Assignment 1: Decision Tree Learning for Cancer Diagnosis

February 27, 2017

Description of the Assignment

In this assignment, you will implement a decision-tree algorithm and apply it to breast cancer diagnosis. For each patient, an image of a fine needle aspirate (FNA) of a breast mass was taken, and eight features in the image potentially correlated with breast cancer were extracted. Your task is to develop a decision tree algorithm, learn from data, and predict for new patients whether they have breast cancer. Dataset is attached.

1. In the dataset (in CSV format), each patient is represented by one line, with columns separated by commas: the last is the class (benign or malignant), the rest are attribute values which are integers ranging from 1 to 10. The first line of the dataset contains the name of the attributes. You can safely discard this line if required.
2. Implement the ID3 decision tree learner, as described in class. You may program in C, C++, Java, Python or Matlab. You can't use any third-party tool or function directly related to the problem. Your program should assume input in the above format.
3. Use information gain for evaluation criterion.
4. Use your algorithm to train a decision tree classifier. Use 5-fold cross-validation. Use accuracy for performance evaluation. Calculate the following measures for your learned model.
    a. True positive rate (sensitivity, recall, hit rate)
    b. True negative rate (specificity)
    c. Positive predictive value (precision)
    d. Negative predictive value ()
    e. False positive rate (fall-out)
    f. False negative rate (miss rate)
    g. False discovery rate
    h. F1 score

Instructions for Report Writing

1. Put the performance measures in tabular format
2. Answer the following questions in your own language:
    a. Why are you using cross validation? Do the dataset justify it?
    b. Besides accuracy, which of the criteria mentioned above should be used in cross validation for the given data set? Explain.
3. Never copy the report. Just answer the questions precisely. Make it as simple as possible. Too much description is not needed.

Special Instructions

1. Don't Copy anything! If you do copy from internet or from any other person or from any other source, you will be severely punished and it is obvious. More than that, we expect Fairness and honesty from you. Don't disappoint us!
2. The report should be in .docx/.pdf (No hardcopy is required). Write precisely in your own language and keep it as simple as possible.
3. Upload the code and report in Moodle within 10:00 P.M. of 12th March, 2017 (Sunday). This is a strict system-imposed deadline for both section A and B.
4. For Python and Matlab, you may not get supporting software in the lab. If you do program in these languages, bring your computer in the sessional.
5. You are allowed to show the assignment in your own laptop during the final submission. But in that case, ensure an internet connection as you have to instantly download your code from the Moodle and show it.

Instructions for Moodle upload

1. Upload the assignment within the specified time. Otherwise, we can't accept it.
2. If you write code in a single file, then rename it as <Student id> <code>.<extension>. For example, if your student id is 1105123 and you have done in java, then your file name should be "1105123 code.java".
3. The report name should be <Student id> <report>.<extension>. For example, if your student id is 1105123 and it is in pdf format, then the report name should be "1105123 report.pdf".
4. Finally make a main folder, put the code and report in it, and rename the main folder as your <Student id> <Programming language>. For example, "1105123 Java". Then zip it and upload it.