TDDE31 - Lab 3

Group C6
olosi122, erisn497

Task 1: Show that your choice for the kernels' widths is sensible

Our choices of kernel widths are 100, 10, and 2 for the distance kernel, day kernel, and hour kernel, respectively. The kernels have been implemented according to the instructions. The kernels have been simulated with realistic input.

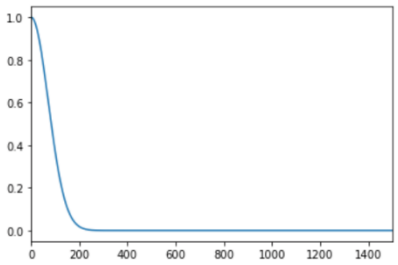

Figure 1 Distance kernel with h = 100

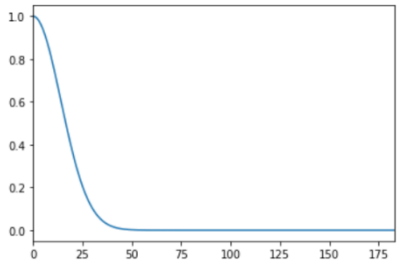

Figure 2 Day kernel with h = 10

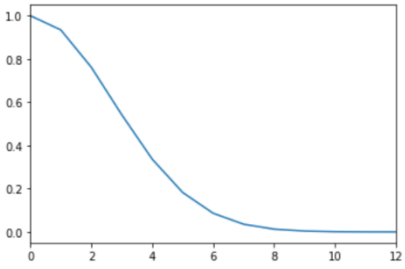

Figure 3 Hour kernel with h = 2

The distance kernel is defined as the physical distance between the data points and the point of prediction, measured in kilometers. The maximum distance between two cities is Sweden is about 1500 km. The kernel will give considerable weight to the data points within 200 km from the prediction point. The closer the distance, the more likely they are to have similar temperature readings. The reason 100 was chosen as the kernel width is that it allows many data points to be considered surrounding the prediction point. This is especially useful if there are no temperature readings close to the area of prediction. Multiplied with the day kernel, as high- and low-pressure systems have an impact on the temperature and travels across the country, surrounding readings leading up to the day of prediction will also be considered.

The day kernel is defined as the difference between the data points and the point of prediction in terms of what day of the year it is, measured in days. The maximum distance between two points is less than 185. Setting the kernel width to 10 gives considerable weight to data points 30 days away from the prediction point. This allows data to be considered leading up to the prediction point, which is especially important to account for seasonality and that, for example, spring and fall can have unusually high or low temperatures depending on the year. A longer tail would smooth the varied results on any given day.

The day kernel is defined as the difference in terms of the clock, measured in hours. The maximum distance between two points is 12 hours. With a kernel weight of 2, the kernel gives considerable weight up to 6 hours. The largest temperature change most commonly stems from the day-night cycle. In case of

quick change either in the morning or in the afternoon, the relatively long tail of the S-curve will hedge the prediction as to not be overly volatile.

Task 2: Compare the results from additive and multiplicative kernels

The point of prediction is:

- Date: 2013-07-04
- Latitude: 59.4113
- Longitude: 18.0578

The results from the additive kernel are the following:

```
('00:00:00', 4.775498573908729)
('22:00:00', 4.9647891390302075)
('20:00:00', 5.118545993018595)
('18:00:00', 5.1370053376509075)
('16:00:00', 5.0886966969163625)
('14:00:00', 4.898448774702406)
('12:00:00', 4.193690523994309)
('10:00:00', 3.8389778973857274)
('08:00:00', 3.296762729419556)
('06:00:00', 2.9521166151379696)
('04:00:00', 2.716238415608755)
```

