San José State University
Computer Science Department
CS156, Introduction to Artificial Intelligence, Spring 2021

Homework #2

Objective:

This homework's objective is to implement KNN classifier from scratch.

Details:

For this assignment implement a KNN (k nearest neighbors) classifier from scratch without using the scikitlearn library. This means, do not use the following import line in your submitted solution:

f*rom sklearn.neighbors import KNeighborsClassifier*

Implement a function called *knn()*, which accepts three input parameters: a single observation, a reference dataset containing more than one observations, and k parameter defining how many nearest neighbors should vote on the classification:

knn(newObservation, referenceData, k=3)

The reference observation data should be a pandas dataframe, with rows as observations and columns as input variables. The last column of this dataframe should be the output/class variable. The parameter k should default to 3 (voting is based on 3 nearest neighbors). Your *knn()* function should not assume any dimensionality of the data and should be able to perform classification on any dimensionality of the data as long as the *newObservation* and the *referenceData* are of the same dimensions. The *newObservation* variable refers to a new observation for which we want to predict a class label.

Use Euclidean distance to compute the distance between the samples. Remember that most similar/nearest observations are going to have the smallest distance.

You can use the code from my notebook examples as a reference to help you get started:
- knn.synthetic_data.ipynb
- kmeans.synthetic_data.ipynb

<u>Your submission should include the following:</u>
1. Working implementation of the described above *knn()* function
2. Test your function on two datasets:

1) 2-D data
    a) Generate 100 synthetic samples from the following distributions:
        a. Class 0:
            (a) Dim1: Gaussian distribution with mean = -2 and sd = 2
            (b) Dim2: Gaussian distribution with mean = 0 and sd = 1
        b. Class 1:
            (a) Dim1: Gaussian distribution with mean = 2 and sd = 2
            (b) Dim2: Gaussian distribution with mean = 0 and sd = 1
    b) Separate this synthetic dataset into training and test sets at 80/20 ratio
    c) Use the training dataset as a reference data for your *knn()* function
    d) Predict each datapoint in the test set
    e) Print out overall accuracy rate in your test data predictions
    f) Plot two scatter plots, one that shows real labels and one that shows predicted labels. Make sure to plot both the training and test points and distinguish them by point markers. Distinguish the two classes by different colors.
2) 3-D data
    a) Generate 1000 synthetic samples from the following distributions:
        a. Class 0:
            (a) Dim1: Gaussian distribution with mean = 0 and sd = 3
            (b) Dim2: Gaussian distribution with mean = -3 and sd = 1
            (c) Dim3: Gaussian distribution with mean = -1 and sd = 1
        b. Class 1:
            (a) Dim1: Gaussian distribution with mean = 0 and sd = 3
            (b) Dim2: Gaussian distribution with mean = 1 and sd = 2
            (c) Dim3: Gaussian distribution with mean = 1 and sd = 1
        c. Class 2:
            (a) Dim1: Gaussian distribution with mean = 0 and sd = 3
            (b) Dim2: Gaussian distribution with mean = 3 and sd = 1
            (c) Dim3: Gaussian distribution with mean = 4 and sd = 1
        d. Class 3:
            (a) Dim1: Gaussian distribution with mean = 0 and sd = 3
            (b) Dim2: Gaussian distribution with mean = 5 and sd = 3
            (c) Dim3: Gaussian distribution with mean = -3 and sd = 1
    b) Separate this synthetic dataset into training and test sets at 80/20 ratio
    c) Use the training dataset as a reference data for your *knn()* function
    d) Predict each datapoint in the test set
    e) Print out overall accuracy rate in your test data predictions

Submission:

Email your assignment submission to me at Yulia.Newton@sjsu.edu and the grader (Akshay Kajale) at akshay.kajale@sjsu.edu. Make sure to email this submission by 11:59pm on the due date listed in Canvas. Your sent email is the proof of submission. The subject of the email should

say “CS156 Assignment 2”. In the body of the email list your name as it appears on the class roster and your student ID. Attach to this email both the pdf of your Jupyter notebook, which contains the solution for this homework assignment, as well as the notebook itself (the notebook file with .ipynb extension). Make sure to submit both files, otherwise the submission will not be considered complete.

Grading:

I will return the grades as fast as we can grade this homework. Normally it should not take more than a few weeks.

A total of 10 points are possible for this homework assignment.