EAI 320

Practical Assignment 7

11 May 2016

Compiled by Dr. Joel Dabrowski

Question 1 (k-Nearest-Neighbours)

The Iris dataset is a popular dataset used in pattern recognition literature. The dataset was created by R. A. Fisher. R. A Fisher was a founding father of modern statistical science. The dataset consists of measurements of 150 iris flowers. There are three classes of flowers; *Iris Setosa*, *Iris Versicolour* and *Iris Virginica*. Each sample in the dataset contains four attributes/features; *sepal length in cm*, *sepal width in cm*, *petal length in cm* and *petal width in cm*.

From the original dataset, a training set and a test set has been sampled for you. The dataset is contained in four data files; 'testData.data', 'trainData.data', 'testLabels.data' and 'trainLabels.data'. The 'testData.data' and 'trainData.data' consist of samples with the four attributes. Each row contains a sample. The attributes of each sample are delimited by commas. The 'testLabels.data' and 'trainLabels.data' f;iles contain the labels associated with the samples in the dataset files. The flower labels are denoted with integers such that; *Iris Setosa* = 1, *Iris Versicolour* = 2 and *Iris Virginica* = 3.

For this question, the k-nearest-neighbours algorithm with the Euclidean distance is to be utilised. Use the algorithm with the training set to classify the samples from the test dataset according to the following procedure:

a) Plot the dataset and the samples. Indicate which belong to which class and which are test samples. (Exclude the last feature in the plot)

b) Apply the k-nearest-neighbours algorithm with $k = 1$ to classify the test samples using the training samples. Note the number of errors.

c) Increment k in question (b) until there are zero errors. Note the number of errors for each value of k.

Question 2 (Linear Regression)

Consider a driver eyesight dataset presented in the 'signdist.data' file. The research firm, Last Resource, Inc. collected data on 30 drivers relating to their age and the distance they can see. The dataset consists of two features, age and distance. The data file is comma delimited. Samples are presented in the rows. Features are presented in columns.

a) Use linear regression to fit a straight line to the dataset.

b) Plot the dataset and the straight line.

c) What is the expected distance a 16 year old can see?

d) What is the expected distance a 90 year old can see?

Question 3 (Logistic Regression)

Consider a dataset relating to student semester test results and exam entrance. The dataset is presented in the ‘examX.data’ and ‘examY.data’ files. The ‘examX.data’ file contains the first and second semester test results for 80 students. The ‘examY.data’ file contains binary labels indicating whether the student had exam entrance or not.

a) Train a logistic regression classifier on the provided dataset.

b) Plot the dataset and the decision boundary.

c) What is the probability that a student gets exam entrance with semester test results of 20% and 80%? Plot this sample along with the dataset and decision boundary.

d) What is the probability that a student gets exam entrance with semester test results of 50% and 50%? Plot this sample along with the dataset and decision boundary.

Deliverables

- Write a technical report on your finding for this assignment.
- Include your code in the digital submission as an appendix, but leave it out for the hardcopy submission.

Instructions

- All reports must be in PDF format and be named report.pdf.
- Place the software in a folder called SOFTWARE and the report in a folder called REPORT.
- Add the folders to a zip-archive and name it EAI320_prac1_studnr.zip.
- All reports and simulation software must be e-mailed to *EAI320.UP@gmail.com* no later than 16:00 on 17 May 2016. No late submissions will be accepted.
- Place a hard copy of your report in the box in front of Eng 3 7-25 before the deadline.
- Submit your report online on ClickUP using the TurnItIn link.

Additional Instructions

- Do not copy! The copier and the copyee (of software and/or documentation) will receive zero for both the software and the documentation. Z-e-r-o.
- For any questions of appointments email me at *EAI320.UP@gmail.com*
- Make sure that you discuss the results that are obtained. This is a large part of writing a technical report.

Marking

Your report will be marked as follow:

- 60% will be awarded for the full implementation of the practical and the subsequent results in the report. For partially completed practicals, marks will be awarded as seen fit by the marker.
- 40% will be awarded for the overall report. This includes everything from the report structure, grammar and discussion of results. The discussion will be the bulk of the marks awarded.