2017/18 COM6012 - Assignment 3

Assignment Brief

Deadline: 11:59PM on Friday 18 May 2018

How and what to submit

Create a .zip file containing three folders. One folder per exercise. Name the three folders: Exercise1, Exercise2, and Exercise3. Each folder should include the following files: 1) the .sbt file, 2) the .scala file(s), 3) the .sh file(s), 4) the files you get as outputs when you run your code in the HPC.

Please, also include in the .zip, a pdf file with your answers to all questions asked or figures required to produce, with explanations. Please, be concise.

Upload your .zip file to MOLE before the date and time specified above. Name your .zip file as NAME_REGCOD.zip, where NAME is your full name, and REGCOD is your registration code.

Please, make sure you are not uploading any dataset. Please check if you are unintentionally uploading unsolicited files.

Assessment Criteria

1. Being able to use scalable PCA to analyse and visualise big data [8 marks]
2. Being able to use scalable k-means to analyse big data [8 marks]
3. Being able to perform scalable collaborative filtering for recommendation systems, as well as analyse and visualise the learned factors. [9 marks]

Late submissions

We follow Department's guidelines about late submissions. Please, see this link

Use of unfair means

"*Any form of unfair means is treated as a serious academic offence and action may be taken under the Discipline Regulations.*" (taken from the Handbook for MSc Students). If you are unaware what constitutes Unfair Means, please carefully read the following link.

Exercise 1 [8 marks] PCA for visualising NIPS Papers

We use the NIPS Conference Papers 1987-2015 Data Set:

https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015

Please read the dataset description to understand the data. There are 5811 NIPS conference papers and we want to visualise them using PCA in a 2D space. We view each of the 5811 papers as a sample, where each sample has a feature vector of dimension 11463. **Note:** you need to carefully consider for input to PCA, i.e., what should be the rows and what should be the columns.

Step 1: Compute the top 2 principal components (PCs) on the NIPS papers. Report the two corresponding eigenvalues and the percentage of variance they have captured. Show the first 10 entries of the 2 PCs. [5 marks]

Step 2: Visualise the 5811 papers using the 2 PCs, with the first PC as the x-axis and the second PC as the y-axis. Each paper will appear as a point on the figure, with coordinates determined by these top 2 PCs. [3 marks]

Reference 1: I found an example on YouTube here for your reference https://www.youtube.com/watch?v=5nkKPfle4bc. You are free/encouraged to find your way though.

Reference 2: Sample submission command used in a similar task last year:

time spark-submit --driver-memory 40g --executor-memory 2g --master local[10] --conf spark.driver.maxResultSize=4g target/scala-2.11/pca_2.11-1.0.jar

Exercise 2 [8 marks] Clustering of Seizure Patterns

We use the Epileptic Seizure Recognition Data Set:
https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition

Please read the dataset description to understand the data.

Step 1: Use K-means to cluster all **data points** in this dataset with K=5. Note that the data provided are labelled (i.e., they are data points + their labels) so you need to **ignore** the labels (the last column) when doing K-means. [2 marks]

Step 2: Find the largest cluster and the smallest cluster by computing the cluster size (number of points). What is the size of the largest cluster and the size of the smallest

cluster? What is the distance between the largest cluster and the smallest cluster [3 marks]

Step 3. In clustering above, the labels available are not used, but they can serve as the ground truth clusters for us to examine the clustering quality. Note that each cluster found in Step 1 may contain data points of more than one labels. Find the majority label (i.e., the label with the most data points) for the largest cluster. And find the majority label for the smallest cluster. Report these two labels and their corresponding number of data points in respective clusters. [3 marks]

Exercise 3 (9 Marks) Movie Recommendation

We use the MovieLens 10M Dataset:

https://grouplens.org/datasets/movielens/10m/

Step 1: Run ./split_ratings.sh (included in the dataset zip file) to get the five splits (r1 to r5) for five-fold cross-validation. See ReadMe (also included in the zip file). [1 mark]

Step 2: Run ALS with default settings on the five splits and report the MSE for each split. [4 marks]

Step 3: Use PCA to reduce the dimensions of both movie factors and user factors (learned above) to 2 for the first split (r1). Plot two figures to visualise movies and users, respectively, using the 2 PCs, similar to the visualisation in Exercise 1-Step 2. [4 marks]