732A54 Big Data Analytics

Lab 3- Machine Learning with Spark

Stylianos Sidiropoulos (stysi607) - Vasileia Kampouraki(vaska979)

Group 17

Note: Code was based on the code of groups A17 & A06 - Machine Learning course(732A99) Lab3 Topic3.

Also, the analysis of the theory behind this assignment is based on the teacher's slides (Jose M. Pena) from Lecture 3a Topic 1.

In this assignment we made temperature predictions for a place in Sweden for the date "2014-07-10" from 4 am to 24 pm in an interval of 2 hours using two kernel models.

The first one consists of the summation of three Gaussian kernels and the second one of the product of those three kernels.

The parameters h which are called smoothing factors where chosen in a sensible way for each kernel. More specifically we have chosen

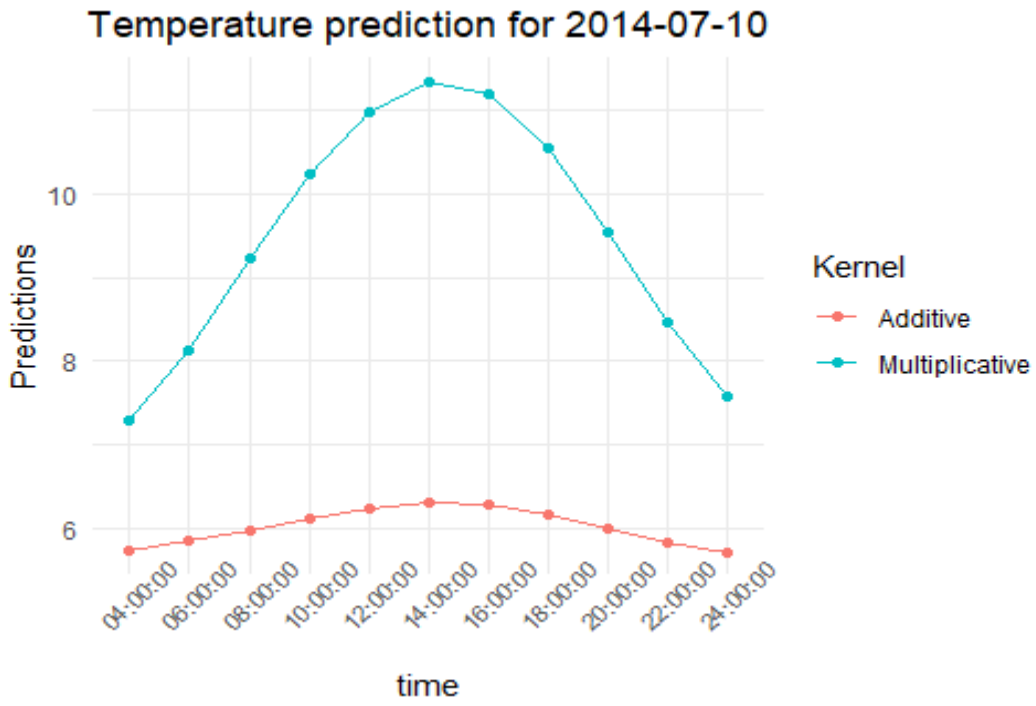
- h_distance = 500000m
- h_date = 100 days
- h_time = 5 hours

When we choose small values for parameter h we consider few data points, whereas when we choose big values we consider a lot. Parameter h inserts actually a bias-variance trade-off and a small value for h means large variance and low bias and the opposite for a big value of h.

That's the reason why we generally get ''noisy'' results for very small values of h and very ''flat'' results for big values of h. Thus, we need to choose some intermediate values for the h parameters of our kernels. The reason we chose those values for the three h smoothing factors is because there can be actually a difference in the temperature for a distance of 500km, a time interval of 100 days (e.g. from summer to autumn) and 5 hours(e.g. early morning to afternoon).

Below is the plot of the predicted temperatures produced by using both kernels.

(The plot was produced using RStudio and ggplot library)



We can see that the Multiplicative kernel model gave better predictions as they are more diverse compared to the ones obtained from the Additive kernel model which has bigger variance as we can see from the tails that are longer.

The reason for this might be because those two models assign ‘’weight’’ to the temperatures in a different way. So for example, if the three kernels have very small values, or even zero for a kernel, then the additive kernel will assign more ‘’weight’’ to the temperatures that are not correct whereas in such a case the multiplicative kernel would assign smaller or 0(if there was a 0 from one of the three kernels) ‘’weight’’ to those temperatures.