Examination

Linköping University, Department of Computer and Information Science, Statistics

Course code and name	732A95 Introduction to Machine Learning
Date and time	2018-04-06, 08.00-13.00
Assisting teacher	Oleg Sysoev
Allowed aids	“Pattern recognition and Machine Learning” by Bishop and “The Elements of Statistical learning” by Hastie
Grades:	A=19-20 points
	B=16-18 points
	C=11-15 points
	D=9-10 points
	E=7-8 points
	F=0-6 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. Use seed 12345 when randomness is present unless specified otherwise.

Assignment 1 (10p)

Data file *iris* present in a default R installation describes measurements of 3 types of flowers. Type *?iris* to read more about this data set. Divide the data into training, validation and test set (1/3, 1/3, 1/3):

```
n=dim(iris)[1]
set.seed(12345)
id=sample(1:n, floor(n*1/3))
train=iris[id,]
id1=setdiff(1:n, id)
set.seed(12345)
id2=sample(id1, floor(n*1/3))
valid=iris[id2,]
id3=setdiff(id1,id2)
test=iris[id3,]
```

1. Perform a series of LASSO regressions by using training data and computing the misclassification errors for the validation data for various values of the penalty factor $\lambda = 0, 0.1, \dots, 0.9, 1$. Produce a plot showing the dependence of the training and test errors on the penalty parameter and choose the optimal value of the penalty parameter. Comment how the training error behaves when λ increases and why it happens in this way. Which features are selected by the optimal LASSO model? Finally, compute the test error, compare it with the training and validation errors and make appropriate conclusions. **(3p)**
2. Fit a Naïve Bayes classification to the combined training and validation data and estimate the prediction error for the test data. Report the confusion matrix for the training data, training and test errors. Present also a scatterplot showing Sepal Length versus Sepal Width in which the observations are colored by the value of Species. Does Naïve Bayes assumption seem to be fulfilled according to this plot? **(2p)**
3. Use another Naïve Bayes model with the following loss matrix: $C = \begin{array}{cc} \begin{matrix} 1 & 2 & 3 \end{matrix} & \\ \begin{pmatrix} 0 & 2 & 1 \\ 6 & 0 & 100 \\ 6 & 2 & 0 \end{pmatrix} & \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} \end{array}$ and compute the new confusion matrix for the training data, where “1”= “Setosa”, “2”=”Versicolor” and “3”=”Virginica”. How do the numbers in the confusion matrix change in the second Naïve Bayes model compared to the first one and why? **(3p)**
4. Compute a PCR regression in which Sepal Width is the target variable and Sepal length, Petal Length and Petal Width are features, use all data to fit this model. Report how many components are needed to explain more than 90% of the variance in the target and how many principal components are needed to explain more than 90% of the variance in the feature space? Why are these numbers normally quite different? **(2p)**

Assignment 2 (10p)

Support Vector Machines (4 p)

In this assignment, you are asked to use the R package `kernlab` to learn SVMs 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 5, 25 och 50.

(2p) Estimate the error for the three values of *C*. Use cross-validation with 2 folds. Hint: Use the argument `cross=2` when calling the function `ksvm`. Use the function `cross()` to print out the error estimate. Use `set.seed(1234567890)`.

(2p) In the previous question, the error estimate may not be mono- tone with respect to the value of *C*. Explain why this happens.

Neural Networks (3 p)

In this assignment, you are asked to use the R package `neuralnet` to train a NN to learn the trigonometric sine function. To produce the learning data, sample 50 points uniformly at random in the interval $[\mathbf{0}, \mathbf{10}]$ and, then, apply the sine function to each point.

Your task is to estimate the mean squared error of a NN with 2 hidden layers of 3 units per layer for the regression task described above. Use cross-validation with 2 folds. For the training, initialize the weights of the NN to random values in the interval $[-1, 1]$. Stop the training when the partial derivatives of the error function are below a threshold value of 0.001.

Hint: Check the argument `threshold` in the documentation. Use the function `compute()` to compute the output of the trained NN for a given input vector. Use the default values for the arguments not mentioned here. Feel free to use the following template.

```
library(neuralnet)
set.seed(1234567890)


Var <- runif(50, 0, 10)
tr <- data.frame(Var, Sin=sin(Var))
tr1 <- tr[1:25,] # Fold 1
tr2 <- tr[26:50,] # Fold 2
```

Ensemble Methods (3 p)

(1p) Interpret the plot resulting from the code below.

```
library(mboost)
bf <- read.csv2("bodyfatregression.csv")


set.seed(1234567890)
m <- blackboost(Bodyfat_percent˜Waist_cm+Weight_kg, data=bf)
mstop(m)
cvf <- cv(model.weights(m),type="kfold") cvm
<- cvrisk(m, folds=cvf, grid=1:100)
plot(cvm)
mstop(cvm)
```

(2p) Estimate the mean squared error of the boosting regression tree in the question above. Use 1/2 of the data for training and 1/2 as hold-out test data. Let the boosting procedure choose the appropriate number of trees by adding the parameter `control=boost control(mstop=mstop(cvm))` to the function `blackboost()`.

Hint: Use the function `predict()` to compute the output of the boosting tree.