732A99/732A68/ TDDE01 Machine Learning
Division of Statistics and Machine Learning
Department of Computer and Information Science

Computer lab 2

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

- **To be solved by TDDE01 students**

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. Make LDA analysis with target Sex and features CL and RW and proportional prior by using lda() function in package MASS. Make a scatter plot of CL versus RW colored by the predicted Sex and compare it with the plot in step 1. Compute the misclassification error and comment on the quality of fit.
3. Repeat step 2 but use priors $p(Male) = 0.9, p(Female) = 0.1$ instead. How did the classification result change and why?
4. Make a similar kind of classification by logistic regression (use function glm()), plot the classified data and compute the misclassification error. Compare these results with the LDA results. Finally, report the equation of the decision boundary and draw it in the plot of the classified data.

732A99/732A68/ TDDE01 Machine Learning
Division of Statistics and Machine Learning
Department of Computer and Information Science

Assignment 2. Analysis of credit scoring

- **To be solved by TDDE01/732A99/732A68/PhD course students**

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: use data partitioning code specified in Lecture 1e.
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. Use the optimal tree and the Naïve Bayes model to classify the test data by using the following principle:

   $$\hat{Y} = 1 \; if \; p(Y = 'good'|X) > \pi, otherwise \; \hat{Y} = 0$$

   where $\pi = 0.05,0.1,0.15, \ldots 0.9,0.95$. Compute the TPR and FPR values for the two models and plot the corresponding ROC curves. Conclusion?

6. Repeat Naïve Bayes classification as it was in step 4 but use the following loss matrix:

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

   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.

732A99/732A68/ TDDE01 Machine Learning
Division of Statistics and Machine Learning
Department of Computer and Information Science

Assignment 3. Uncertainty estimation

- **To be solved by 732A99/732A68/PhD course students**

The data file **State.csv** contains per capita state and local public expenditures and associated state demographic and economic characteristics, 1960, and there are variables

- MET: Percentage of population living in standard metropolitan areas
- EX: Per capita state and local public expenditures ($)

1. Reorder your data with respect to the increase of MET and plot EX versus MET. Discuss what kind of model can be appropriate here. Use the reordered data in steps 2-5.
2. Use package **tree** and fit a regression tree model with target EX and feature MET in which the number of the leaves is selected by cross-validation, use the entire data set and set minimum number of observations in a leaf equal to 8 (setting *minsize* in *tree.control*). Report the selected tree. Plot the original and the fitted data and histogram of residuals. Comment on the distribution of the residuals and the quality of the fit.
3. Compute and plot the 95% confidence bands for the regression tree model from step 2 (fit a regression tree with the same settings and the same number of leaves as in step 2 to the resampled data) by using a non-parametric bootstrap. Comment whether the band is smooth or bumpy and try to explain why. Consider the width of the confidence band and comment whether results of the regression model in step 2 seem to be reliable.
4. Compute and plot the 95% confidence and prediction bands the regression tree model from step 2 (fit a regression tree with the same settings and the same number of leaves as in step 2 to the resampled data) by using a parametric bootstrap, assume $Y \sim N(\mu_i, \sigma^2)$ where μ_i are labels in the tree leaves and σ^2 is the residual variance. Consider the width of the confidence band and comment whether results of the regression model in step 2 seem to be reliable. Does it look like only 5% of data are outside the prediction band? Should it be?
5. Consider the histogram of residuals from step 2 and suggest what kind of bootstrap is actually more appropriate here.

Assignment 4. Principal components

- **To be solved by TDDE01/732A99/732A68/PhD course students**

The data file **NIRspectra.csv** contains near-infrared spectra and viscosity levels for a collection of diesel fuels. Your task is to investigate how the measured spectra can be used to predict the viscosity.

1. Conduct a standard PCA by using the feature space and provide a plot explaining how much variation is explained by each feature. Does the plot show how many PC should be extracted? Select the minimal number of components explaining at least 99% of the total variance. Provide also a plot of the scores in the coordinates (PC1, PC2). Are there unusual diesel fuels according to this plot?
2. Make trace plots of the loadings of the components selected in step 1. Is there any principle component that is explained by mainly a few original features?
3. Perform Independent Component Analysis with the number of components selected in step 1 (set seed 12345). Check the documentation for the fastICA method in R and do the following:
    a. Compute $W' = K \cdot W$ and present the columns of W' in form of the trace plots. Compare with the trace plots in step 2 and make conclusions. What kind of measure is represented by the matrix W'?
    b. Make a plot of the scores of the first two latent features and compare it with the score plot from step 1.

Special task 3 (individual, optional)

Refer to Assignment 1.

1. 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.
2. 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.

Special task 4 (individual, optional)

1. Refer to Assignment 1 and divide data into training/test as 50/50, use seed 12345.
2. Implement Naive Bayes (**use only basic R functions)** that uses nonparametric density estimation method (use density()) to predict Species of different crabs based on all available crab measurement variables. Report training and test misclassification errors.
    a. **Hint**: density() function does not have predict() function but it evaluates predictions on a given grid. To make prediction for a vector of new values, you may call density() several times and specify one prediction point at a time, i.e. interval [a,b]=[x(i),x(i)].

732A99/732A68/ TDDE01 Machine Learning
Division of Statistics and Machine Learning
Department of Computer and Information Science

Submission procedure

First read 'Course Information.PDF' or 'Course Information- PhD.PDF' at LISAM, folder 'Course documents'

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 want to submit special tasks, use *Lab X special tasks* item in *Submissions.*

If you are a speaker for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- If you want to submit special tasks, use *Lab X special tasks* item in *Submissions.*
- 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.
- If you want to submit special tasks, use *Lab X special tasks* item in *Submissions.*
- 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.