The results from the multiplicative kernel are the following:

```
('00:00:00', 12.189220782846501)
('22:00:00', 13.543447848580493)
('20:00:00', 16.436599209889383)
('18:00:00', 17.72949149515628)
('16:00:00', 18.033571883827687)
('14:00:00', 18.108075197842698)
('12:00:00', 17.467259014876102)
('10:00:00', 16.36673006717434)
('08:00:00', 14.509204440031917)
('06:00:00', 12.99519909332644)
('04:00:00', 11.607346912735883)
```

The first thing to establish is that we are predicting the temperature in the Stockholm area, in the summer, around the time when some of the highest temperatures are measured. The additive kernel only reaches about 5℃, which is highly unlikely, based on personal experience. There is also very low variance over the hours of the day. The temperature also usually changes more than 1.5℃ as the sun comes up and goes down. This could indicate that the additive kernel has both a bias and lacking variance.

The multiplicative kernel reaches a temperature of about 18℃, which is more reasonable. There is also more variance, as the lowest temperature is measured arguable just around the time the sun is expected to rise – around 11.5℃.

The only difference between the kernels is in the addition or multiplication of the individual kernels. This causes some differing behavior.

The additive kernel adds the results from each individual kernel, a sum between 3 and 0. As long as at least one kernel has weight close to 1, the additive kernel will consider it in the prediction. Therefore, it will weigh a lot of data points. For the data points that are close with regard to all 3 kernels, the additive

kernel will be closer to 3. This is not considerably different from the large number of points that will score high on at least one of the kernels. Thus, this explains the bias and the low variance during the day.

The multiplicative kernel multiplies the result from each individual kernel, a product between 1 and 0. In this case, as long as one kernel is close to zero, then the absolute weighting of that data point will be relatively small. This results in giving data points a considerably higher weighting when all kernels have a low distance to the prediction point. Thus, the prediction will be based much more on the recent data, explaining the lower bias. As for the higher variance intra-day, compared to the additive kernel, the day kernel has a larger impact on the product as it drastically reduces it for hours of the day far away from the prediction point.

Task 3: Compare results with MLlib models (Decision tree, Random Forest)

The models chosen for the exercise are a decision tree regressor and a random forest regressor. The decision tree and the decision tree within the random forest have a maximum depth of 10. The number of trees in the random forest is 5.

The results for the decision tree are:

```
('24:00:00', 14.562022972542886)
('22:00:00', 14.562022972542886)
('20:00:00', 14.562022972542886)
('18:00:00', 18.121297683393045)
('16:00:00', 18.121297683393045)
('14:00:00', 18.121297683393045)
('12:00:00', 18.121297683393045)
('10:00:00', 18.121297683393045)
('08:00:00', 18.121297683393045)
('06:00:00', 14.519707549079131)
('04:00:00', 13.191813573112466)
```

The results for the random forest are:

```
('24:00:00', 8.416619266017708)
('22:00:00', 8.551832973010557)
('20:00:00', 9.05147305196098)
('18:00:00', 12.32937710349339)
('16:00:00', 13.18657798768874)
('14:00:00', 13.18657798768874)
('12:00:00', 13.18657798768874)
('10:00:00', 13.18657798768874)
('08:00:00', 12.894808284544222)
('06:00:00', 10.62253686267304)
('04:00:00', 7.806613240461869)
```

Beginning with the decision tree. The first observation is that the results are most similar to the multiplicative kernel. Many of the values are repeating, for example 18.12℃ between hours 08:00 and 18:00. This is because of the architecture of the decision tree. With a relatively low depth, similar prediction points intra-day will be grouped to the same result i.e., end up in the same child node.

The random forest is also most similar to the multiplicative kernel. A rough interpretation would be that the results are about 5℃ lower than the multiplicative kernel – arguably too low based on personal experience. The random forest makes predictions based on multiple decision trees, 5 in this case, and then makes the

prediction based on the average of the results from the decision trees. The result is biased but has similar variance to the multiplicative kernel. The bias could stem from some of the randomly generated trees. A good mitigation for this would be to increase the number of trees in the future.