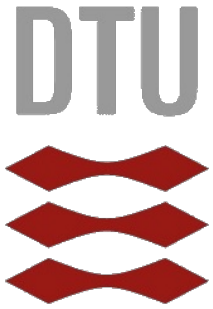

Danmarks
Tekniske
Universitet

42186 Model Based Machine Learning, Report I	
Date:	May 15, 2019
Name	Student no.:
Alexander Bilton	s165635
Gustav Hartz	s174315

Contents

1 Introduction

Congestion is a major problem for cities all around the world. Amongst other things, it erupts from the problem that too many people want to move at the same time of day. This is something that is visible in many different datasets e.g. freeway occupancy, public transport 'check in's' and bike lane activity. This project will analyze one of such phenomenons using New York taxi pickup data. The goal is to make accurate demand predictions that the taxi management can use when determining what the supply of taxi's should be. The modelling will be based upon the use of graphical models of temporal structure as well as other tools from the Model-Based Machine Learning toolbox.

Research question
We are a new NYC-based taxi company looking to beat the competition. Somehow, one of our co-founders have gotten her hands on the data of all pickups from the competitors between 2009 - 2016.
Our strategy is to be much better than our competitors at matching supply and demand and thereby operate at lower costs.
Our drivers are rather flexible, which means that we can call them in and send them home almost instantly. To solve this issue, the co-founder lets the data science team tackle the following question:

How can we predict the number of pickups, such that we know how many drivers we need for a given hour?

2 Data

We have worked with the dataset *pickups+weather_ wallstreet.csv*, which was made available during week 5 of the course. The dataset consists of the number of taxi pickups in the Wall Street area from 2009 through 2016. With 65712 observations, each observation corresponds to the number of pickups by taxi in the Wall Street area in the given hour, along with the given weather conditions for that hour.

Feature	**Type**	**Comment**
Datetime	Discrete	
Pickups	Discrete	
Date	Discrete	
Min_temp	Continous	
Max_temp	Continous	
Wind_speed	Continous	
Wind_gust	Continous	
Visibility	Continous	
Pressure	Continous	
Precipitation	Continous	
Snow_depth	Continous	
Fog	Binary	
Rain_drizzle	Binary	
Snow_ice	Binary	
Thunder	Binary	

Furthermore, we experimented with a supplementary data set, *holidays.csv*, which consists of major US holidays in the same period as we have taxi data for. The reason for adding this dataset was to test the hypothesis that demand during special dates (E.g Christmas, New Years, Boxing Day etc.) is higher.

Feature	Type	Comment
Date	Discrete	
Holiday	Discrete	Christmas, NYE etc..

2.1 Results from exploratory data analysis

Contrary to our initial hypothesis about special dates being important for modelling the number of pickups, the exploratory data analysis was not able to find a significant difference between the number of pickups on special days and the number of pickups on regular days. A reason for this could be either that the dataset contains too many "special" days, or that the actual demand for taxis has actually been higher, but not been met. Another reason could be the investigation: We only look at the overall travel pattern through the boxplot, and one could imagine that these special dates affect only the patterns of cetain hours (Who wants a taxi christmas eve at 6 pm?) Because of this result, the special dates was not taken into consideration during modelling.

3 Modelling

Since the NYC taxi data is a time series, we have chosen to model the data using markovian models. Namely, we will apply autoregressive models and state-space models. For all the applied models, the common ground is that and observations y_t is a linear combination of earlier observations, either explicitly in case of autogressive models, or through a latent variable in the state space models.

3.1 Model I - Autoregressive model of order 3 - AR(3)

For the first of our three models, we assume that the number of pickups for a given hour t, is a linear function of the previous three hours with some gaussian noise.

$$y_t \sim \mathcal{N}(\beta_1 y_{t-3} + \beta_2 y_{t-2} + \beta_3 y_{t-1}, \sigma^2)$$

Generative story
Given T, μ_β, σ_β, μ_σ, σ_σ, μ_y, σ_y and σ

1. Draw $\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta)$
2. Draw $\sigma \sim Cauchy(\mu_\sigma, \sigma_\sigma)$
3. Draw $y_1..y_3 \sim \mathcal{N}(\mu_y, \sigma_y)$
4. For each t in 4..T
    - Draw $y_t \sim \mathcal{N}(\beta_1 y_{t-3} + \beta_2 y_{t-2} + \beta_3 y_{t-1}, \sigma^2)$

