ACM Winter Workshop: Summary

-Dimpy Chhabra

January 9, 2018

1 Introduction

The ACM-IGDTU student chapter conducted a five day workshop on data science and machine learning. The workshop included of working with the Twitter API, Facebook API, for data collection, preparation and data analysis in R, and machine learning concepts and applications using Scikit learn in python. Guest lecturers from Precog were invited to take various sessions too. The workshop culminated in a project discussion seminar and certificate distribution ceremony.

2 What did we learn?

Over the course of five days we went through the following laid out curriculum:

- Registration and Preparing Computing Environment
- Preliminaries: Python and R
- Data Collection from Social Media Platforms
- Issues and Challenges in Data Collection
- Data Analysis: Univariate and Bivariate
- Data Visualization
- Graph Data Analysis
- Application of Data Analysis
- Data Distributions, Parameter Estimation
- Hypothesis Testing
- Research Skills: Reading and Documentation
- Applications in Real World Problems
- Supervised Learning
- Regression
- Unsupervised Learning
- Applications of Machine Learning
- Key Research Areas
- Hearing from Participants
- Certificate Distribution and Valedictory Ceremony

Figure 1: Day 5 ceremony.

2.1 Day 1 Takeaways

Data Collection:

- Datasets
  ==>Ques Uno. Data Entities and attributes.
  ==>Publically Available

- APIS
  ==>Website's provide with API Access.
  ==>Ex. Graph API(Facebook)

- Web Scraping
  ==>Scraping the internet

- Selenium
  ==>Automating handled here

Above is the exhaustive list of getting data! Thus, one must be able to work on getting data out of APIs or the internet in general.

2.2 Day 2 Takeaways

On our second day we discussed various concepts:

- Data Representation
  We learnt about the matrix representation of data. ie. Data can be presented in a data Matrix in two dimension that is n and d. where the coloumns are the attributes and the rows are the tuples! Further discussed the two broad classifications of attributes: Numerical

and Categorical. We also discussed the two viewpoints: Probabilistic and Algebraic 'nd Geometric.
Thusly leading into various distributions like Bernoulli, binomial, normal and Poisson.

- Analysis: Univariate and Bivariate
  Waltz through the various univariate analysis like mean, mode and median; followed by implementation on IRIS dataset in R language.
  In the latter we discussed correlation and thusly, the development of graphs out of datasets on the basis of similarity degrees.

- Graph Data Analysis
  Studied degree distributions, walks, paths, connectedness, graph representations etc, to eventually dwell into developing similarity graphs. ex. Gaussian Similarity graph for Iris dataset.
  Ideas like
  -> Clustering Coefficient
  -> Centrality Analysis
  -> Eccentricity Analysis
  -> Closeness Centrality
  were discussed with real time possible applications of each, simply based on the twitter network.

- Other Topics discussed included of:
  ==> Baseline // GroundTruth
  ==> Small World Property and Ultra Small World Property
  ==> Power Law Relationship

2.3 Day 3 Takeaways

Day 3 included of detailed discussions and uses of various distributions. Namely, Bernoulli, binomial, Poisson and Normal.
This discussion was followed by the Hypothesis testing, and a simple hands-on session on git and github.

2.4 Day 4 Takeaways

On our fourth day we went head-first into machine learning. Theoretical sessions were followed by hand-on sessions, where we worked in python using Scikit learn.
The codes and representation of the work done can be found here.

2.5 Day 5 Takeaways

Certificate Ceremony and a session on 'how to read and write research papers', thus studying [Kau15]
The pictures of the same can be found here.

3 Conclusion

On and All, the workshop was a great learning experience! The various assignment problems and my solutions can be found on my github portal here.

Reach me @ http://www.dimpychhabra.co :)

References

[Kau15] Rishabh Kaushal. Improving Spam Detection in Online Social Networks. *2015 International Conference on Cognitive Computing and Information Processing(CCIP)*, (6):3–5, 2015.