Examination

Linköping University, Department of Computer and Information Science, Statistics

Course code and name	TDDE01 Machine Learning
Date and time	2020-08-26, 8.00-13.00
Assisting teacher	Oleg Sysoev
Allowed aids	"Pattern recognition and Machine Learning" by Bishop and "The Elements of Statistical learning" by Hastie
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.

- FIRST READ THE FILE **732A99_TDDE01_EXAM_ REGULATIONS.PDF** UNLESS YOU HAVE ALREADY DONE THAT IN ADVANCE
- **Use seed 12345 when randomness is present unless specified otherwise.**
- **Specify RNGversion("3.5.2") in the code**

Assignment 1 (10p)

Part 1

File ***jobs.csv*** contains information about the percentage employed in different industries in Europe countries during 1979 and some geographical information. Variables are the following:

- Country: Name of country
- Block: Eastern Europe, Western Europe
- Agr: Percentage employed in agriculture
- Min: Percentage employed in mining

- Man: Percentage employed in manufacturing
- PS: Percentage employed in power supply industries
- Con: Percentage employed in construction
- SI: Percentage employed in service industries
- Fin: Percentage employed in finance
- SPS: Percentage employed in social and personal services
- TC: Percentage employed in transport and communications

1. Compute a classification tree in which Block is the target variable and all other numerical variables are features in such a way that the length of the tree is selected by cross-validation. Provide the resulting tree and interpret it. **(2p)**
2. Assume the following loss matrix $\begin{matrix} True \\ \\ East \\ West \end{matrix} \begin{pmatrix} & Predicted & \\ & East & West \\ East & 0 & 5 \\ West & 1 & 0 \end{pmatrix}$ and revise the classification made by the trees. How have the tree labels changed and why? **(2p)**

Part 2

The data file **CYGOB1.csv** contains information about the parameters of Star Cluster:

- "Logst" shows log surface temperature of the star
- "Logli" shows log light intensity of the star

1. Standardize each column of the data and then split your data into train (50%) and test (50%). Assuming that Logst is a feature (X) and Logli is a target variable (Y) in a Ridge regression, why is it conventional to standardize data before doing the regression? (**1p)**
2. Assuming that Y is Laplace distributed, i.e. $Y = w_1 X + \epsilon,\ p(\epsilon) = \frac{1}{2} e^{-|\epsilon|}$, implement an R function that depends on parameters w and computes the log-likelihood for the **training** data. By using this function create a new one which would depend on parameters w and λ and compute penalized a log-likelihood by using Ridge penalty factor λ. Finally, create and interpret two plots: one that shows a dependence of the penalized log-likelihood on w for $\lambda = 1$, and another plot of the same kind for $\lambda = 10$. Which w parameter value is optimal for each λ ? **(3p)**
3. Use the test set to predict Y values by the models corresponding to the two selected λ values. Which model appears to be more optimal and why? **(2p)**

Assignment 2 (10p)

SUPPORT VECTOR MACHINES – 5 POINTS

The code below trains an 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)
```

NEURAL NETWORKS – 5 POINTS

The code below trains a NN to predict the sine function in the interval [0, 10]. You are provided with three custom activation functions (myAct, myAct2 and myAct3). You are asked to run the code below with the three of them. Do not change any other parameter between runs. Explain the results that you obtain (even if the result is an error or warning).

```
library(neuralnet)

Var <- runif(50, 0, 10)

tr <- data.frame(Var, Sin=sin(Var))

myAct <- function(x) x
```

```
myAct2 <- function(x) max(0,x)

myAct3 <- function(x) log(1+exp(x))

nn <- neuralnet(formula = Sin ~ Var, data = tr, hidden = c(2,2), act.fct = myAct)

plot(tr[,1],predict(nn,tr), col="blue", cex=3)

points(tr, col = "red", cex=3)
```