EAI 320

Practical Assignment 7

Concept by Dr J. Dabrowski
Compiled, edited and reviewed by Johan Langenhoven

19 April 2018

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.

1. Iris Setosa.
2. Iris Versicolour.
3. Iris Vriginica.

Each sample in the dataset contains four attributes/features.

1. Sepal Length in cm.
2. Sepal Width in cm.
3. Petal Length in cm.
4. 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 namely: 'testData.data', 'trainData.data', 'testLabels.data' and 'trainLabels.data' (2 training files, 2 testing files). The 'testData.data' and 'trainData.data' files 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' files 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.
- Iris Virginica = 3.

The k-nearest neighbours algorithm is a simple algorithm that is quite powerful. When a new testing point is evaluated in a classification based scenario, the algorithm simply calculates the distance between the testing point and *all* of the training points, and selects the closest k values (the smallest distance). The algorithm then assigns a class to the new testing point based on the class of the majority of the nearest neighbours.

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

a) Using the first three attributes of the Iris dataset, plot the dataset and the samples. Indicate which belong to which class and which are test samples. This image will be a 3D image, as you are plotting the first three attributes with respects to each other. You can use colours or symbols to define the different classes.

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 the errors reach a minimum. Take note of the amount of errors for each value of k.

Question 2: Linear Regression

Regression is a category of problems where one wants to map continuous input variables to continuous output variables. Linear regression accomplishes this by finding a straight line that is the closest to all the points in a given dataset. Us

Even though the k-nearest neighbours algorithm is usually used for classification, it can also be used for regression - where continuous values are predicted instead of classes. Where the linear regression method takes the whole dataset into account, the kNN algorithm uses the local average of the k-nearest points to predict the output of the input variables.

For this question, you are expected to compare the linear regression and kNN regression methods.

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. This means that you should find a straight line going through the data that has the smallest overall error. The error is determined by evaluating the distance between the sample point and the straight line.

b) Plot the dataset and the straight line.

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

d) What is the expected distance that an 85 year old can see?

e) Repeat steps (a) - (d), using the kNN algorithm, instead of linear regression.

f) Comment on the distribution of the dataset and the manner in which k-nearest neighbours implements regression with regards to the amount of training samples.

Question 3: Logistic Regression

Logistic regression is a manner of classification using the logistic function and probabilities for the predictions of a binary outcome. The logistic function's equation is as follows:

$$\sigma(x) = \frac{1}{1 + e^{-(\alpha x + \beta)}}$$

Where α and β are parameters that determine the slope and intersection of the logistic function, respectively. The first task for this question is to determine:

a) How to use a logistic regression model for classification, and

b) How to calculate the values of α and β for the logistic model.

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.

After you understand what a logistic classifier is and how to implement and use it, the following tasks need to be done.

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

b) Plot the dataset and the decision boundary. A decision boundary is the curve in the input space that separates one class from another.

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.

Report

You have to write a short technical report for this assignment. Your report must be written in LaTeX. In the report you will give your results as well as provide a discussion on the results. Make sure to follow the guidelines as set out in the separate questions to form a complete report.

Reports will only be handed in as digital copies, but a hard copy plagiarism statement needs to be handed in at the following week's practical session (on the final day of the practical submission).

Deliverable

- Write a technical report on your finding for this assignment.
- Include your code in the digital submission as an appendix.

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_prac7_studnr.zip`.
- All reports and simulation software must be e-mailed to **up.eai320@gmail.com** no later than 19:00 on 27 April 2018. No late submissions will be accepted.
- Use the following format for the subject header for the email: EAI 320 Prac 7 - studnr.
- Bring your plagiarism statements to the practical session on Thursday, 26 April 2018, where they will be collected.
- 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.
- For any questions or appointments email me at **up.eai320@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. **Commented code allows for easier marking!**
- 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.