Computer Based Test 2 - Machine Learning (extended)

Bayesian Classification

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 11th November, 2016 (end of week 7). 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 solutions for all the tasks. The PDF file should contain the MATLAB code developed, with due credits (to any borrowed part from the book) and summary of your findings including any graphical plots for each task. The first page should include the student name and ID number.

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: **CBT2.zip** (https://canvas.bham.ac.uk/files/3305621/download?download_frd=1). This includes sample data (**bc_data.mat**) and sample MATLAB code (**bayesclass.m** and **BayesTrainTest.m**) for exploring Bayesian classification. Run these MATLAB code files and try to fully understand their operation by associating MATLAB commands with results and plots 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 **cbt2data.mat** data file which contains the data for developing solution to the tasks for this assignment.

Solve the following tasks:

A new test for a disease is being developed. The test involves the measurement of the concentration of two chemicals in a urine sample. The test is less accurate than an existing blood test, but much faster and cheaper. In order to determine whether a test is "positive" or "negative", a classifier is desired that can make this decision.

In order to train the classifier, measurements on 500 patients were taken and for each, both blood and urine tests were performed. The blood test was used to canonically determine the "actual" target class (i.e. ground truth) of these training data, and from this initial labelling, the classifier is to be trained on the urine data.

The data file **cbt2data.mat** contains the discriminated training data in variables 'diseased' and 'healthy', with one variable per row, one data point per column. Your task is to determine a classification rule on this data, and to use it to classify 2000 unclassified data points (in variable "newpts"), and thus "diagnose" the disease state of the individual from whom they came.

T1: Train the Bayesian classifier, with maximum likelihood estimate, on the training data with and without Naïve assumption [2]

T2: Train the Bayesian classifier, with maximum a posteriori estimate, on the training data with and without Naïve assumption [2]

T3: Evaluate each classifier against the new data points and determine the class of each point ("diseased" or "healthy"). Compare and discuss the results. Note that you'll have a total of 4 classifiers from above 2 steps. [4]

T4: Visualise the results by plotting: [1]

- "Diseased" training points as red circles.
- "Healthy" training points as blue circles.
- "Diseased" new points as red x's.
- "Healthy" new points as blue x's.

T5: Comment on the overall results. [1]