Lab 3

Members: jinya425 (Jin Yan), siyli424 (Siyu Liu)

In this lab, we test three models from MLlib library, which are `GeneralizedLinearRegression` `RandomForestRegressor` and `GBTRegressor` (Gradient-boosted Tree).

To check whether our choice for the kernels' width is sensible, we find a station which contain temperature info in the day of '2013-07-04' near the target location.

The station info in the data is

```
# 74460;Jönköpings Flygplats;2.0;57.7514;14.0733;1962-01-01 00:00:00;2016-10-01 07:00:00;226.0
```

The whole results are below.

	Linear Regression	Random Forest	Gradient-boosted Tree	Sum Kernel	Multiply Kernel	Near Station
04:00:00	7.77	12.96	14.09	5.95	15.23	12.8
06:00:00	7.57	12.96	14.20	5.76	15.69	14.0
08:00:00	7.37	12.96	15.31	5.73	16.65	16.0
10:00:00	7.16	13.16	17.28	5.99	17.55	18.9
12:00:00	6.96	13.51	18.53	6.40	18.29	20.4
14:00:00	6.76	13.51	18.55	6.70	18.71	21.0
16:00:00	6.55	13.51	18.55	6.67	18.55	21.7
18:00:00	6.35	13.51	18.36	6.21	17.70	19.8
20:00:00	6.14	13.44	16.97	5.48	16.34	15.4
22:00:00	5.94	11.44	13.27	4.85	15.00	14.3
24:00:00	5.73	11.25	12.85	4.54	13.90	13.5

From the table, we can see multipy kernel and Gradient-boosted Tree has the best result comparing with the real temperature. And it shows that our kernels' width are sensible. Random Forest is better than Linear Regression and Sum Kernel. But we can see that some time periods have the same value because the random forest predictions are jagged. The performence of Linear Regression is bad since it is also sum the kernels, same as the Sum Kernel.

Multiply kernel is better than sum kernel. Since multiplication has a "scaling effect" which sum did not have. When we multiply two numbers, the result varies more dramatically depending on the values of the two numbers. This is more indicative of the fact that all three kernel distances are small in order to be similar.

For the definition of closeness, we use the same strategy for days and hours. These values should be seen as a ring, as the temperature changes are continuous and correlated on day and hour. For example the distance between 2012-12-31 and 2013-01-01 should be 1, not 365. the same applies to hours. About location distance, we use the default function.