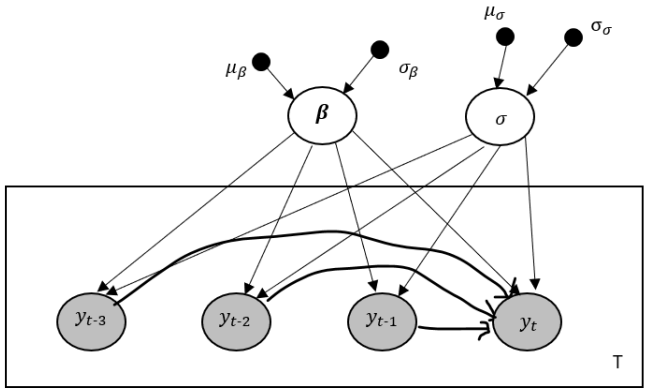

Figure 1: Probabalistic Graphical model for AR(3) model

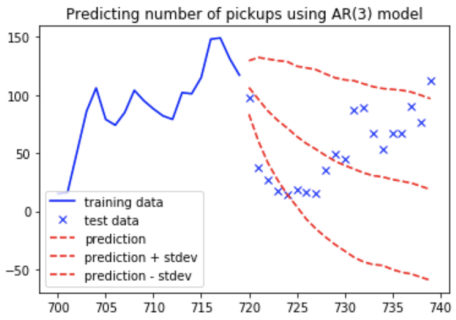

Figure 2: Resulting prediction based upon PGM (left) implemented in STAN

As can be seen from the above PGM, the temporal connectivity is modelled explicitly, as opposed to some of the other temporal models (E.g State-Space-Models). The prediction above reveals an important learning about the model. It looks like the model is not able to capture the periodic structure of the data, since observation y_t only 'knows' about its three previous observations, and not how it usually looks at the *same* time of the day.

3.2 Model II - State-Space-Model: AR(3)

The second model is a State-Space-Model, meaning that we introduce a set of hidden variables **H**, which we assume are generating our observations. That is, y_t is independent of other variables given h_t.

Generative Story
Given T, μ_β, σ_β, μ_τ, σ_τ, μ_h, σ_y and σ_h

1. Draw $\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta)$
2. Draw $\tau \sim \mathcal{C}auchy(\mu_\tau, \sigma_\tau)$
3. Draw h_1 .. $h_{24} \sim \mathcal{N}(\mu_h, \sigma_h)$
4. For each t in 25..T:
   - Draw $h_t \sim \mathcal{N}(\beta_1 h_{t-24} + \beta_2 h_{t-2} + \beta_3 h_{t-1}, \tau^2)$
   - Draw $y_t \sim \mathcal{N}(h_t, \sigma_y)$

As can be seen from the generative story, we have updated the temporal connectivity, such that h_t is a linear function of the last two hours and the *same* hour the day before. Thereby we try to capture the special events through the variables from the two earlier hours $h_{t-1..t-2}$, and get the periodicity through the h_{t-24} variable. The idea behind this intuition is that every hour has a somewhat standard travel pattern. The different days then vary from each other due to special events such as holidays, football games, festivals, people going out and so on. These fluctuations in demand is supposed to be captured by h_{t-1} and h_{t-2}, whereas the standard travel pattern for the hour as captured by h_{t-24}. Below is the PGM and the resulting prediction:

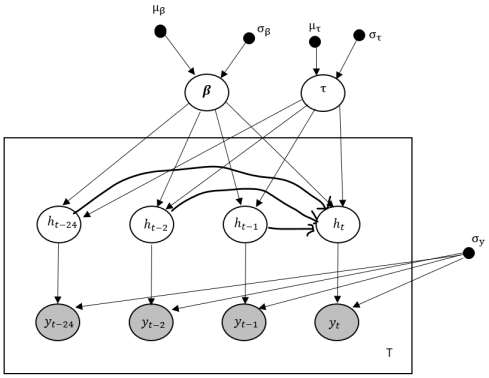

Figure 3: Probabalistic Graphical model for State-Space-AR(3) model.

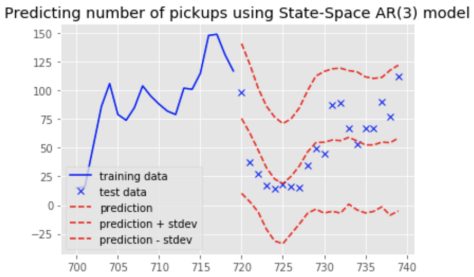

Figure 4: Resulting prediction based upon PGM (left) implemented in STAN

As we can see from the prediction plot, this model is better at capturing the structure of the data, but definitely not perfect. From the traceplots in the notebook, we can see that this model still results in a wide posterior which brings a lot of uncertainty into y_{hat} (the predictions). Reflecting over this model, there is one naive assumption in the model which can be improved. The model has no way to tell the difference between the weekdays. The h_{t-24} variable will try to take care of that, but imagine the prediction of demand for monday morning at 2 AM. The h_{t-24} variable will correspond to sunday morning, which definitely has a different demand than a night where most people have to get up to work. An effort to include this fact is what leads us to the next model.

