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

Computer lab 1

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.
- Some exercises that require implementations of methods are provided with a skeleton of the code (see LISAM), you are free to use this code or implement your own code from scratch.
- **Include all your codes as an appendix into your report.**
- 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. Spam classification with nearest neighbors

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

The data file **spambase.xlsx** contains information about the frequency of various words, characters etc for a total of 2740 e-mails. Furthermore, these e-mails have been manually classified as spams (spam = 1) or regular e-mails (spam = 0).

1. Import the data into R and divide it into training and test sets (50%/50%) by using the following code:

```
n=dim(data)[1]
set.seed(12345)
id=sample(1:n, floor(n*0.5))
train=data[id,]
test=data[-id,]
```

2. Use logistic regression (functions *glm(), predict()*) to classify the training and test data by the classification principle

$$\hat{Y} = 1 \; if \; p(Y = 1|X) > 0.5, otherwise \; \hat{Y} = 0$$

and report the confusion matrices (use table()) and the misclassification rates for training and test data. Analyse the obtained results.

3. Use logistic regression to classify the test data by the classification principle

$$\hat{Y} = 1 \; if \; p(Y = 1|X) > 0.9, otherwise \; \hat{Y} = 0$$

and report the confusion matrices (use table()) and the misclassification rates for training and test data. Compare the results. What effect did the new rule have?

4. Use standard classifier *kknn()* with *K=30* from package **kknn**, report the the misclassification rates for the training and test data and compare the results with step 2.

5. Repeat step 4 for *K=1* and compare the results with step 4. What effect does the decrease of K lead to and why?

Assignment 2. Inference about lifetime of machines

- **To be solved by TDDE01 students**

The data file **machines.xlsx** contains information about the lifetime of certain machines, and the company is interested to know more about the underlying process in order to determine the warranty time. The variable is following:

- Length: shows lifetime of a machine

1. Import the data to R.
2. Assume the probability model $p(x|\theta) = \theta e^{-\theta x}$ for $\boldsymbol{x}$=*Length* in which observations are independent and identically distributed. What is the distribution type of *x*? Write a function that computes the log-likelihood $\log p(\boldsymbol{x}|\theta)$ for a given θ and a given data vector $\boldsymbol{x}$. Plot the curve showing the dependence of log-likelihood on θ where the entire data is used for fitting. What is the maximum likelihood value of θ according to the plot?
3. Repeat step 2 but use only 6 first observations from the data, and put the two log-likelihood curves (from step 2 and 3) in the same plot. What can you say about reliability of the maximum likelihood solution in each case?
4. Assume now a Bayesian model with $p(x|\theta) = \theta e^{-\theta x}$ and a prior $p(\theta) = \lambda e^{-\lambda\theta}, \lambda = 10$. Write a function computing $\mathrm{l}(\theta) = \log\big(p(\boldsymbol{x}|\theta)p(\theta)\big)$. What kind of measure is actually computed by this function? Plot the curve showing the dependence of $l(\theta)$ on θ computed using the entire data and overlay it with a plot from step 2. Find an optimal θ and compare your result with the previous findings.
5. Use θ value found in step 2 and generate 50 new observations from $p(x|\theta) = \theta e^{-\theta x}$ (use standard random number generators). Create the histograms of the original and the new data and make conclusions.

Assignment 3. Feature selection by cross-validation in a linear model.

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

1. Implement an R function that performs feature selection (best subset selection) in linear regression by using k-fold cross-validation without using any specialized function like lm() **(use only basic R functions**). Your function should depend on:
    - X: matrix containing X measurements.
    - Y: vector containing Y measurements

- Nfolds: number of folds in the cross-validation.

You may assume in your code that matrix X has 5 columns. The function should plot the CV scores computed for various feature subsets against the number of features, and it should also return the optimal subset of features and the corresponding cross-validation (CV) score. Before splitting into folds, the data should be permuted, and the seed 12345 should be used for that purpose.

2. Test your function on data set **swiss** available in the standard R repository:
   - Fertility should be Y
   - All other variables should be X
   - Nfolds should be 5

Report the resulting plot and interpret it. Report the optimal subset of features and comment whether it is reasonable that these specific features have largest impact on the target.

