Examination

Linköping University, Department of Computer and Information Science, Statistics

Course code and name	732A95 Introduction to Machine Learning
Date and time	2017-04-18, 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 (6p)

The data file **australian-crabs.csv** contains measurements of various crabs, such as Frontal lobe (FL), Rear width (RW), Carapace Length (CL), Carapace Width (CW), Body depth (BD) as well as indication of which kind of crab it is (Species).

1. Plot the dependence of CW versus BD where the points are colored by Species. Are CW and BD good predictors of the Species? **(1p)**
2. Create a Naïve Bayes classifier model with Species as target and CW and BD as predictors. Present the confusion matrix and comment on the quality of the classification. Based on the assumptions of the Naïve Bayes, explain why this model is not appropriate for these data **(2p)**
3. Fit the logistic regression now with Species as target and CW and BD as predictors and present the equation of the decision boundary. Plot the classified data and the decision boundary and comment on the quality of the classification **(2p)**

4. Scale variables CW and BD and perform principal component analysis with these two variables. Present the proportion of variation explained by PC1 and PC2 and based on results from step 1 explain why the first principal component contains so much variation. **(1p)**

Assignment 2 (4p)

File **bank.csv** shows the number of customers (Visits) that arrived to a bank during various time slots (Time) between 9.00 and 12.00.

1. Fit a generalized linear model in which response is Poisson distributed, and the canonical link (log) is used for regression. Report the probabilistic expression for the fitted model (how the target is distributed based on the feature) **(1p)**
2. Compute a prediction band for the values of Time=12,12.05,12.1,…,13.0 by using the model from step 1 and the parametric bootstrap with B=1000. Plot the original data values and the prediction band into one figure and comment whether the band seems to give a correct forecasting. How many customers (report a range) should the bank expect at 13.00? **(3p)**

Assignment 3 (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 1, 10 and 100.

(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 a single hidden layer of 10 units 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 $[-\mathbf{1}, \mathbf{1}]$. 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 2/3 of the data for training and 1/3 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.