Statistical Learning Methods, summer semester 2017/18 – Practical exam, groups 100 & 101

**During exam you can use Internet, codes, books or notes but communication with each other (and other people) is forbidden.**

**Ex. 1 (10 p)**

Investment fund XYZ considers buying personalized software which can be used for forecasting of stock prices. Price of such software amounts to $20.000. If the company decides to buy this software, it has to pay the price and can potentially gain in profit, if they pass on this opportunity there is no cost nor profit.

Company’s IT team is not so experienced and the configuration of software is complicated. Because of this, if the company decides to buy the software, there is only 10% chance, that configuration will be successful and company will gain $50.000 thanks to it.

If IT department fails to implement the software, the company can hire external consultants to help them. Cost of such service is additional $15.000. Unfortunately, the software is a new product so there is only one company which offers installation of it, and it is a new company with no reputation – therefore there is a 5% chance that even consultants will fail to configure software correctly. If consultants do their work, our investment fund will gain $50.000, if not, there will be no additional gains/losses.

Based on provided information please prepare decision tree using SilverDecisions (<http://www.silverdecisions.pl/>) and based on it suggest whether investment fund XYZ should buy the software? If yes, should it be prepared to hire consultants? In other words, provide optimal decision path. Company wants to maximize its profit.

Please write your answer in R script as a comment and additionally send .pdf file with exported tree (include your student number in one of nodes’ labels as a proof of work).

**Ex. 2 (15p)**

Please use **Participation** dataset from **Ecdat** library. Your object is to prepare predictive model with variable ***lnnlinc*** as dependent variable**.**

1. (2p) Plot the distribution of dependent variable.
2. (2p) Divide data into training (80% of observations) and validation (20% of observations) sets (with seed equal to your student number).
3. (2p) Fit original least squares (OLS) with all variables as predictors on training set. Is variable *age* statistically significant (write answer as comment)?
4. (2p) Train a neural network consisting of 5 neurons using *nnet* package. Please use following parameters: decay=3, lineout=TRUE, rest as default.
5. (3p) Find optimal value of decay parameter. Use MSE (Mean Squared Error) as criterion and minimize it on validation set. Consider 20 potential values, sequenced from 0 to 5, with equal step (write optimal value as comment).
6. (4p) Make forecasts for validation set for models from points 3,4 and 5. Assess quality of these models using MAE (Mean Absolute Error) and point best one (write answer as comment).

**Ex. 3 (25 p)**

Use ‘**college.csv’** dataset with statistics for US Colleges. Your dependent variable is ***Private****,* which takes value 1 for private college and 0 for public college. Values in the dataset are separated by comma and dot is a decimal point.

1. (1p) Check if there are any missing values in the dataset.
2. (2p) Check the percentage of observations where college is private (write answer as comment).
3. (2p) Prepare a plot presenting the distribution of *Private* variable.
4. (2p) What is the mean value of variable *Expend* for *Private* equal to 1 (write answer as comment)?
5. (1p) Divide data into train (70% of observations) and validation (30% of observations) sets (with seed equal to 0).
6. (1p) Prepare a decision tree with rpart package on training set (dependent variable is *Private* and your goal is to predict if school was private; take default parameters, take cp=0.00001, minsplit=2, rest default).
7. (3p) Find optimal value of cp parameter using cross validation (write optimal value in comment).
8. (2p) Prune your tree for optimal value of cp.
9. (3p) Write one decision rule for unpruned tree (if… then…, write as a comment).
10. (5p) Check model quality (pruned) on the validation set: prepare confusion matrix, calculate model accuracy and true positive rate and false positive rate.
11. (2p) Draw a ROC curve for pruned tree.
12. (1p) Check the AUC and interpret it (write as a comment).

**Your submission should contain:**

* **pdf with the decision tree from SilverDecisions (with your student number in one of nodes’ labels)**
* **executable R script with codes used and answers to questions as comments where specified.**

**Please send your submission via e-mail on:** [**nosarzewski.aleks@gmail.com**](mailto:nosarzewski.aleks@gmail.com)**. Late submissions will not be accepted.**

**Good luck!**