Internship Report

presented by

Djibril Ndiaye

iCloud Big Data Analytics (Apple)

Apple is in the business of

* electronic devices
* computer softwares
* online services

iCloud teams work on
e different softwares of those devices that will be used to

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

ke something a little complex and put it in GitHub with a link copied

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

s: Bias-Variance tradeoff, Generalization error, Model capacity, Res

Next dataset: iCloud user subscription

Example of weekly report

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

Softwares and tools

Tools used

- Programming language: Python, Bash
- Libraries: Numpy, Pandas, MatplotLib, Scikit-learn, Tensorflow, st
- 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

END