Internship Report

presented by
Djibril Ndiaye
Apple is in the business of:
- electronic devices
- computer softwares
- online services

iCloud teams work on harnessing data from the different softwares of those devices that will be used to develop better services.
Search Process

* Applied to internships and jobs found online
  * *Interview with Oracle (data scientist)*
* Went to career fairs in Student Union
  * *ThoughtSpot contacted me a year later to interview for a different position*
* Applied to positions suggested in Dr. Bremer’s mail
  * *Interviewed with Alloy about four months later*
* Visited a data scientist at LinkedIn who promised to refer me to jobs
* Got referred to at Apple by a friend; got the job after interviewing

* Leave something a little complex and put it in GitHub with a link copied on your resume*
Mail size prediction

Data: daily utilization per user
- Features: user_id, mail_size, #of-mails, weekday, month
- Insight: yearly and weekly cycles, linearly increasing trend

Techniques used
- Linear Regression (Polynomial and trigonometric) {class}
- Neural Networks (3 layers, tanh activation) {Coursera Ng}
- Time Series Model (ARIMA and VAR) {In office}
- Hybrid models: LR and NN
- Embedding weekdays into weekend/businessDays

Key concepts: Bias-Variance tradeoff, Generalization error, Model capacity, Residual analysis

Next dataset: iCloud user subscription
<table>
<thead>
<tr>
<th>Application</th>
<th>Activities</th>
<th>Status</th>
<th>% Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>TensorFlow</td>
<td>Running deep neural nets on transformed data</td>
<td>Daily Task</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Using discretized feature columns to facilitate learning of daily mail averages</td>
<td>In Progress</td>
<td>On-going</td>
</tr>
<tr>
<td></td>
<td>Deeper exploration of the Tensorflow API: Checkpointing, Progressively loading big datasets into disk, random Mini batch generation…</td>
<td>Done</td>
<td>Completed</td>
</tr>
<tr>
<td></td>
<td>Writing a parameter search algo (exploring bayesian search optimization) using Hyperopt library</td>
<td>In Progress</td>
<td>On-going</td>
</tr>
<tr>
<td>Time series</td>
<td>Comparing ARIMA to vector autoregressive models</td>
<td>In Progress</td>
<td>On-going</td>
</tr>
<tr>
<td>Pandas</td>
<td>Data pipelines for another non parametric model using two dimensional averaging</td>
<td>Done</td>
<td>Completed</td>
</tr>
<tr>
<td></td>
<td>Combining dataframes usage and TensorFlow in a single reusable pipeline</td>
<td>In Progress</td>
<td>On-going</td>
</tr>
<tr>
<td>Scripter</td>
<td>Using scripter for more data exploration</td>
<td>Daily Task</td>
<td>NA</td>
</tr>
<tr>
<td>TuriBolt</td>
<td>Writing scripts for distributed training with real time reporting on master node</td>
<td>In Progress</td>
<td>On-going</td>
</tr>
<tr>
<td>Turi Blobby</td>
<td>Wrapping the current API to a higher level (2 to 3 main functions)</td>
<td>In Progress</td>
<td>On-going</td>
</tr>
<tr>
<td></td>
<td>Moving data from clusters to blobby via turibolt (Attempt)</td>
<td>In Progress</td>
<td>On-going</td>
</tr>
</tbody>
</table>
Softwares and tools

**Tools used**

- Programming language: Python, Bash
- Libraries: Numpy, Pandas, Matplotlib, Scikit-learn, Tensorflow, statsmodels, Boto3 (aws), Turi
- Access to a Spark cluster
- Access to a compute machine and internal storage

**Softwares used**

- Coding environment: Atom, Jupyter lab {out of class}
- Company apps: HipChat, Apple Directory, Radar {On the job}

Most communication through apple email and hipChat