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

R data file “**iris**” (present in the default R installation) shows the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

1. Make a scatterplot of Sepal Width versus Sepal Length where observations are colored by Species. Do you think that this data is easy to classify by linear discriminant analysis? Motivate your answer.
2. Use basic R functions only to implement Linear Discriminant Analysis between the three species based on variables Sepal Length and Sepal Width:
   a. Compute mean, covariance matrices (use cov()) and prior probabilities per class and report them
   b. Compute overall (pooled) covariance matrix and report it
   c. Report the probabilistic model for the LDA
   d. Compute discriminant functions for each class
   e. Compute equations of decision boundaries between classes and report them

   Do estimated covariance matrices seem to fulfill LDA assumptions?
3. Use discriminant functions from step 2 to predict the species from the original data and make a scatterplot of Sepal Length versus Sepal Width in which color shows the predicted Species. Estimate the misclassification rate of the prediction. Comment on the quality of classification. Afterwards, perform the LDA analysis with lda() function and investigate whether you obtain the same test error by using this package. Should it be same?
4. Use Models reported in 2c to generate new data of this kind with the same total number of cases as in the original data (hint: use sample() and rmvnorm() from

package **mvtnorm**). Make a scatterplot of the same kind as in step 1 but for the new data and compare it with the plots for the original and the predicted data. Conclusions?

5. Make a similar kind of classification by logistic regression (use function multinom() from **nnet** package), plot the classified data and compute the misclassification error. Compare these results with the LDA results.

Assignment 2. Decision trees and Naïve Bayes for bank marketing

The data file **bank-full.csv** is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Input variables:
# bank client data:
1 - age (numeric)
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')
# related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: 'cellular','telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
# other attributes:
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
# social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (target):
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

1. Import the data to R, **remove variable "duration"** and divide into training/validation/test as 40/30/30: use data partitioning code specified in Lecture 1e.
2. Fit decision trees to the training data so that you change the default settings one by one (i.e. not simultaneously):
    a. Decision Tree with default settings.
    b. Decision Tree with smallest allowed node size equal to 7000.
    c. Decision trees minimum deviance to 0.0005.

    and report the misclassification rates for the training and validation data. Which model is the best one among these three? Report how changing the deviance and node size affected the size of the trees and explain why.

3. Use training and validation sets to choose the optimal tree depth in the model 2c: study the trees up to 50 leaves. Present a graph of the dependence of deviances for the training and the validation data on the number of leaves and interpret this graph. Report the optimal amount of leaves and which variables seem to be most important for decision making in this tree. Interpret the information provided by the tree structure (not everything but most important findings). Estimate the confusion matrix and misclassification rate for the test data and comment whether the model has a good predictive power.
4. Perform a decision tree classification of the test data with the following loss matrix,

$$L = {}_{Observed} \begin{array}{cc} & Predicted \\ \begin{array}{c} yes \\ no \end{array} & \begin{pmatrix} 0 & 5 \\ 1 & 0 \end{pmatrix} \end{array}$$

    and report the confusion matrix for the test data. Compare the results with the results from step 3 and discuss how the rates has changed and why.

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?

Assignment 3. Principal components for crime level analysis

The data file **communities.csv** contains the results of studies of the crime level in the united states based on various characteristics of the given location. The main variable that is studied is ViolentCrimesPerPop which represents the total number of violent crimes per 100K population. The meaning of other variables can be found at:
https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

1. Scale all variables except of ViolentCrimesPerPop and implement PCA by using function eigen(). Report how many features are needed to obtain at least 95% of variance in the data. What is the proportion of variation explained by each of the first two principal components?
2. Repeat PCA analysis by using princomp() function and make the score plot of the first principle component. Do many features have a notable contribution to this component? Report which 5 features contribute mostly (by the absolute value) to the first principle component. Comment whether these features have anything in common and whether they may have a logical relationship to the crime level. Also provide a plot of the PC scores in the coordinates (PC1, PC2) in which the color of the points is given by ViolentCrimesPerPop. Analyse this plot (hint: use **ggplot2** package).
3. Assume a second order polynomial regression model in which ViolentCrimesPerPop is target and PC1 is the feature. Compute this model using lm() function (hint: use poly() function within the *formula*), make a scatterplot of the target versus the feature and present also the predicted values in this plot. Can the target be well explained by this feature? Does the model seem to capture the connection between the target and the feature?
4. Use parametric bootstrap to estimate the confidence and prediction bands from the model from step 3 and add these bands into the plot from step 3. What can be concluded by looking at a) confidence intervals b) prediction intervals?

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

Submission procedure

First read 'Course Information.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 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 *Sumissions* and opens item *Password X*. 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
    - 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 *Submission* →*Password X* item, 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.