1/04/16 DA - EXAM QUESTIONS RULEFIT

- A mixture of regression and trees. Random Forest only uses trees
- RF are v good at picking up linear relationships, but rulefit uses trees for non-linear relationship and linear regression for detecting linear relationships
- Rulefit grow a series of trees like RF but then use individual from the variables to form logistic regression equation
- The importance metric that rulefit uses include a combination of the importance from trees and linear regression RF just use trees.

Output of Rulefit

- The rulefit program outputs all prediction function (i.e, rules and linear functions) included in the ensemble with their respective coefficients.
- The coefficients can be interpreted as regular regression coefficients: they represent the increase in the mean predicted value for a unit increase in the prediction function
- The output provides a measure of importance of every prediction function and input variable in the ensemble
- Importances for prediction functions are given by the absolute value of the coefficient multiplied by their support
- The importance of an input variable is calculated as a weighted sum of the importances of the prediction functions in the ensemble in which the variable appears
- Importances are rescaled to have max 100 and min 0.
- Inspection of importance provides an overview of the relative importance of each predictor variable to the predictive model

- First component produces "rules" and the second component fits a linear model with the rules as inputs
- highly interpretable, decision rules have an easy understandable format
- Rules have simple form: if $X2<3$ and $X5<7$ then 1 else 0
- Rules are generated from the covariate matrix X
- Derived from decision trees by basically ripping them apart - every path in a decision tree can be used into a rule
- 4 rules generated from tree with 3 terminal nodes

- Another way to see this step is as a black box, that generates a new set of features x' out of your original features x. These features are binary and can represent quite complex interactions of your original x. The rules are chosen to maximize the prediction/classification task at hand.

Part II Fit the rules

- You will get alot of rules from the first step (what you want)
- Want to reduce the number of rules. - Lasso or regularized regression.
- Next to the rules add all numeric observations from your original dataset will be used in lasso linear model
- Every rule and numeric variable get a coefficient (beta)
- Thanks to regularization, a lot of the betas will be estimated to zero
- Outcome is a linear model that has linear effects to all of the numeric variables and also linear terms for the rules
- Tells us: variable importance, degree of relevance of original input variables and interaction effects between variables