3.3 Model III - State-Space-Model: AR(3) with inputs

The third and final model seeks to solve some of the flaws in the earlier model through input data consisting of one-out-of-k-coding the weekdays. Model II was created under the assumption that all days are similar and only differ due to special events. If we look at the initial data analysis, namely the boxplot for weekdays, this assumption does not quite hold. To incorporate this in our predictions the state-space-model is provided with aforementioned input. Below the outcome of this is shown, along with the generative story.

Generative Story

Given T, μ_β, σ_β, $\mu_{\beta x}$, $\sigma_{\beta x}$ μ_τ, σ_τ, μ_y, σ_y, μ_h and σ_h

1. Draw $\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta)$
2. Draw $\beta_x \sim \mathcal{N}(\mu_{\beta x}, \sigma_{\beta x})$
3. Draw $\tau \sim \mathcal{C}auchy(\mu_\tau, \sigma_\tau)$
4. Draw $\sigma_y \sim (\mu_y, \sigma_y)$
5. Draw h_1 .. $h_{24} \sim \mathcal{N}(\mu_h, \sigma_h)$
6. For each t in 4..T:
   - Draw $h_t \sim \mathcal{N}(\beta_1 h_{t-1} + \beta_2 h_{t-2} + \beta_3 h_{t-3}, \tau^2)$
   - Draw $y_t \sim (h_t + \beta_x \cdot \mathbf{x}, \sigma_y)$

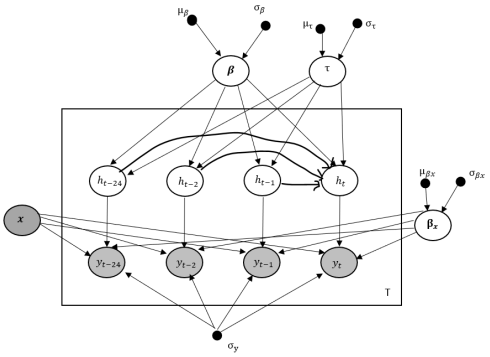

Figure 5: Probabalistic graphical model for State-Space-AR(3) model.

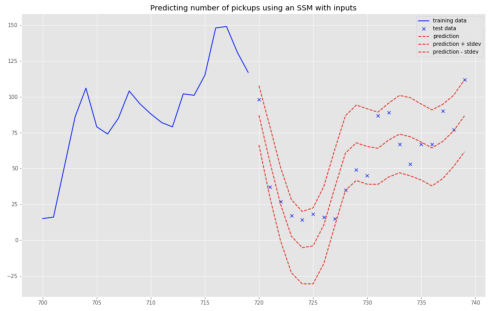

Figure 6: Resulting prediction based upon PGM (left) implemented in STAN

As we can see from the prediction plot above model performance dramatically increases when information about day of the week is added to the model. From the traceplots in the notebook, one can see that the posterior for y_hat is somewhat more narrow than before, which means that we reduced the uncertaincy, which is great.

4 Reflections

This project saught to model the demand of NYC taxi pickups through three different models. We saw an increase in performance for each model, with the third model ending up with fairly good predictions. We chose not to include the weather data available, because our focus was mainly on the time series aspect of the dataset. Including this might have been able to improve predictions a little. Also, a better domain knowledge on the area might have helped in selecting better priors, and thereby improving predictions. Lastly, it is worth mentioning that we experimented with different distributions for the pickup, amongst them the poisson, which seemed reasonable, but did not result in any useable model.

Other modelling ideas

A simple model which might also have performed well on this dataset is a pure regression model, which is widely used for these types of predictions. Another type of model, which we unsuccessfully tried to implement, is the gaussian processes. Through the covariance matrix in the gaussian process, dependencies between earlier states can be modelled. The model can learn these from the data, which means that it could include some interesting depedencies on earlier states, that we did not try. For instance, one could imagine that y_{t-168} (I.e the same hour on the same day, the week before), might be a good state to depend on.

5 Group work

Part	**Primary responsible**
Research question definition	Both
Exploratory data analysis	Alexander
Model I	Gustav
Model II	Alexander
Model III	Both
Report summaries	Gustav
Report conclusions	Both