732A95 Introduction to Machine Learning/ TDDE01 Machine Learning
Division of Statistics and Machine Learning
Department of Computer and Information Science

Computer lab 3

Instructions

- Create a report to the lab solutions in PDF.
- Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments.
- **Include all your codes as an appendix into your report.**
- **Use set.seed(12345) for every piece of code that contains randomness**
- A typical lab report should 2-4 pages of text plus some amount of figures plus appendix with codes.
- The lab report should be submitted via LISAM before the deadline.

Assignment 1. LDA and logistic regression

The data file **australian-crabs.csv** contains measurements of various crabs, such as Frontal lobe, Rear width and others

1. Use australian-crabs.csv and make a scatterplot of carapace length (CL) versus rear width (RW) where observations are colored by Sex. Do you think that this data is easy to classify by linear discriminant analysis? Motivate your answer.
2. Sometimes it can be interesting to see a decision boundary between classes, and lda() function does not provide this information directly. Thus, implement LDA with proportional priors, inputs RW and CL and output Sex for this data yourself (**use only basic R functions)**, classify the observations and extract the discriminant functions and equation of the decision boundary.
3. Make a plot of the original data **australian-crabs.csv** coloured by the classification label obtained and plot also the decision boundary. Comment on the quality of fit.
4. Make a similar kind of classification by logistic regression (use function glm()), plot the classified data and present the equation of the decision boundary. Compare this result with the LDA result.

Assignment 2. Analysis of credit scoring

The data file **creditscoring.xls** contains data retrieved from a database in a private enterprise. Each row contains information about one customer. The variable good/bad indicates how the customers have managed their loans. The other features are potential

predictors. Your task is to derive a prediction model that can be used to predict whether or not a new customer is likely to pay back the loan.

1. Import the data to R and divide into training/validation/test as $50/25/25$
2. Fit a decision tree to the training data by using the following measures of impurity
    a. Deviance
    b. Gini index

    and report the misclassification rates for the training and test data. Choose the measure providing the better results for the following steps.

3. Use training and validation sets to choose the optimal tree depth. Present the graphs of the dependence of deviances for the training and the validation data on the number of leaves. Report the optimal tree, report it's depth and the variables used by the tree. Interpret the information provided by the tree structure. Estimate the misclassification rate for the test data.
4. Use training data to perform classification using Naïve Bayes and report the confusion matrices and misclassification rates for the training and for the test data. Compare the results with those from step 3.
5. Repeat Naïve Bayes classification but use the following loss matrix:

$$L = \begin{matrix} & \\ Observed & \begin{matrix} good \\ bad \end{matrix} \end{matrix} \begin{matrix} Predicted \\ \begin{matrix} good & bad \end{matrix} \end{matrix} \begin{pmatrix} 0 & 1 \\ 10 & 0 \end{pmatrix}$$

    and report the confusion matrix for the training and test data. Compare the results with the results from step 4 and discuss how the rates has changed and why.

Submission procedure

Assume that X is the current lab number, Y is your group number.

If you are neither speaker nor opponent for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members submits the group report using *Lab X group report* in the *Submissions* folder before the deadline

If you are a speaker for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members does the following before the deadline:
    - submits the group report using *Lab X group report* in the *Submissions* folder before the deadline
    - Goes to Study room *Group Y*→*Documents* and opens file *Password X.txt*. Then the student should put your group report into ZIP file *Lab X_Group Y.zip* and protect it with a password you found in Password X.txt
    - Uploads the file to *Collaborative workspace* →*Lab X* folder

If you are opponent for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members submits the group report using *Lab X group report* in the *Submissions* folder before the deadline
- After the deadline for the lab has passed, go to Collaborative workspace→*Lab X* folder and download the appropriate ZIP file. Open the PDF in this ZIP file by using the password available in *Course Documents*→*Password X.txt*, read it carefully and prepare (in cooperation with other group members) **at least three questions/comments/improvement suggestions per lab assignment** in order to put them at the seminar.