Foundations of Data Science (CS F320)
Assignment 4

Submission by 11:59 PM on 15-Nov-2018

The Problem

Conejo and Tortuga are Spanish exchange students from Madrid who've come to BPHC to study Machine Learning.
Using their knowledge, they want to be able to help their city's Municipal Corporation get rid of pollution in the area.
The Professor suggested them to look into using KNNs for Time Series Analysis. However, they'd never heard of such an application before. Your job is to help Conejo and Tortuga model the pollution levels in Madrid. You will be doing this by implementing the K Nearest Neighbors routine.

Dataset

Download the following dataset from Kaggle:
`https://www.kaggle.com/decide-soluciones/air-quality-madrid`

You will make use of data from 2016 to predict the PM10 and PM25 values for the year 2017. Specifically, you will be using the remaining pollutant levels and the specific station to make your prediction.
(Station is a categorical variable. Handle Appropriately)
Missing values can be replaced with yearly average for each of the columns. Any other form of imputation you see fit can also be used.

For further description of the dataset, refer to the overview tab in the Kaggle Dataset page.

Tip: Don't just blindly use the imported dataset. The dates might not be aligned in the right way.

Time Series Analysis Using KNNs

When using KNNs, you have to combine supervised modelling with a Time-Series staple called the Lookback period.
Essentially, one just considers all the values that fall in the lookback period when fitting a KNN model on the explanatory variables over the time-period for predicting the next value.

When the next forecast is being done i.e the lookback window is modified, the actual value is added to make the new Lookback period, not the predicted value.
For this task, you will use a lookback period starting from 2016-01-01 00:00:00 and lasting 1 year, i.e till 2016-12-31 23:00:00. The very first prediction you make should be for 2017-01-01 00:00:00
For each successive prediction, the lookback period changes as mentioned below, hence the model has to be refit with new instances. So, make sure you're thorough with dataframe indexing.

Part A: Recursive Window

In the recursive approach, one makes use of an increasing window of points to re-estimate the model for the next forecast i.e the number of points on which the KNN is trained on increases by 1 after each forecast
Experiment with multiple values of k and see which one gives best fit

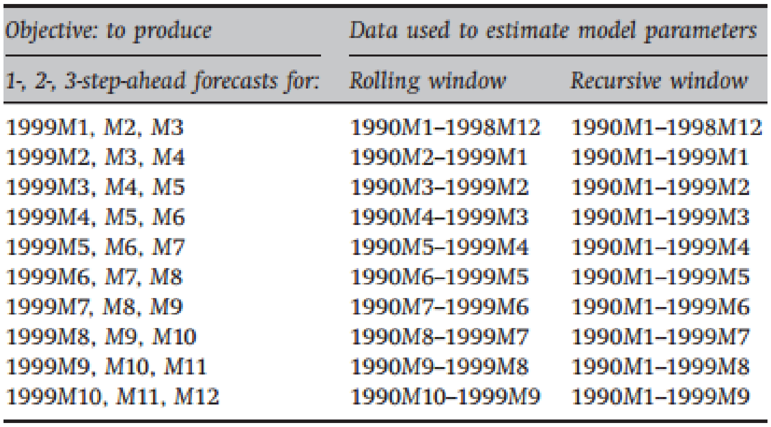
Part B: Rolling Window

In the rolling approach, one always makes use of a window of the same size. After each forecast, the window is shifted forward by 1 unit of time. This effectively means that 1 point leaves while another point is added to our model.
Experiment with multiple values of k and see which one gives best fit

Part C: Normalized Distance

Redo parts A and B by normalizing each vector to a unit vector. This makes it such that it is not biased towards features with large values.

An example



Objective: to produce	Data used to estimate model parameters	
1-, 2-, 3-step-ahead forecasts for:	Rolling window	Recursive window
1999M1, M2, M3	1990M1–1998M12	1990M1–1998M12
1999M2, M3, M4	1990M2–1999M1	1990M1–1999M1
1999M3, M4, M5	1990M3–1999M2	1990M1–1999M2
1999M4, M5, M6	1990M4–1999M3	1990M1–1999M3
1999M5, M6, M7	1990M5–1999M4	1990M1–1999M4
1999M6, M7, M8	1990M6–1999M5	1990M1–1999M5
1999M7, M8, M9	1990M7–1999M6	1990M1–1999M6
1999M8, M9, M10	1990M8–1999M7	1990M1–1999M7
1999M9, M10, M11	1990M9–1999M8	1990M1–1999M8
1999M10, M11, M12	1990M10–1999M9	1990M1–1999M9

Figure 1: Recursive Window vs Rolling Window, from Introductory Econometrics for Finance by Chris Brooks

Implementation Details

You will have to implement this in *python* using *pandas*, *numpy*, *scipy* and *matplotlib*. No other libraries are allowed. This means you must implement the K Nearest Neighbors routine yourself.

Submission

Submission will be via CMS. Gather all your code into a single .py file. Along with that, make a document showing your prediction vs the actual value for each day. Include any plots you find suitable. Zip the code and this document and name it with your ID numbers.

Helpful Links

`https://people.revoledu.com/kardi/tutorial/KNN/KNN_TimeSeries.htm`

`https://datascience.stackexchange.com/questions/12365/what-is-the-correct-way-to-apply-knn-to-a-time-series-using-a-rolling-window`

For Queries

Phani Shankar Ede (2015B3A70420H)
Anirudh Srinivasan (2015A7PS0382H)