**ALY6020 PREDICTIVE ANALYTICS**

**INSTRUCTOR NAME- DR. MARCO MONTE DE OCA**

REETIKA CHATURVEDI

Northeastern University



Note:

Reetika Chaturvedi, College of Professional Studies, Northeastern University, Boston, MA 02115.

Report based on Parameter space exploration of a K-Nearest Neighbors Algorithm on a Time Series Problem.

Contact: chaturvedi.re@northeastern.edu

**INTRODUCTION**

This assignment gave us an opportunity to predict nearest neighbors using exponential models. The machine learning technique used to find the nearest neighbor is called K- Nearest Neighbors (KNN) which is based on regression. Using this supervised technique, we were able to identify the arrangement of our dataset. Changes occurred over a period of time that can be can be seen in our dataset ‘Predictive Analytics”, downloaded from Google trends. We considered the data of United States.

Data changed (on a weekly basis) over the past five years. We were asked to predict the values of a week (Jan 10-16), wherein we have actual weekly value reported on January 17th as 24. Considering the past weeks for prediction we will try to predict the values for the entire week using Euclidean and Manhattan distance.

**ANALYSIS**

The model used for prediction is in the following form:

*xi+1=m(xi,xi-1,...,xi-n)*

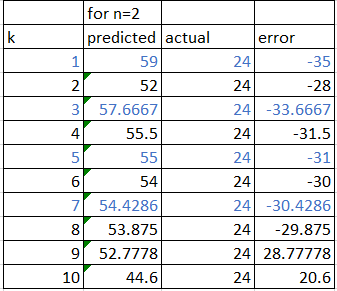
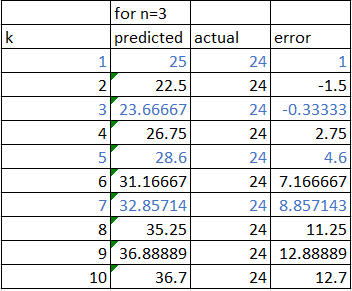
where n+1 is the dimensionality of the model, ‘m’ is the regression slope and xi+1 is the value that is predicted, xi-n is the lag value generated due to the given dimensionality.

Lags were generated in order to calculate the distance using Euclidean and Manhattan distance functions.

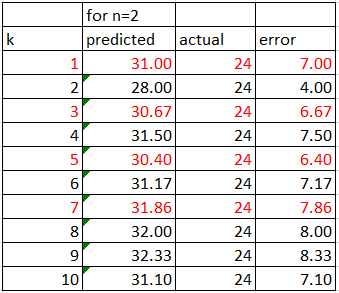
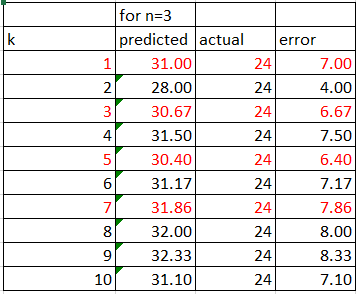
Distance was computed using lagged values and sorted from largest to smallest to analyze the closest predictions. The Euclidean distance was calculated using square root of the differences in distances whereas Manhattan distance was computed using the absolute values of the differences of distances.

Screenshots attached to represent the prediction:

1. **Euclidean Distance-** When the lag was taken as 3 the best prediction was when k=3 as the margin of error was the lowest (-0.33). On the other hand, when the lag was taken as 2, the best prediction was when k=7 as the margin of error was -30.4. It can be concluded from the predictions that as n increases, the maximum value of k might decrease because of the small size of the dataset. The point was given in the question and is observed during analysis.



1. **Manhattan Distance-** To calculate the distance in a grid like path, Manhattan distance was used which used absolute sum of differences. It can be seen that at different lag values it yielded the same error results in the dataset. It may be because of the dimensionality of the data. Since the data does not have many dimensions, Manhattan distance yielded similar results.



**CONCLUSION**

The following can be concluded:

1. The most favorable k value while computing Euclidean distance when(n)=2 was 7, whereas it decreased to 3 when (n)=3. Lower k value suggests that noise has a higher influence on the result. In this case it is due to the size of our dataset.
2. Manhattan distance is effective in high dimensional data. It is a grid like path that can be explained as the total sum of differences between the x and y coordinates. It is most favorable when k=5 as the margin of error was the lowest i.e., 6.40.
3. For this particular dataset, Euclidean distance makes more sense as the predicted values were more closer to the actual values.

**REFERENCES**

1. Martinez, F., Frias, M. P., Charte, F., & Rivera, A. J. (2019, May 31). Time Series Forecasting with KNN in R: the tsfknn Package. Retrieved from <https://cran.r-project.org/web/packages/tsfknn/vignettes/tsfknn.html>
2. TiagoTiago 42911 gold badge44 silver badges1010 bronze badges, Pablo SuauPablo Suau 1, Jacques KvamJacques Kvam 18111 silver badge44 bronze badges, Damian MelniczukDamian Melniczuk 55933 silver badges1919 bronze badges, Saurabh JainSaurabh Jain 12144 bronze badges, & Jaffar NaqshbandiJaffar Naqshbandi 1. (1966, June 01). When would one use Manhattan distance as opposed to Euclidean distance? Retrieved January 29, 2021, from <https://datascience.stackexchange.com/questions/20075/when-would-one-use-manhattan-distance-as-opposed-to-euclidean-distance>
3. forecast. (n.d.). Retrieved from <https://www.rdocumentation.org/packages/forecast/versions/8.12/topics/accuracy>