# TDDE31 Lab3

David Albrekt daval064, Stian Lockhart Pedersen stilo759

**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.**
By using a smaller sample of the data in R we could find values for the kernels width that were sensible. This included a simple hyperparameter search which gave the resulting findings:

The distance values are the least contributing factor as the month/day/hourofday impacts the temperature more greatly. This is true because the Swedish temperatures vary more on the time of the day and the month rather than the distance (I.e., the distance between Norrköping and Linköping).

**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.**

For the multiplied kernels each of the kernels depend on each other. That said, when one of the kernels are far off, I.e., the distance kernel is way off but the time and date are spot on, the distance kernel will neglect the two others. In the sum kernel the contribution from the two kernels that are spot on will still contribute significantly.

**Repeat the exercise using at least two MLlib library models to predict the hourly temperatures for a date and place in Sweden. Compare the results with two Gaussian kernels. If they differ, explain why.**

We have tested 3 different regression models from MLlib

LogisticRegressionWithSGD:
[('24', 0.18616402892235281), ('22', 0.18616279950641182), ('20', 0.18616157009047085), ('18', 0.18616034067452986), ('16', 0.18615911125858889), ('14', 0.18615788184264789), ('12', 0.18615665242670693), ('10', 0.18615542301076593), ('08', 0.18615419359482493), ('06', 0.18615296417888397), ('04', 0.18615173476294297)]
LogisticRegressionWithSGD compared to the gaussian kernels used in part 1) differ quite substantially. That is because the data that we use in the logisticregression model is not representative and cannot capture the prediction of the temperature.

RidgeRegressionWithSGD:
[('24', 1.0696820830155594e+66), ('22', 1.069676039156565e+66), ('20', 1.0696699952975717e+66), ('18', 1.069663951438579e+66), ('16', 1.069657907579584e+66),

('14', 1.0696518637205902e+66), ('12', 1.0696458198615964e+66), ('10', 1.0696397760026025e+66), ('08', 1.0696337321436087e+66), ('06', 1.0696276882846148e+66), ('04', 1.069621644425621e+66)]
Tge results from RidgeRegressionWithSGD are even closer to zero. The problem persists that the values that we send into the model (year month and date) are neusence paramters for the model.

LassoWithSGD:
[('24', 1.0696820533542038e+66), ('22', 1.0696760094953776e+66), ('20', 1.0696699656365513e+66), ('18', 1.0696639217777251e+66), ('16', 1.0696578779188989e+66), ('14', 1.0696518340600727e+66), ('12', 1.0696457902012463e+66), ('10', 1.06963974634242e+66), ('08', 1.0696337024835938e+66), ('06', 1.0696276586247676e+66), ('04', 1.0696216147659414e+66)]

We even tested a third alternative MlLib model, but the same problems as described above persist even here