Mathias Fredholm, matfr953, 961121-8158
David Gumpert Harryson, davha130, 960302-4937

Lab 3 - Machine Learning with Spark

INSTRUCTIONS. Hand in a lab report that includes the name and LiU-id for each group member. The report should include your code, results from the code execution, and written answers to the questions in the assignment. Comment each step in your program to provide a clear picture of your reasoning when solving the assignment.

ASSIGNMENT *Implement in Spark (PySpark) a kernel model to predict the hourly temperatures for a date and place in Sweden. To do so, you should use the files temperature-readings.csv and stations.csv from previous labs. Specifically, the forecast should consist of the predicted temperatures from 4 am to 24 pm in an interval of 2 hours for a date and place in Sweden.*

Use a kernel that is the sum of three Gaussian kernels:

- *The first to account for the distance from a station to the point of interest.*
- *The second to account for the distance between the day a temperature measurement was made and the day of interest.*
- *The third to account for the distance between the hour of the day a temperature measurement was made and the hour of interest. Choose an appropriate smoothing coefficient or width for each of the three kernels above. You do not need to use cross-validation.*

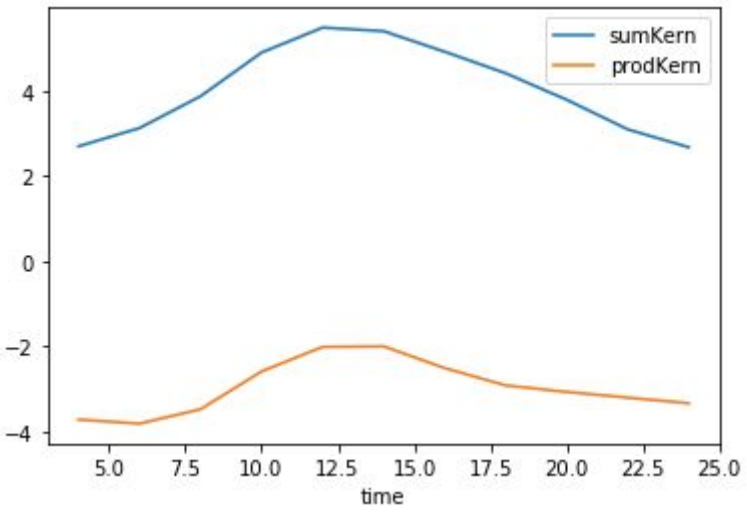
Our code:

```
h_distance = 100 # Up to you
h_date = 5 # Up to you
h_time = 2.5 # Up to you
a = 58.4274 # Up to you
b = 14.826 # Up to you
date = “2013-01-24” # Up to you
givenDate = toDateTime(date)
timeStamps = [24, 22, 20, 18, 16, 14, 12, 10, 8, 6, 4]
```

Output:

Mathias Fredholm, matfr953, 961121-8158
David Gumpert Harryson, davha130, 960302-4937

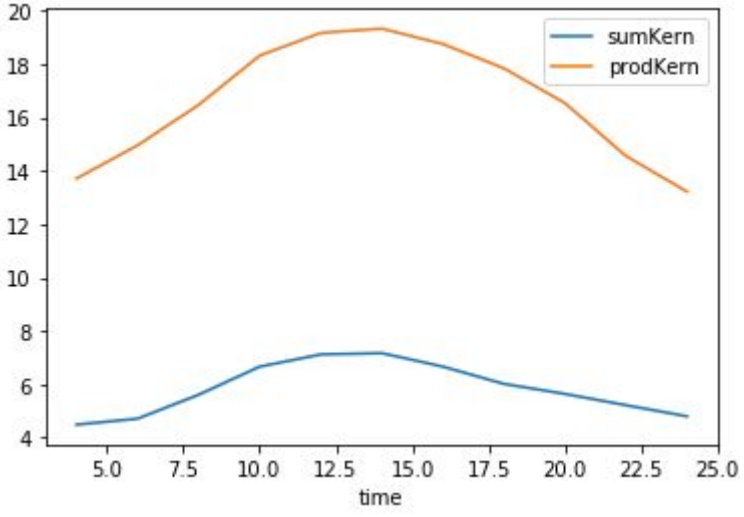
	time	sumKern	prodKern
0	24	2.673455	-3.328526
1	22	3.091985	-3.198374
2	20	3.786627	-3.063857
3	18	4.408017	-2.911114
4	16	4.913126	-2.505046
5	14	5.394297	-1.995264
6	12	5.479816	-2.007215
7	10	4.894223	-2.581827
8	8	3.870674	-3.468681
9	6	3.124059	-3.807008
10	4	2.695120	-3.712995




```
h_distance = 100 # Up to you
h_date = 5 # Up to you
h_time = 2.5 # Up to you
a = 58.4274 # Up to you
b = 14.826 # Up to you
date = "2013-07-04" # Up to you
givenDate = toDateTime(date)
timeStamps = [24, 22, 20, 18, 16, 14, 12, 10, 8, 6, 4]
```

