732A95 Introduction to Machine Learning/ 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.
- **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

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). Your task is to develop a K-nearest neighbor model that can be used as a spam filter.

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,]
```

It is required to fit a K-nearest neighbor classifier to the training data. However, 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 X and 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 $d(X,Y) = 1 - \mathrm{c}(X,Y)$

2. 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 $D = 1 - C$

3. Classify the training and test data by using $K=5$ and the classification principle
$$\hat{Y} = 1 \; if \; p(Y = 1|X) > 0.5, otherwise \; \hat{Y} = 0$$
and report the confusion matrix (use table()) and the misclassification rate for training and test data
4. Repeat step 3 with $K=1$ and compare the results.
5. Use standard classifier *kknn()* with $K=5$ from package **kknn**, report the confusion matrix and the misclassification rate for test data and compare the results with steps 3 and 4.
6. Use *knearest()* and *kknn()* functions with $K=5$ and classify the test data by using the following principle:
$$\hat{Y} = 1 \; if \; p(Y = 1|X) > \pi, otherwise \; \hat{Y} = 0$$
where $\pi = 0.05, 0.1, 0.15, \ldots 0.9, 0.95$. Compute sensitivity and the specificity values for the two methods and plot the corresponding ROC curves. Conclusion?

Assignment 2. Inference about lifetime of machines

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 **yo**u 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, 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.

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.