Project Report

Ye Zhang

1 Introduction

This report is about the progress I made during the last month. There were totally three processes. First, I computed baseline of predicting error without using any machine learning methods. Second, I read some paper about DBN(Dynamic Bayesian Network) and studied some toolkits in Python to deal with DBN. Third, I used ARMA(Autoregressive-moving-average model) to do prediction and also built the DBN model.

2 Work Progress

2.1 Compute the baseline

Without using any Machine Learning algorithm, I just use the average waiting time in a certain bucket to predict the future waiting time. Firstly, I divided the whole datasets into half and used one as training data and the other as testing data. And I used 10 minutes as a bucket and also considered the weekday information. That is, Monday 7:00-7:10 is a bucket, Monday 7:10-7:20…Tuesday 7:00-7:10, Tuesday 7:00-7:10…Friday 7:00…Friday 17:50 all are distinct buckets. Then for each office, I calculated average waiting time in each bucket to predict the corresponding waiting time in the test dataset. Finally, I calculated the average prediction error over all the buckets over all offices which was 0.30256 hours. Later, I just considered hour and minute information without considering weekday. So Monday 7:00, Tuesday 7:00, Friday 7:00 and so on all belong to the same bucket. The average prediction error I obtained was 0.292847 hours.

2.2 ARMA model

As I mentioned above, there are two cases I considered. One is "Weekday+Time of day" and the other is "Time of day". I first considered "time of day model". When I only had time information without any additional spatial features like mobile logs or ringtone states, I don't think that DBN will be a good option. So I tried ARMA which is autoregressive-moving-average model. In this model, I used order 2 relations which was $x_t = c + \sum_{i=1}^{2} \varphi_i X_{t-i} + \varepsilon_t + \sum_{i=1}^{2} \theta_i \varepsilon_{t-i}$. I used "Statistics in Python" toolbox to train the model and predict the future waiting time. Unfortunately, the result was more than 0.4 hours which was worse than the baseline.

2.3 DBN model

In DBN model, I firstly discretized the waiting time. Specifically, I reserved only one decimal point for each value, and since the maximum waiting time is 3 hours, so there are only 31 possible discrete values for waiting time. Now, I am using MLE to compute the parameters in DBN. Next step, I'll do inference on waiting time.

3 Future Work

In the future, I might try linear regression and regression tree. I'm still confused with DBN model. In this DBN model, the observation value is determined. For example, if in this step, the time is "Monday 7:00", then the next observation must be "Monday 7:10" and the next observation must be "Monday 7:20". So the observation sequence is the same as the time sequence of the hidden state. It's weird to use time information of the hidden state as feature as well. So after trying DBN, I'll try other models.

4 Difficulty

I'm new in Python and I don't have strong research background, so I have to spend a lot of time learning Python and reviewing some math knowledge. What's more, I have a course which had substantial assignment in February. So I did not progress much during February. I'll have much more time after 5th March. I hope I can make more progress after that.