Ira A. Fulton Schools of Engineering
Arizona State University

CSE 572: Data Mining
Fall 2022 (Kerner)

Homework 4: Novelty detection

In Lab 17, we used statistical and distance-based approaches to detect anomalous changes in the daily closing prices of various stocks. The input data `stocks.csv` contains the historical closing prices of stocks for 3 large corporations (Microsoft, Ford Motor Company, and Bank of America). In the lab, we used anomaly detection techniques to detect anomalies in the changes in daily closing prices over the entire dataset (entire time period).

In this homework, you will re-frame this problem to instead use techniques for *novelty* detection. Instead of scoring each sample based on its anomalousness compared to all other samples, you will score every sample based on its anomalousness compared to all *previous* samples in time. You will step through each record in order of time and at each step construct an updated model that will be used to score the new sample. Use the kth nearest neighbor approach used in Lab 17, but instead of using the distance to the 4th nearest neighbor as in Lab 17, use the *average* distance to the four nearest neighbors.

Using the provided Colab notebook as a starting point, you will:

- Compute the novelty score for each date in the dataset
- Plot the novelty scores over time
- Identify which dates had the 5 highest novelty scores

Additional notes:

- You will not be able to compute the novelty score for the first date since there are no prior dates to use for fitting the model. Use 0 for the first date's score.
- For the kth nearest neighbors approach, you must have at least as many samples as neighbors to compute. For the first few dates, the number of samples used to fit the model (n_samples) will be smaller than the number of nearest neighbors (n_neighbors=4). When n_samples < n_neighbors, set n_neighbors = n_samples.

Submission

You will add your code to the notebook provided in the assignment instructions which contains starter code for loading the dataset (`cse572-homework4.ipynb`). Rename the notebook to `cse572-homework4-<lastname>.ipynb` and submit the following three deliverables:

1. a link to your Colab notebook (as a comment on the submission)
2. your .ipynb file (`cse572-homework4-<lastname>.ipynb`)
3. a pdf of the executed notebook (`cse572-homework4-<lastname>.pdf`)

Grading

Grading will be based on your code and the correctness of your outputs. You may receive partial credit if you come to the incorrect conclusion but parts of your work are correct, so make sure to show your work. Grading will be based on the following rubric.

- Notebook runs without errors (25 points)
- Correctly computed novelty scores (25 points)
- Correct plot of novelty scores over time (25 points)
- Correct output for/identification of dates with top 5 novelty scores (25 points)