=[1,2,5,10,50,100,200];
2 errTrain = [None]*len(K)          # (preallocate storage for training error)
3 for i,k in enumerate(K):
4     learner = ml.knn.knnClassify(...) # TODO: complete code to train model
5     Yhat = learner.predict(...)      # TODO: predict results on training data
6     errTrain[i] = ...                # TODO: count what fraction of predictions are wrong
7     #TODO: repeat prediction / error evaluation for validation data
8
9 plt.semilogx(...)                  #TODO: average and plot results on semi-log scale

```

Plot the resulting error rate functions using a semi-log plot (`semilogx`), with training error in red and validation error in green. Based on these plots, what value of K would you recommend? (10 points)

3. Create the same error rate plots as the previous part, but with all the features in the dataset. Are the plots very different? Is your recommendation for the best K different? (5 points)

Problem 3: Naïve Bayes Classifiers (35 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	x_2	x_3	x_4	x_5	y
know author?	is long?	has ‘research’	has ‘grade’	has ‘lottery’	\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

I decide to try a naïve Bayes classifier to make my decisions and compute my uncertainty. In the case of any ties where both classes have equal probability, we will prefer to predict class $+1$.

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 . (7 points)
2. Which class would be predicted for $\underline{x} = (0\ 0\ 0\ 0\ 0)$? What about for $\underline{x} = (1\ 1\ 0\ 1\ 0)$? (7 points)
3. Compute the posterior probability that $y = +1$ given the observation $\underline{x} = (0\ 0\ 0\ 0\ 0)$. Also compute the posterior probability that $y = +1$ given the observation $\underline{x} = (1\ 1\ 0\ 1\ 0)$. (7 points)
4. Why should we probably not use a “joint” Bayes classifier (using the joint probability of the features x , as opposed to the conditional independencies assumed by naïve Bayes) for these data? (7 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. Do we need to re-train the model to classify based solely on the other four features? If so, how? If not, what changes about how our trained parameters are used? Hint: what parameters do I need for a naïve Bayes model over only features x_2, \dots, x_5 ? Do I need to re-calculate any new parameter values in our new setting? What, if anything, changes about the parameters or the way they are used? (7 points)

Problem 4: Gaussian Bayes Classifiers (15 points)

Now, using the NYC Housing data, we will explore a classifier based on Bayes rule. Again, we'll use only the first two features of NYC Housing, shuffled and split in to training and validation sets as before.

1. Splitting your training data by class, compute the empirical mean vector and covariance matrix of the data in each class. (You can use `mean` and `cov` for this.) (5 points)
2. Plot a scatterplot of the data, coloring each data point by its class, and use `plotGauss2D` to plot contours on your scatterplot for each class, i.e., plot a Gaussian contour for each class using its empirical parameters, in the same color you used for those data points. 5 points
3. Visualize the classifier and its boundaries that result from applying Bayes rule, using

```
1 bc = ml.bayes.gaussClassify( Xtr, Ytr );  
2 ml.plotClassify2D(bc, Xtr, Ytr);
```

Also compute the empirical error rate (number of misclassified points) on the training and validation data. (5 points)

Problem 5: Statement of Collaboration (5 points)

It is **mandatory** to include a *Statement of Collaboration* in each submission, that follows 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 students to organize (perhaps using Piazza) to discuss the task descriptions, requirements, possible bugs in the support 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 (no photographs of the blackboard, written notes, referring to Piazza, etc.). Especially *after* you have started working on the assignment, try to restrict the discussion to Piazza as much as possible, so that there is no doubt as to the extent of your collaboration.