Exercises, Lecture 8

1. The first few exercises concern the fictitious movie dataset below. A real database with information on movies and TV-series is available in different formats from http://www.imdb.com/interfaces. The goal here is to predict the user rating of an upcoming movie as soon as the information on cast and budget is known, and to be able to explain in words why the movie will go straight to the Oscar's – or straight to oblivion.

Observation	Actor	Budget ($million)	IMDb User Rating
1	Nicholas Cage	100	2.8
2	Scarlett Johansson	50	8.3
3	Scarlett Johansson	150	4.0
4	Nicholas Cage	20	2.9
5	Al Pacino	75	7.8
6	Al Pacino	150	8.1
7	Al Pacino	115	3.0
8	Nicholas Cage	115	3.0

   a. Is the described problem a classification or regression problem? Motivate your answer.
   b. Which input variables are categorical and which are continuous?
   c. For a categorical variable with k unique categories, what is the number of possible splits into two groups? Note that empty groups are not allowed, and that groupings are commutative in the sense that e.g. the split {1,2,3},{4,5} is equal to the split {4,5},{1,3,2}.
   d. What is the total number of splits to investigate at the root node for the movie dataset?
   e. Use Matlab to build a tree that predicts IMDb ratings. Notice that you can prune the tree directly in the window.

2. Matlab has excellent functions for building, pruning, evaluating and viewing classification and regression trees. We are going to use them to diagnose heart problems based on a set of 13 clinical variables from 303 patients and healthy controls.
   a. Read the help files (using `doc`) for `fitctree`, `cvLoss`, `predict` and `view` to familiarize yourself with the possibilities
   b. We have a data set from healthy controls and patients with heart problems in the Excel file `ClevelandHeartData.xls`. The first 13 columns are different features and the 14th column is an indicator for heart problem/healthy. You can read more about the data in `ClevelandHeartDataDescription.txt.`
   c. Import data to Matlab using `xlsread`.
   d. Convert output from xlsread into varibles that Matlabs tree function needs.
   e. Build a large tree with `'MinLeaf'` set to 1 and view the tree.
   f. Choose optimal tree size by tuning the `'MinLeaf'` value using cross validation.
   g. View the optimal tree and try to interpret it such that it makes sense for a doctor.