

INF632 Homework 3

Statistical Learning & Prediction

Assigned: February 19.

Due: March 19 at 11:59pm. Submissions turned in after this time and still within two weeks of this deadline will be automatically docked 50% of the possible points.

Submission: Share a direct link to your repo / folder on canvas.nau.edu by the deadline.

Points: EE499 students, this is worth 15% of your final grade. EE599 students, this is worth 10% of your final grade.

Background:

Analysis of wearable device data isn't just about making predictions, it's also about understanding what the data reflects of real world data and how it changes with time.

Assignment:

Your assignment will be to write several functions to analyze how data is clustered, changes, and how one data source can be used to predict the observations of the other. You may use Octave/MATLAB, Python, R, or any of these in Notebooks, but you must write the functions to compute and perform the statistics outlined below from scratch. You shouldn't need any libraries beyond basics for reading CSVs.

Functions

K Means

kmeans for some arbitrary number of clusters. Your function should accept two variables, a data source X and number of clusters to look for k. X should be formatted as an arbitrary number of columns (dimensions of measure) and some arbitrary number of rows (records). The function should compute a single pass clustering analysis and return the clusters, with centroids calculated and point locations (records) for each cluster.

K Nearest Neighbors

knn for an arbitrary odd number of k. Your function should accept at training set, k, and a single unclassified new data point (must have the same number of columns as your training set). This training set should be structured the same way as X above.

Change Point Analysis

cpa (change point analysis) for a single variable time series data set. This function should calculate and return the first (dominant) change point from a vector data set of observations. You can assume that the time step between each data point is equal.

Application of your Functions

Using the functions you have developed from above, and the data provided, compute the following and answer the questions posed.

0.0.1 K Means and KNN

Using the sample data we have been using in class, demonstrate that both your kmeans() and knn() perform as expected.

0.0.2 CPA

Using the dailySteps data from HW2, with the zero step days removed, identify all (up to 8) change points in the two-year daily steps. Do any of these change points make sense, relative to the time of year?

Expectations and Grading Rubric:

Great news, you get another break from IEEE format submissions on this homework! This time I need to see your code, with comments that make it clear what you are doing. If you choose to complete this assignment in a Notebook (e.g. Jupyter Notebook), you can show your answers to the “Application of your Functions” in markdown cells after each step. If you choose not to use a Notebook, prepare a markdown file (.md) for each or your responses to the “Application of your Functions” questions.

Your submission will be graded as follows:

Area	Frac. of Points	UG: Meets Expectations	UG:Exceeds, G:Meets Expectations	G:Exceeds Expectations
C	Code Requirements	40% The code takes inputs and provides outputs as specified	The code also accepts optional input and makes use of defaults.	The code can self identify characteristics that change how the calculations will be performed.
	Code Clarity	40% Minimal comments, but sufficient to understand the basics of what is happening.	The code make good use of comments, including outlining some assumptions and where errors could happen (or how they are being handled).	The comments make reading the code so easy that one doesn't even have to know the programming language to understand the steps.
	Analysis and Response	20% The analysis is performed correctly and the findings are accurate.	The analysis and findings are correct, and the explanation shows an understanding of what was done and why.	The analysis and findings are correct, the explanation shows depth of thought, and further analysis is suggested in detail.