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 tocloser 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).

COMMENT-1

Predicated temperature using Gaussian models

date = "2013-01-01"

Sum

[
('24:00:00', -5.207852008009206),
('22:00:00', -5.2239740187650625),
('20:00:00', -5.24917935670155),
('18:00:00', -5.243884711556564),
('16:00:00', -5.278938012419783),
('14:00:00', -5.346101690843786),
('12:00:00', -5.564098809763531),
('10:00:00', -5.698776761532565),
('08:00:00', -5.817451562729036),
('06:00:00', -5.746884985820088),
('04:00:00', -5.810810152320789)
]

Multiply

[
('24:00:00', -4.941761042102876),
('22:00:00', -5.000779194604902),
('20:00:00', -5.070462845636481),
('18:00:00', -4.699608153303366),
('16:00:00', -4.703773844447404),
('14:00:00', -4.69256837935204),
('12:00:00', -5.008432809625627),
('10:00:00', -5.511424455252313),
('08:00:00', -5.827919229508379),
('06:00:00', -5.691842243775445),
('04:00:00', -5.836192665025088)
]

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.

Comparison between sum and mult (COMMENT 2)

For the multiplied kernels each of the kernels depend on each other. That said, when one of the kernels is 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 either way

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.
Date = "2013-01-01"

LogisticRegressionWithSGD:
[
('24', -0.35022191892201388),
('22', -0.35133643259247727),
('20', -0.35296334313920852),
('18', -0.3536741534998335),
('16', -0.35701260168690274),
('14', -0.36281004403080019),
('12', -0.37649625067086678),
('10', -0.38404220104497455),
('08', -0.39018878031414256),
('06', -0.3865308735956724),
('04', -0.39030712859345801)
]

Comparison between sum and LogisticRegressionWithSGD (COMMENT 2)
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. Although compared to the other mllib this model is substantially better.

Comparison between mult and LogisticRegressionWithSGD (COMMENT 2)
the same can be said for the comparison between multiplication gaussian kernel and LogisticRegressionWithSGD. The fluxuation between the different values between the days in percentage is more closely related to the multiplication gaussian kernel than the sum gaussian kernel.

RidgeRegressionWithSGD:
[
('24', 1.072523544106797e+66),
('22', 1.0679106806653234e+66),
('20', 1.0647353404273139e+66),
('18', 1.0715318144385685e+66),
('16', 1.0665653480827197e+66),
('14', 1.063994340894109e+66),
('12', 1.078350848215841e+66),
('10', 1.0714622322795078e+66),
('08', 1.0686211169240269e+66),
('06', 1.101684982919555e+66),
('04', 1.0874690020196999e+66)
]

The 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 nuisance parameters for the model.

(COMMENT 2)

Comparison between mult and RidgeRegressionWithSGD

The model RidgeRegressionWithSGD performs worse than the sum gaussian kernel, although it seems to just predict values close to zero. This we believe has to do with the values we actually send into the ridgeregression, which are not parameterized in any sensible way.

(COMMENT 2)

Comparison between mult and RidgeRegressionWithSGD

The same conclusion can be made for the comparison to the multiplication gaussian kernel, although the multiplication is much better in performance over the sum.

LassoWithSGD:
[
('24', 1.0725235143902545e+66),
('22', 1.0679106510646795e+66),
('20', 1.0647353109070218e+66),
('18', 1.071531784751008e+66),
('16', 1.0665653185199509e+66),
('14', 1.0639943113969048e+66),
('12', 1.0783508183630628e+66),
('10', 1.071462202599432 5e+66),
('08', 1.0686210873162323e+66),
('06', 1.1016849524908322e+66),
('04', 1.087468971945611e+66)
]

(COMMENT 2)

Comparison between sum mult and LassoWithSGD

The LassoWithSGD with predefined default parameters seems to perform the exact same way as the RidgeRegressionWithSGD. The same compariosons can be made towards the multiplication and sum kernels.

(COMMENT 2)

Comparison between MLLIB models

All the mllib models seem to predict values close to zero, even though less prevalent in the LogisticRegressionWithSGD. This we believe has to do with the model input parameters not being formatted or converted to reasonable values. For instance, we now only send in the year as an int, and for the models they can't interpret this in a sensible way. The more accurate way to do this is to copy the method for the sum/mult gaussian kernels where the time/date/year are transformed into a scale more accuratly representing their impact.