Examination

Linköping University, Department of Computer and Information Science, Statistics

Course code and name	TDDE01 Machine Learning
Date and time	2024-03-15, 14.00-19.00
Assisting teacher	Oleg Sysoev
Allowed aids	Contents of “help_materials” folder + 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 (allowed formats: 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)

Data file **women.csv** contains clinical records about female cancer patients and variable Death that represent the survival status (0=survived, 1=not survived)

1. Scale all variables except of Death and then use these scaled features to perform PCA. Report how much of the total variation is captured by the first two principal components. Report an equation of the first principal component in terms of the scaled original variables. Finally, report a scatter plot of the data in the coordinate system (PC1, PC2) where observations are colored by Death and comment which of these PCs is best in discriminating between the survival groups. **(3p)**
2. Split the original dataset without Death variable into training and test sets (50/50) and then consider the training sets that have first 100 observations, first 200 observations, …, first 2400 observations, 2500 observations. For each of these data sets, estimate a) a decision tree with 10 leaves and b) K-nearest neighbor with K=10 where Cholestrol is used as target and all remaining variables as features. Compute MSE for each of the training and test sets and each of the two model types and then plot a dependence of training and test errors on the size of the training set a) for decision tree models b) for K-nearest neighbor models. Comment on the trends you observe in the decision tree plot and the theoretical reasons behind these trends. By comparing the two plots, report which of the two models can be considered as more complex one for these data and why. **(4p)**
3. Split original data into training and test sets (50/50) and estimate a Ridge classification model where Death is target and all other variables are features so that optimal penalty parameter is selected by the cross-validation from the training data. Report the optimal penalty factor, and confusion matrix for the test dataset. Use also the following cost matrix to update the confusion matrix for the test data:

$$\begin{array}{c} \\ True\ "0" \\ True\ "1" \end{array} \begin{array}{c} \begin{array}{cc} Pred\ "0" & Pred\ "1" \end{array} \\ \begin{pmatrix} 0 & 1 \\ 10 & 0 \end{pmatrix} \end{array}$$

and comment on what kind of changes in the confusion matrix you observed and why they happened.
(3p)

Assignment 2 (10p)

Read the exercise entirely before starting. The support vector machines (SVMs) that you saw in the course assume that the learning data is available when starting the training process. This is known as batch learning. As opposed, online learning assumes that the data points arrive one by one as a (possibly endless) stream of data, and that we cannot wait until the end of the stream to start the training process. Instead, we have to update our predictor with each new point's arrival. This is useful in weather prediction, spam filtering, financial applications, etc.

Your task is to **implement the so-called budget online SVM classifier**. You have the pseudocode below (|S| means the cardinality of the set S, i.e. the number of elements in S).

Budget online SVM (input: β and M)

1 $\mathcal{S} = \varnothing$
2 $b = 0$
3 Receive a random example $(\boldsymbol{x}_i, t_i)$
4 Compute $y(\boldsymbol{x}_i) = \sum_{m \in \mathcal{S}} a_m t_m k(\boldsymbol{x}_i, \boldsymbol{x}_m) + b$
5 If $t_i y(\boldsymbol{x}_i) \leq \beta$ then
6 $\quad \mathcal{S} = \mathcal{S} \cup \{i\}$
7 $\quad a_i = 1$
8 $\quad$ If $|\mathcal{S}| > M$ then $\mathcal{S} = \mathcal{S} \setminus \{\arg\max_{m \in \mathcal{S}} t_m (y(\boldsymbol{x}_m) - a_m t_m k(\boldsymbol{x}_m, \boldsymbol{x}_m))\}$
9 Go to step 3

You are requested to use the template below.

```
set.seed(1234567890)

spam <- read.csv2("spambase.csv")

ind <- sample(1:nrow(spam))

spam <- spam[ind,c(1:48,58)]

h <- 1

beta <- # Your value here

M <- # Your value here

N <- 500 # number of training points

gaussian_k <- function(x, h) { # Gaussian kernel

# Your code here

}

SVM <- function(sv,i) { # SVM on point i with support vectors sv

# Your code here

# Note that the labels in spambase.csv are 0/1 and SVMs need -1/+1. Then, use 2*label-1

# to convert from 0/1 to -1/+1

# Do not include the labels when computing the Euclidean distance between the point i

# and each of the support vectors. This is the distance to use in the kernel function
```

```
# You can use dist() to compute the Euclidean distance

}

errors <- 1

errorrate <- vector(length = N)

errorrate[1] <- 1

sv <- c(1)

for(i in 2:N) {

# Your code here

# Assume that you receive the i-th data point in spam

}

plot(errorrate[seq(from=1, to=N, by=10)], ylim=c(0.2,0.4), type="o")

length(sv)

errorrate[N]
```

Run your code on the spambase.csv file for the (M, beta) values (500,0) and (500,-0.05). **Comment your results**. Finally, **answer the following two questions**. First, explain why the removal criterion in step 8 of the budget online SVM makes sense. Second, name two similarities between the budget online SVM and the (batch) SVMs that you have seen in the course.

The exercise will be graded as follows: Implementation 4 p, results 3 p, and questions 3 p.