Computer Based Test 3 - Machine Learning (extended)

Clustering

Notes: This is a Computer Based Test. You will need a computer and MATLAB to solve it. This test is compulsory for students registered for the Machine Learning (extended) module. It contributes 1/3rd towards the 20% allocated for Computer Based continuous assessment marks. Recall that ML Extended has 40% continuous assessment, of which half (i.e. 20%) is Computer Based Test.

Students registered for the Machine Learning module are encouraged to try it, however they do not need to submit solution and it will not be counted towards their grade.

Submission Deadline: 23.59 on Friday 2nd December, 2016 (end of week 10). Late submission will carry a penalty of 10% loss per day. Submissions more than 72 hours delay will not be accepted. Canvas will not accept late submissions – if you're late, submit by email.

Deliverables: Submit a single PDF file on Canvas containing your report for all the tasks. The report should be structured as an academic report with a summary of: (i) investigations conducted, (ii) findings including any graphical plots, and (iii) your analysis of the findings. The MATLAB code developed, with due credits (to any borrowed part from any other source), should be included in appendices.

How to submit?: via Canvas by uploading a file.

Plagiarism Policy: Any work submitted in your name should be YOUR work, except where due credit is cited.

https://intranet.birmingham.ac.uk/as/studentservices/conduct/plagiarism/index.aspx

https://intranet.birmingham.ac.uk/as/studentservices/conduct/plagiarism/guidance-students.aspx

http://www.birmingham.ac.uk/Documents/university/legal/plagiarism.pdf

Marking Scheme: The submitted work will be graded on the basis of completeness, correctness, and neatness. You should include sufficient comments in MATLAB code to make it clear and understandable.

Any questions?

Please ask in good time, not in the last minutes before deadline! The best time to ask questions is during the weekly office hours on Tuesdays 9.30am-11am. Alternatively, you may ask or get appointment via email (k.m.rajpoot@cs.bham.ac.uk).

Test Tasks

Download the MATLAB code and data associated with this test: **CBT3.zip** (https://canvas.bham.ac.uk/files/3379390/download?download_frd=1). This includes example data and images (**kmeansdata.mat**, **kmeansnonlindata.mat**, **horse.jpg**, **MRI.jpg**, and **rgbeye.jpg**) and example MATLAB code (**kmeans_cluster2.m**, **kmeans_cluster3.m**, **script_kmeans.m**, **script_kmeans2.m** and **script_kmeans3.m**) for exploring clustering. Run the MATLAB code files (**script_kmeans.m**, **script_kmeans2.m** and **script_kmeans3.m**) in order to understand their operation by associating MATLAB commands with results obtained. This will provide you useful insights and MATLAB background to solve the tasks below. You are free to develop your own code or use this code with appropriate referencing.

You will also find **cbt3data.mat** data file which contains the functional magnetic resonance imaging (fMRI) data for developing solution to the tasks for this assignment. This data set consists of fMRI scans over 200 time points from 90 brain cortical regions for 20 subjects each of healthy and diseased groups. Your task is to explore clustering on the basis of finding similarity between temporal data (i.e. across the 200 time points) by grouping 90 brain regions into K groups for each subject. The results of clustering each subject can then be combined to form a community matrix of brain regions inter-activity for healthy group or diseased group (see for example bottom left and bottom right of Figure 1 in this paper http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0036733). You should read this paper to understand the application of k-means clustering in this problem and also to determine the brain region community matrix (Figure 1 bottom half).

Solve the following tasks:

T1: Perform k-means clustering on this data to group brain regions into K=10 clusters. [3]

T2: Perform k-means clustering by varying the number of clusters (K=20, and K=30). [1]

T3: Generate community matrix for K=10, K=20, and K=30 for healthy as well as for diseased groups. [3]

T4: Compare and comment on the results by looking at the differences between community matrices of healthy and diseased groups and the effect of K. [3]