Examination

Linköping University, Department of Computer and Information Science, Statistics

Course code and name	TDDE01 Machine Learning
Date and time	2023-08-23, 08.00-13.00
Assisting teacher	Oleg Sysoev
Allowed aids	PDF of the course book + your help file (if submitted to LISAM in due time)
Grades:	
	5=18-20 points
	4=14-17 points
	3=10-13 points
	U=0-9 points

Provide a detailed report that includes plots, conclusions and interpretations. Give motivated answers to the questions. If an answer is not motivated, the points are reduced. Provide all necessary codes in the appendix.

Note: seed 12345 should be used in all codes that assumes randomness unless stated otherwise!

To start work in RStudio, type this in the Terminal application:

```
module add courses/TDDE01
rstudio
```

To submit your report:

1. Create one file (DOC, DOCX, ODT, PDF)
2. Use Exam Client to submit, and choose Assignment 1 in the drop box
3. Attach your report
4. Submit.
5. “Request Received” status implies that your report is successfully submitted.

Assignment 1 (10p)

The data file **glass.csv** contains information about the chemical components of two different glass types. The Type of glass is represented by variable Class coded by “1” and “-1”.

1. Use chemical components only to run PCA analysis on scaled data. How much variation is explained by the first two components? If we shall treat this operation as a data compression task where we keep only first two principal components but might want to restore the data in the original feature scale, compare how many numbers need to be stored in the original data, and how many numbers we need to keep in the compressed data. **(3p)**
2. Report equations of the first two principal components in terms of centered original features. Report also equations of original features in terms of the first two principal components. Finally, comment on which variables contribute mostly to each of the two principal components, respectively. **(3p)**
3. Scale the original data (except of Class) and divide it into training and test (50/50). Assume that we want to predict Class from all available chemical components by the following model: $y = sign(f(x)), f(x) = w^T x$, where x are all chemical variables, w is parameter vector and y is Class variable. Write a function that depends on feature matrix x, target vector y, parameter vector w and computes the cost function for this model given the following loss function: $L(y, f(x)) = \exp(-yf(x))$. Implement a code that optimizes training error (cost) through this cost function by doing early stopping with BFGS optimizer (start with $w_0 = 0$). Report a graph of training and validation errors dependence on iteration number, choose the optimal model and report confusion matrix for test data based on this model. **(4p)**

Assignment 2 (10p)

SUPPORT VECTOR MACHINES – 2 POINTS

The code below trains a support vector machine (SVM) for classification. The problem consists of two continuous inputs, one binary target and 1000 training points. Thus, the problem may seem rather easy. However, the SVM does not perform great. Explain why.

```
library(kernlab)

# Create training data (x1, x2: predictors, x3: target)

x1 <- sample(0:1,1000,replace = TRUE)

x2 <- sample(0:1,1000,replace = TRUE)

x3 <- as.numeric(xor(x1,x2))

foo <- runif(1000,min = -0.2,max = 0.2)

x1 <- x1 + foo

foo <- runif(1000,min = -0.2,max = 0.2)
```

```
x2 <- x2 + foo

# Visualize training data.

plot(cbind(x1,x2),type = "n")

text(cbind(x1,x2),labels = x3)

# Learn SVM and check training error.

foo <- ksvm(cbind(x1,x2),x3,kernel = "vanilladot",type = 'C-svc')

foo

# Visualize predictions for training data.

prex3 <- predict(foo,cbind(x1,x2))

plot(cbind(x1,x2),type = "n")

text(cbind(x1,x2),labels = prex3)
```

SUPPORT VECTOR MACHINES – 3 POINTS

You are asked to use the function **ksvm** from the R package **kernlab** to learn a support vector machine (SVM) for classifying the **spam** dataset that is included with the package. Consider the radial basis function kernel (also known as Gaussian) with a width of 0.05. For the C parameter, consider values 0.5, 1 and 5. This implies that you have to consider three models.

- Perform model selection, i.e. select the most promising of the three models (use any method of your choice).
- Estimate the generalization error of the SVM selected above (use any method of your choice).
- Produce the SVM that will be returned to the user, i.e. show the code.

KERNEL METHODS – 5 POINTS

One of the labs in the course consisted in implementing a kernel method to predict the hourly temperatures for a date and place in Sweden. Modify your lab solution or implement it anew to classify the **spam** dataset used in the previous exercise. Use the Gaussian kernel and show results for different kernel widths, e.g. use 2/3 of the data for learning and 1/3 for testing. Use only the first 48 attributes in the dataset, in addition to the last one which is the class label. Assume that the class label is a continuous random variable (so that you actually solve a regression problem, like you did in the lab).