Temperature Prediction

Using KNN, Naïve Bayes, Linear Logistic Regression and Tree

by **Rajat Chakraborty**

1. **KNN, Naïve Bayes and Linear Regression**

Validation Error:

	KNN	Naive Bayes	Linear Regression
RMS Error	3.23	3.78	2.44
Median Abs Error	1.44	1.57	1.28

Test Error:

	KNN	Naive Bayes	Linear Regression
RMS Error	3.4	3.65	5.21
Median Abs Error	1.5	1.54	1.76

Most Important Features

13 feature coefficients have magnitude greater than 0.001.

Top 10 Features:

Feature Rank	Feature number	City	Day
1	334	Chicago	-1
2	347	Minneapolis	-1
3	405	Grand Rapids	-1
4	366	Kansas City	-1
5	361	Cleveland	-1
6	307	Omaha	-2
7	367	Indianapolis	-1
8	264	Minneapolis	-2
9	9	Boston	-5
10	236	Springfield	-3

Errors using only the 10 most important features:

Validation Error:

	KNN	Naive Bayes	Linear Regression
RMS Error	2.82	2.79	2.26
Median Abs Error	1.25	1.23	1.11

Test Error:

	KNN	Naive Bayes	Linear Regression
RMS Error	2.5	2.66	4.91
Median Abs Error	1.26	1.25	1.73

Improve Performance:

The following methods are used:

- Selection of important features using the Lasso method
- Scaling the features with zero mean and unit variance.
- Adding gaussian noise to the given training data and augmented that to the current training set to increase data size.
- Tuned the alpha value of the Ridge().

Validation RMSE: 2.235
Test RMSE: 2.048

2. Model Complexity with Tree Regressors

The following observations are made:

- Boosted tree has the lowest bias (based on training error).
- For regression tree, the minimum validation error is achieved when the max tree depth is 4 (solid red curve).
- Random Forest is least prone to overfitting as seen by the green solid line. The validation error does not go up with complexity unlike regression tree or boosted tree. RF uses a subset of the complete dataset to train a tree and then averages the decision

of the trees to get the final result which decreases variance. This process prevents overfitting.

- Boosted trees seems to perform better with smaller trees (based on validation error seen in the solid blue line (depth =2,4,8)). Large trees tend to overfit the training data as they try to fit the training data too closely with increasing complexity which might not be general for the validation/test data. On the other hand, small trees can capture the underlying pattern and have less variance. Combining prediction from many trees then decreases the bias.