Machine Learning Assignment

In this assignment you will be required to demonstrate your knowledge of machine learning techniques, your ability to apply technical skills to build actual machine learning models, and your capacity to evaluate state-of-the-art machine learning approaches on real world data. In addition you'll have the opportunity to win a share of a $50,000 pot!

H&M Personalised Fashion Recommendations

https://www.kaggle.com/c/h-and-m-personalized-fashion-recommendations

The H&M personalised fashion recommendations dataset is an open competition asking participants to build models capable of using customer and product metadata, as well as previous transactions, to make predictions about what customers will buy next. The training dataset consists of 3 files: a **customers.csv** containing (limited) metadata about each customer, **articles.csv** containing metadata about each product available for sale and **transactions_train.csv**, a list of purchases made by each customer.

The test data is not available to competition participants, but participants may make up to 5 submissions per day which will be evaluated on a hold-out testing set. A leaderboard tracks the highest performing models to date. When the competition finishes, all submissions will be re-evaluated on a new test set, and prizes are awarded to the top ranking submissions.

The Brief

To complete this assignment, you are required to develop and evaluate a machine learning solution to the H&M personalised fashion recommendations competition using Jupyter Notebook. Unlike the competition, which is interested only in results, in order to succeed in this assignment you will need to demonstrate your understanding of the concepts and techniques covered throughout the module. Your commentary and discussion of your approach and findings is more important than the code itself, and a high scoring competition submission does not guarantee a high result for the assignment, nor is it a prerequisite. Please read the rubric for a detailed breakdown of how marks will be awarded.

Evaluation

The Kaggle competition uses Mean Average Precision as its evaluation metric of choice. Although the Kaggle submission will give you a ranking, for the purposes of the assignment you are required to produce a local evaluation of your solution. You are not limited to the metric chosen by Kaggle for your own evaluation and may use whatever performance measures you feel would be appropriate when deploying a model such as this.

Additional Resources

By creating a free account on Kaggle you will have access to Kaggle notebooks. Kaggle notebooks allow you to execute your code in the cloud using public servers. This may be a useful option if you feel your hardware isn't up to the task of crunching through the data.

Ideas

Kaggle is a website which hosts Machine Learning competitions, but these competitions also tend to foster a spirit of collaboration, so there are lots of ideas floating around the discussion and code sections of the competition. These can be very useful resources, but please do bear in mind that the standard rules regarding plagiarism remain in force so attribution is essential. You will also need to demonstrate an understanding of any code you submit so treat these examples as guides rather than templates

Marking Scheme

Please **see the rubric** (attached as a separate file to the Brightspace assignment) for a full breakdown of how marks will be allocated for this assignment.

Submission

Submissions should include the following:

1. A single Jupyter notebook. Any required Python modules should be installed by the notebook. The notebook should expect to find the training data at the following locations (i.e. the input directory is located in the directory above the notebook)
   ../input/articles.csv
   ../input/customers.csv
   ../input/transactions_train.csv
2. A document of no more than **eight pages** (11pt Calibri or equivalent, 1.15 line spacing), including all diagrams and references, that describes what you did and why you did it. It should describe:
   (i) the approach you used;
   (ii) your model(s) in detail including features used, prediction algorithm(s) used, parameters, etc.;
   (iii) your local evaluation strategy;
   (iv) your findings with an appropriate discussion and analysis.
3. A screenshot of your Kaggle performance report