THESIS LOG

STEFAN SABEV

1. Week 3

In the model have:

- last 4 friday weights
- the twitter counts

Those would be the things that change.

(1) Implement the method in the Ryan Adams & David MacKay paper.
(2) Run this on some toy data.
(3) Run this on the search volumes without the Twitter data.
(4) Histogram of the change point to the nearest twitter peak.

Non-linear function of the features.
And SV regression with polynomial kernel.
Use relative error or use absolute error.
Average across days and destinations.
Do a histogram of median, box plot or any ways to drill down and show it clearly.

(1) take quantiles across destination.
(2) Median/mean absolute error.

Take mean over days or median over destinations. Cluster the destinations by popularity - split into 5 groups. report the mean absolute error for each of the groups.
(Add something to the prior that makes change point more likely if there is a twitter spike)

2. Week 5

What's been tried so far:

- I have used smoothing to smooth the weekly seasonality component out of both the twitter and searches data. That has yielded small improvement in the correlation coefficients.
- I have also used a very basic method of prediction which works as follows:

  > Calculate the mean and the standard deviation of the searches. If the standard deviation is more than the mean, then there has been a spike which has pushed it higher.
- I've also used that to determine which destinations should have a classifier built. That has yielded very small improvements.

The fact that the simple combination of LASSO + Ridge regression does not perform miraculously well has led my supervisor and I to believe that perhaps we should investigate more sophisticated models that will perhaps model the problem better.

I am currently doing:

- Reading the Adams & MacKay paper on Bayesian change point models
- Will look into the matlab implementation and try to port it to Python.

Notes from today:

In the model have:

- last 4 friday weights

- the twitter counts

Those would be the things that change.

(1) Implement the method in the Ryan Adams & David MacKay paper.
(2) Run this on some toy data.
(3) Run this on the search volumes without the Twitter data.
(4) Histogram of the change point to the nearest twitter peak.

Non-linear function of the features.

And SV regression with polynomial kernel.

Use relative error or use absolute error.

Average across days and destinations.

Do a histogram of median, box plot or any ways to drill down and show it clearly.

(1) take quantiles across destination.
(2) Median/mean absolute error.

Take mean over days or median over destinations.

Cluster the destinations by popularity - split into 5 groups.

report the mean absolute error for each of the groups.

(Add something to the prior that makes change point more likely if there is a twitter spike)

1/T (sum (y prediction - y actual) / y actual)

Having read the Adams and MacKay paper the well-log data seems to be the most relevant, because of the step changes observed.

After reading the paper I started the experimental part.

Both of online and offline work perfectly fine.

Currently doing some other assignments, but will make a start on this soon.

3. Week 6

The algorithm does not work out of the box which is expected.

What should be used for evaluation: 1/T (sum (y prediction - y actual) / y actual)

Things to try:

(1) AR(1)
(2) and then AR(K)

do it on weekly average or other data.