Lab 9

1. **Complete reference to the IEEE or ACM paper you are trying to outperform. Only articles published by IEEE or ACM will be accepted.**

   Lippi, M., Bertini, M., & Frasconi, P. (2013). Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. IEEE Transactions on Intelligent Transportation Systems, 14(2), 871-882

2. **Describe the results of the paper you have chosen.**

   The paper we've selected presents an experimental view of the different statistical & machine-learning approach to short-term traffic-flow forecast. They follow the approach of SARIMA and have proposed 2 new SVR models using a seasonal kernel to determine the similarity with the time-series examples. The results that they present confirm that seasonality is the key feature to achieve a high-accuracy. Though, the more accurate models usually require high computational resources – both while the training and prediction phase. Therefore, the seasonal kernel approach may be a reasonable compromise between the forecast accuracy and computational complexity. The SARIMA version that doesn't include a Kalman filter and the ANNs performed worst than SVR with an RBF kernel. This in-turn is less accurate than seasonal kernel variant.

   Furthermore, another important direction of the research paper that has been indicated by the experimental results presented in this paperconsists of investigating the covariate shift in traffic

3. **Describe your results so far, and your next objective in terms of data analysis.**

   We've compared the descriptive statistics of the datasets we've using various aggregation operators like mean, median, etc. Furthermore, we've applied different algorithm to the dataset and compared the different methods as well as compare them with the competitor's methods.

4. **Describe how you will use the concepts of aggregation and group operations**

   Wewill use the aggregation and group operations to look at the descriptive statistics of the dataset which will be helpful to understand what the mean average speed of traffic is by hour of the day to see if there is any difference

between average speeds by hour of day. The same analysis can be done with vehicle count.

5. Create a table showing the code for every operation in the left column and the time measurements for every operation in the right column. You need at least 10 operations. Highlight in bold the lines of code where you use aggregation or group operation concepts.

| Code | Time |
|---|---|
| roadtraffic = pd.concat(map(pd.read_csv, glob.glob(os.path.join("/Users/Amit/Downloads/traffic_feb_june", "*.csv")))) | 48 sec |
| %pyspark<br>roadtraffic.count() | 4 secs |
| %pyspark<br>import re<br>roadtraffic['hour'] = roadtraffic['TIMESTAMP'].str[11:13] | 5 secs |
| grouped_avgspeed_byhour =<br>roadtraffic['avgSpeed'].groupby(roadtraffic['hour'])<br>print(grouped.mean()) | **<1 sec** |
| grouped_vehicleCount_byhour =<br>roadtraffic['vehicleCount'].groupby(roadtraffic['hour'])<br>grouped_vehicleCount_byhour.mean() | <1 sec |
| %sql<br>select count(_id) from road_traffic | 10 secs |
| %pyspark<br>close_px = pd.read_csv('/Users/Amit/Downloads/stock_px.csv',<br>parse_dates=True, index_col=0)<br>close_px[-4:] | <1 sec |
| %pyspark<br># Annual correlation of Apple with Microsoft<br>by_year.apply(lambda g: g['AAPL'].corr(g['MSFT'])) | <1 sec |
| %pyspark<br>import statsmodels.api as sm<br>def regression(data, yvar, xvars):<br>  Y = data[yvar]<br>  X = data[xvars]<br>  X['intercept'] = 1.<br>  result = sm.OLS(Y,X).fit()<br>  return result.params | 3 secs |

| | |
|---|---|
| by_year.apply(regression,'AAPL',['SPX']) | |
| %sql<br>select * from road_traffic limit 15 --see the first 15 rows in the table | 1 sec |