Output:

	time	sumKern	prodKern
0	24	4.801511	13.229538
1	22	5.219054	14.563409
2	20	5.641548	16.551095
3	18	6.018928	17.849645
4	16	6.666353	18.759548
5	14	7.169995	19.337706
6	12	7.120047	19.175632
7	10	6.662847	18.325369
8	8	5.609774	16.477010
9	6	4.713006	14.960227
10	4	4.489839	13.718917



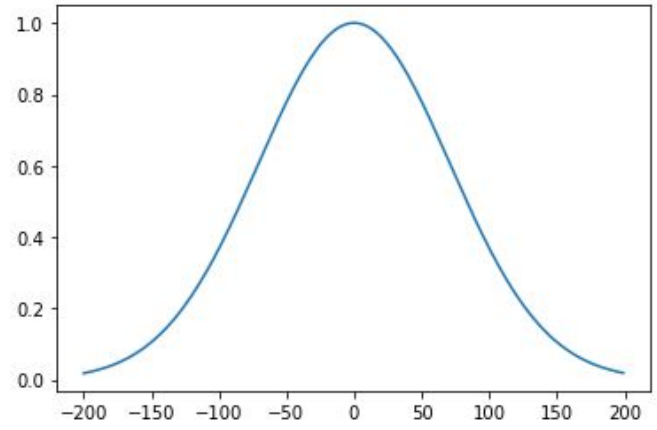
Mathias Fredholm, matfr953, 961121-8158
David Gumpert Harryson, davha130, 960302-4937

Answer to questions:

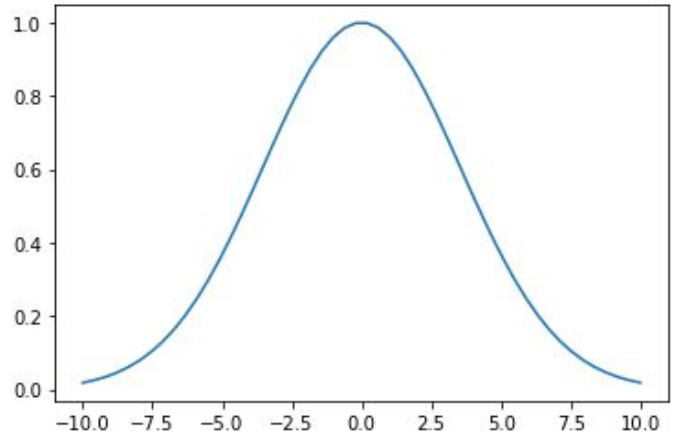
- Show that your choice for the kernels' width is sensible, i.e. it gives more weight to closer points. Discuss why your definition of closeness is reasonable.

As can be seen by the plotted gaussian distribution below, representing out distance kernel, distances above 200km will not have a large impact on the final prediction of the temperature. The same reasoning can be made for the other two kernel. For the date time kernel we can see that readings further away than 10 days (not taking to account the year) will have little impact and for the time kernel readings more than 10 hours from the prediction time will also have little impact. After tweaking and trying different h_values we feel this is a reasonable representation of the problem.

Distance Kernel:
h_distance = 100

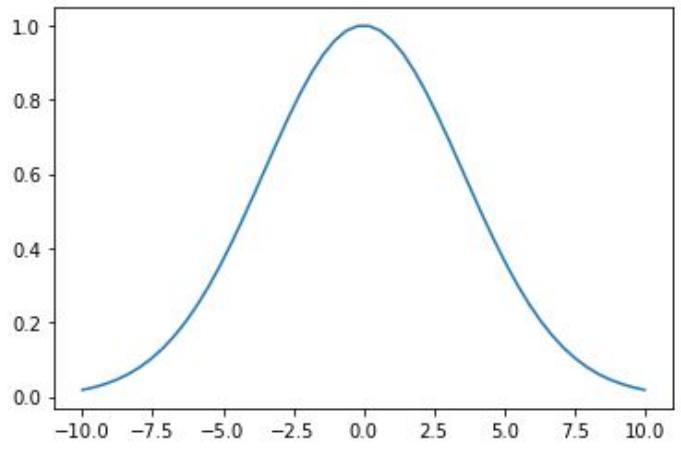


Date Kernel:
h_date = 5



Time kernel:
h_time = 2.5

Mathias Fredholm, matfr953, 961121-8158
David Gumpert Harryson, davha130, 960302-4937



- Repeat the exercise using a kernel that is the product of the three Gaussian kernels above. Compare the results with those obtained for the additive kernel. If they differ, explain why.

Using product rather than sum makes each of the three kernels more dependent on each other. If one of the variabels are far from the one we are trying to predict it will make all of the three have less influence on what we are trying to predict. For example, even though the time and place of the historical value we are using is exactly the same as the one we are trying to predict, if the historic date is from the winter when we are trying to predict a date in the summer, it makes the time and place rather irrelevant. With that reasoning, it makes sense that using the product of the three Gaussian kernels rather than the sum of the three Gaussian kernels.