Assignment 4. Linear regression and regularization

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

The Excel file **tecator.xlsx** contains the results of study aimed to investigate whether a near infrared absorbance spectrum can be used to predict the fat content of samples of meat. For each meat sample the data consists of a 100 channel spectrum of absorbance records and the levels of moisture (water), fat and protein. The absorbance is -log10 of the transmittance measured by the spectrometer. The moisture, fat and protein are determined by analytic chemistry.

1. Import data to R and create a plot of Moisture versus Protein. Do you think that these data are described well by a linear model?
2. Consider model M_i in which Moisture is normally distributed, and the expected Moisture is a polynomial function of Protein including the polynomial terms up to power i (i.e M_1 is a linear model, M_2 is a quadratic model and so on). Report a probabilistic model that describes M_i. Why is it appropriate to use MSE criterion when fitting this model to a training data?
3. Divide the data into training and validation sets(50%/50%) and fit models $M_i, i = 1 \ldots 6$. For each model, record the training and the validation MSE and present a plot showing how training and validation MSE depend on i (write some R code to make this plot). Which model is best according to the plot? How do the MSE values change and why? Interpret this picture in terms of bias-variance tradeoff.

   Use the entire data set in the following computations:

4. Perform variable selection of a linear model in which *Fat* is response and *Channel1-Channel100* are predictors by using stepAIC. Comment on how many variables were selected.
5. Fit a Ridge regression model with the same predictor and response variables. Present a plot showing how model coefficients depend on the log of the penalty factor λ and report how the coefficients change with λ.
6. Repeat step 5 but fit LASSO instead of the Ridge regression and compare the plots from steps 5 and 6. Conclusions?
7. Use cross-validation to find the optimal LASSO model (make sure that case $\lambda = 0$ is also considered by the procedure) , report the optimal λ and how many variables were chosen by the model and make conclusions. Present also a plot showing the dependence of the CV score and comment how the CV score changes with λ.
8. Compare the results from steps 4 and 7.

Special task 1 (optional, individual)

Use the training and test data from assignment 1. The existing packages for nearest neighbor classification do not allow for using a distance measure which is often used to compare two text documents. This measure is based on a so called cosine similarity, which for two vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as:

$$\mathrm{c}(X,Y) = \frac{X^T Y}{\sqrt{\sum_i X_i^2}\sqrt{\sum_i Y_i^2}}$$

The corresponding distance is then defined as $d(X,Y) = 1 - \mathrm{c}(X,Y)$

1. Implement from scratch the K-nearest neighbors method which is based on distance function $d(X,Y)$ (**use only basic R functions).** Your code should be presented as a function *knearest*(*data*, *K*, *newdata*) that uses *data* as training data and then returns the predicted class probabilities for *newdata* by using K-nearest neighbor approach.
    - **Note**: R implementations can be very slow if inner loops are used. In order to efficiently compute $d(X,Y)$ for several observations *X* and *Y* represented by rows of matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, you may do the following:
        i. Compute $\widehat{\boldsymbol{X}}$ by dividing each row X_i of matrix $\boldsymbol{X}$ by $\sqrt{\sum_j X_{ij}^2}$ (use function rowSums())
        ii. Compute $\widehat{\boldsymbol{Y}}$ by dividing each row Y_i of matrix $\boldsymbol{Y}$ by $\sqrt{\sum_j Y_{ij}^2}$ (use function rowSums())
        iii. Compute matrix $C = \|c(X_i, Y_j)\|$ as $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{Y}}^{\boldsymbol{T}}$
        iv. Compute distance matrix $D = 1 - C$
2. Computer training and test misclassification errors for *K=30* and compare these with errors obtained in assignment 1, step 4.

Special task 2 (optional, individual)

Data set “cars” describes speeds of some cars in 1920. Type ?cars in R for more information. Implement k-nearest neighbor density estimation (**use only basic R functions, no specialized packages**) for the variable 'speed' from cars data set. Use k=6 and present the histogram of the ‘speed’ overlaid by the density curve computed by your density estimation algorithm. Conclusions?

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.