Homework 2 (due 2/9)

DS-210 @ Boston University

Spring 2022

Before you start...

Collaboration policy: You may verbally collaborate on required homework problems. However, you must write your solutions independently without showing them to other students. If you choose to collaborate on a problem, you are allowed to discuss it with at most 2 other students currently enrolled in the class.

The header of each assignment you submit must include the field "Collaborators:" with the names of the students with whom you have had discussions concerning your solutions. If you didn't collaborate with anyone, write "Collaborators: none." A failure to list collaborators may result in credit deduction.

You may use external resources such as software documentation, textbooks, lecture notes, and videos to supplement your general understanding of the course topics. You may use references such as books and online resources for well known facts. However, you must always cite the source.

You may **not** look up answers to a homework assignment in the published literature or on the web. You may **not** share written work with anyone else.

Submitting: Solutions should be submitted via Gradescope. The entry code is . Please submit a solution to this homework as a single IPython notebook (`.ipynb`) as much as this is possible.

Grading: Whenever we ask for a solution, you may receive partial credit if your solution is not sufficiently efficient or close to optimal. For instance, if we ask you to solve a specific problem that has a polynomial–time algorithm that is easy to implement, but the solution you provide is exponentially slower, you are likely to receive partial credit.

Late submission policy: No extensions, except for extraordinary circumstances. We accept submissions submitted up to one day late, but we may deduct 10% of points.

Questions

1. First read the following tutorials on NumPy:

   - https://numpy.org/devdocs/user/absolute_beginners.html
   - https://numpy.org/devdocs/user/quickstart.html

   Then:

(a) Create a one–dimensional array A that consists of 35 independent random real numbers distributed uniformly in the range $[0, 1]$.

(b) Output all entries that are at least 0.6.

(c) Square all entries of the array.

(d) Reshape it to have 5 rows and 7 columns. (We follow the convention that the first index is the row number and the second is the column number.) Let us refer to this new array as B.

(e) Output the median of each row in B.

(f) Output the mean of each column in B.

(g) Set the second column of B to be all 0.

(h) Set rows from the second to fourth to be all filled with 1.

(i) Did this modify the original array A?

(j) What is the difference between `numpy.resize` and `numpy.reshape`?

2. In this question you will work with the same data set as in Homework 1. We want to explore how useful k–means clustering (to be covered in the Friday lecture) could be for predicting the same values or labels as last time.

   (a) Select a subset of attributes that you think may be relevant to the value being predicted (similarly to last time). It suffices to consider one subset of attributes. You don't have to try subsets of different sizes (but if you want you can).

   (b) Split your data into the training and test set in the same way as before.

   (c) As opposed to decision trees, it is now important to decide how much weight is assigned to each attribute. For instance, you can rescale each of them so that the variance of each attribute in the training set is similar. Whatever you do, make sure that no single attribute dominates all other attributes and is the sole basis for clustering.

   (d) Compute a k–means clustering of points in the training set for different values of k. (For instance, $k = 1 \ldots 20$. Select a range that makes sense for your data set.)

   (e) For each considered k, explore how helpful the clustering you get is for the classification or regression question you considered last time. For each point in the test set, see to which cluster it gets assigned and make the prediction for this point be

   - the mean of values for points in the training set assigned to the same cluster, *if you are doing regression*,
   - the most popular label for points in the training set assigned to the same cluster, *if you are doing classification*.

   Now use the same measure as last time to see how accurate your resulting predictions are. Plot a graph that displays the relationship between accuracy and k, the number of clusters.

   (f) How does the accuracy of predictions you obtain this time compare to the accuracy of decision tree based predictions last time?

   (g) Did you gain any new insights into the data set? Did the results surprise you in any way? Any other interesting thoughts?

3. (Optional, no credit) How much time did you spend on this homework? The answer will have no impact on the credit you receive, but it may help us adjust the difficulty of future homework assignments.