Unsupervised Learning

Sarabjot Singh Anand, ssanand@tatrasdata.com

Problem Statement

You have been recruited as data scientists by a start-up, JhakaasNewsVala, based out of Mumbai. The company is developing an app that promises to deliver a unique news experience to its app users.

The company has identified it target market as working professionals in the age group 21-40.

Recognising the fact that retention (defined here as a visit after the first visit) is a huge issue for apps, they understand the need to make an impact on the first visit itself. The problem however is that they know nothing about the user interests or demographics at the time to personalise the news feed to them.

The company has acquired a corpus of 4829 news stories.

The real estate available for providing news stories (the mobile phone's screen) is limited and so without a scroll, only 10 stories can be displayed. Statistics show that the number of users scrolling beyond the first set of stories drops off very quickly unless a story on the first page catches the users eye (that is, results in a clickthrough).

You have been tasked with the job of building two intelligent bots.

1. The article recommender: This bot selects articles to serve a user. Inputs to the bot is the corpus of new articles and a user profile if available.
2. The user profiler: Once the user starts consuming news stories, (s)he leaves behind a clickstream of the form below:

UserId	SessionID	ArticleID Served	Article Rank	Click	Time Spent (seconds)
1	1	28	1	No	
1	1	66	2	No	
1	1	45	3	Yes	69
1	…	…	…	…	…
1	1	16	10	No	
1	2	36	1	Yes	46
1	…	…	…	…	…

   The bot must extract user interests from such data that can then be used for further personalisation for (her)his news feed.

The ultimate objective is to increase clickthrough and the frequency with which the user opens the app to consume stories. However, the objective in the first visit is to:

- Reduce bias in data collection (Example Bias: Stories that get served often and ranked higher, have a higher likelihood of being consumed (obtaining a clickthrough))
- Learn as much as possible about the users on their first visit
- Maximise coverage of the news corpus

Having learned about the trade-off between Exploiting what you know and Explore the space for what you don't during the CBA programme. How would you implement a strategy for the same within the project?

Team Size

3 students per team

Deliverables

Assignment: The Strategy and Implementation Plan (20% of the mark)

- Research
- Exploratory Data Analysis: What do you look for and what conclusions you draw
- Build Strategy
- Implementation Plan
- Anonymous Peer Review

Marks based on

- Depth of research
- Breadth of Application of what has been learned to date in the module
- Quality of proposed solution
- Individual Participation of team members

Submission Guidelines

An outline to be submitted by 10am on the **26th of February, 2018**. This outline should be no more than 2 A4 sheets outlining how you will tackle each of the aspects of the system using what you have learned from the first residential and the research you conducted.

Feedback will be provided to you by the 6th of March and you will be expected to incorporate the feedback within your proposed solution that you will submit on the 16th of March, 2018.

Mid-report: Management Presentation (10% of your mark)

- In the form of a 10 minute presentation delivered in second residential to full cohort
- To be submitted in Powerpoint form no later than **10pm on 16th March, 2018**

Marks based on

- Coherency of presentation
- Achievability of plan proposed in the time available

Final report (20% of your mark)

You are welcome to adapt your strategy based on what you learn from your peers or based on what you learn in the second residential. Acknowledge these changes in the report.

- Code
- Final Report focussed on what worked and what did not work as expected
- Future Work:
    - What would you do if you had more time?
    - What would you do differently?
- Anonymous Peer Review

Marks based on

- Inferences made from findings
- The completeness of what was delivered compared with what was planned

Submission Guidelines

The final report is to be submitted by 10am on the **31st of March, 2018**. This report should be no more than 6 A4 sheets.