Assignment 1

Shivakanth Thudi

10/30/2016

Analysis

Part 1 - Imputer Strategy

For missing values, I used the mean strategy for imputations. When loading in the data, I replaced all flag values of "-9999" with NaNs. These NaNs were then imputed using global means across the columns specific to that flagged value.

For all missing values, the mean strategy for imputation was used.

Part 2 - Improving missing temperature readings

For missing temperature readings, it would be better to replace them with the previous temperature reading for the previous hour at that **same station** as the value, rather than using a global mean across all stations for that hour instead.

If this is not possible, then the temperature reading at that same hour for the previous day should be used to impute missing values.

This would help ensure that the model fits well to the training data, while also making good predictions.

Part 3 - Model Performance and MSEs

For the Boston dataset, the MSE was 10.960 when using linear regression. Using k-nearest neighbors regression, the lowest MSE was 18.063 with k, the number of neighbors, as 114. When considering a k-range of 1 to 50 however, the lowest MSE was 22.003, with k as 3.

Table 1: Boston data

Model	MSE	Number of Neighbors, k	Range considered for k
Linear Regression	10.960	Not applicable	Not applicable
k-Nearest Neighbors	18.063	114	1 to 456
k-Nearest Neighbors	22.003	3	1 to 50

Linear Regression fit the data better than k-nearest neighbors, and had a lower MSE on the test set.

First, we consider the testing set to consist of the stations USW00023234, USW00014918, USW00012919, USW00013743 and USW00025309. The following results are the aggregate MSE.

For the Climate dataset, the MSE was 6.957 when using linear regression. Using k-nearest neighbors regression and considering a range of 1 to 10 for k, the lowest MSE was 13.233 and the k was 10, when considering a range of 1-10.

Table 2: Climate data for Stations USW00023234, USW00014918, USW00012919, USW00013743 and USW00025309.

Model	MSE	Number of Neighbors, k	Range considered for k
Linear Regression	6.957	Not applicable	Not applicable
k-Nearest Neighbors	13.233	10	1 to 10

Station USW00023234

Table 3: Climate data for Station USW00023234

Model	MSE	Number of Neighbors, k	Range considered for k
Linear Regression	1.475	Not applicable	Not applicable
k-Nearest Neighbors	11.385	10	1 to 10

Station USW00014918

Table 4: Climate data for Station USW00014918

Model	MSE	Number of Neighbors, k	Range considered for k
Linear Regression	16.132	Not applicable	Not applicable
k-Nearest Neighbors	102.549	10	1 to 10

Station USW00012919

Table 5: Climate data for Stations USW00012919

Model	MSE	Number of Neighbors, k	Range considered for k
Linear Regression	3.706	Not applicable	Not applicable
k-Nearest Neighbors	28.556	10	1 to 10

Station USW00013743

Table 6: Climate data for Station USW00013743

Model	MSE	Number of Neighbors, k	Range considered for k
Linear Regression	8.843	Not applicable	Not applicable
k-Nearest Neighbors	11.265	10	1 to 10

Station USW00025309

Table 7: Climate data for Station USW00025309

Model	MSE	Number of Neighbors, k	Range considered for k
Linear Regression	4.507	Not applicable	Not applicable
k-Nearest Neighbors	17.845	10	1 to 10

Part 4 - Model performance and expectations

In all cases, linear regression achieved a better fit and a lower MSE on the test data as compared to the k-nearest neigbors algorithm. Also, while considering k-values for both datasets, the k-nearest neighbors algorithm had the lowest MSE on high k-values : k = 114 for the Boston data set, with high k on the Climate dataset as well (these values were not explicitly found because the computations would take a lot of time).

This aligns with my expectation that among k-NN procedures, the smaller k is, the better the performance. We observe that since the k-values were high here, the performance was not good. Here, linear regression outperforms k-Nearest Neighbors.

Part 5 - Improving model performance

I would improve the performance of these models by exploring whether more features needed to be added to the models, and by selecting the best subset of features using stepwise selection techniques. Alternatively, I would explore the use of other models and algorithms as well, apart from linear regression and k-NN, that might perform better with our data.