QBUS6850: Tutorial 7 – Advanced Classification Techniques I

Tutorial Tasks

If you have completed the main tasks under the tutor's guidance, you may continue the following tasks

Task 1 - Classification Criteria

In the section "Lecture06_Example01_Updated.py" you were shown Python code to calculate a single layer decision tree.

This code used entropy as the classification criteria. Adapt the original code to a different classification criteria in Lecture 6 or from this list http://scikit-learn.org/stable/modules/tree.html#classification-criteria

Does the different criteria produce a different tree? Why or why not?

Task 2 - Lending Club

Lending Club is a peer-to-peer lending organisation. On the platform you can create a loan and lenders can browse loans and select loans that they would like to invest in.

As an investor on the Lending Club platform you wish to make safe choices with your investment and want to predict whether a loan is likely to be repaid or fall into default.

Fortunately Lending Club makes loan default data available publicly.

Your task is to:

- Build a decision tree to classify whether a loan will be repaid or not

You can download the 2007-2011 LOAN dataset and data dictionary from https://www.lendingclub.com/info/download-data.action.

The data "LoanStats3a_2.csv" is also available on Canvas.

You should:

- split the data into train and test sets
- optimise your tree by finding the best tree depth and criteria
- print our the optimal classifier
- visualise/display the tree structure
- evaluate your tree's performance on the test set

You should think about whether you need to prune any features from the data.

Hint: decision trees split by finding the feature that maximises purity. In other words only features that are useful to the decision making are retained.