Examination

Linköping University, Department of Computer and Information Science, Statistics

Course code and name	TDDE01 Machine Learning
Date and time	2020-08-25, 08.00-13.00
Assisting teacher	Oleg Sysoev
Allowed aids	See "732A99_TDDE01_exam_regulations.PDF"
Grades:	5=18-20 points plus passed oral defense
	U=18-20 points plus failed oral defense
	4= 18-20 points without oral defense
	4=14-17 points with or without oral defense
	3=10-13 points with or without oral defense
	U=0-9 points with or without oral defense

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!

Assignment 1 (5p)

Part 1 (5p)

1. Implement a function that for given numbers n and p generates a data set with n observations and p features and one target where each observation (row) is generated according to the following:

$$X_i \sim U[0,1], i = 1, \dots, p$$

$$Y = \begin{cases} 1 \; if \; \sum_{i=1}^{p} X_i < 0.5p + \epsilon, & where \; \epsilon \sim N(0,0.1) \\ 0, otherwise \end{cases}$$

By using this function, implement a loop for $p = 2, \dots, 100$ where you generate training data with 100 observations, test data with 200 observations (specify set.seed before the loop) and then fit a 3-nearest neighbor classifier to these data in the loop. Present a dependence of the test accuracy on p as a scatter plot and comment on the trend you observe and which phenomenon this trend demonstrates. **(3p)**

2. Use same function as in step 1 to simulate training data with 100 points and test data with 1000 points and dimension $p = 3$ (use set.seed once before simulating the training data) and then fit a K-nearest neighbor classifier to these data where $K = 1, \dots ,99$. Provide a plot showing dependence of the test accuracy on K. What kind of trend can you see in this plot and what phenomenon is demonstrated by this trend? **(2p)**

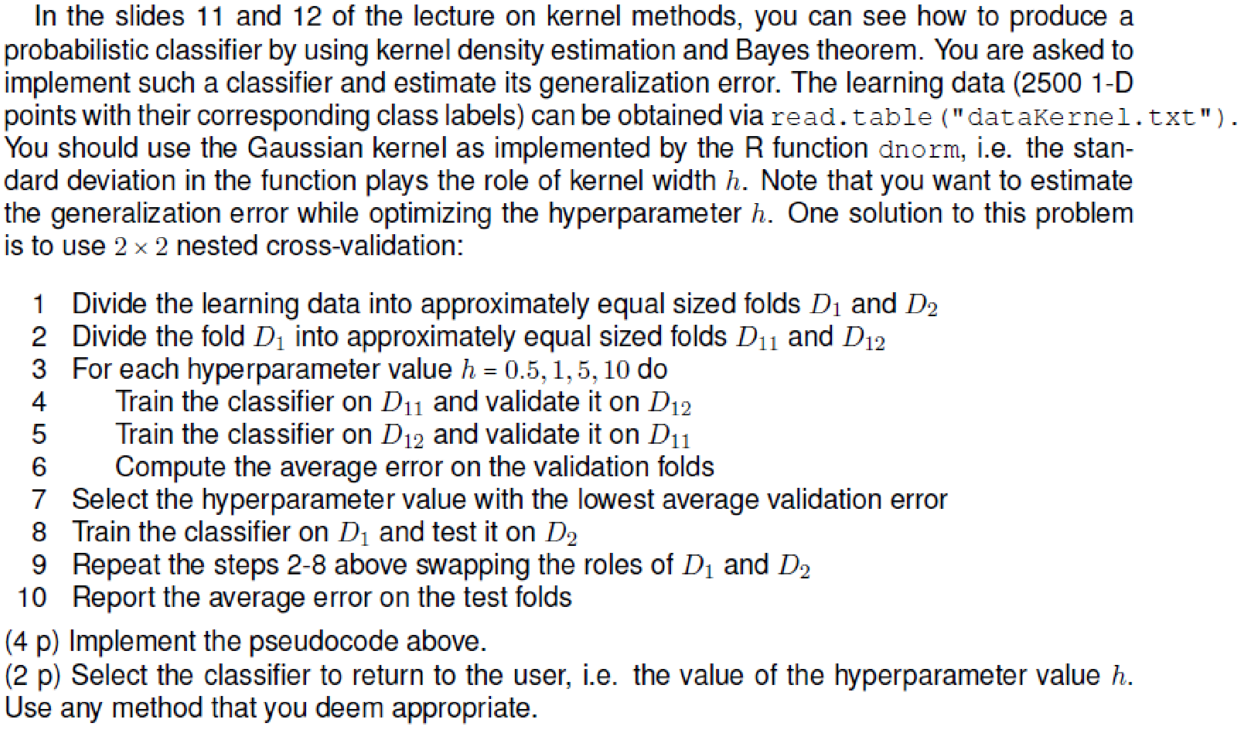
Part 2 (5p)

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. Divide these data into training and test data (50/50) and compute a classification tree which predicts Death variable based on the remaining variables; do this by first growing the full tree with minimum deviance parameter 0.003 and then computing the optimal tree by the cross-validation. Report a) the plot showing the dependence of the cross-validation error on the tree size b) how many features are selected by the tree c) confusion matrix for the test data and d) misclassification error. **(3p)**
2. Use the same data partitioning to fit a logistic regression model and compute the confusion matrix for the test data, the test misclassification error and how many features are chosen by the model. Which of these two models – logistic or tree - would you prioritize here and why? What kind of problem is demonstrated by these data when computing both models? **(2p)**

Assignment 2 (10p)

Part 1: Kernel methods (6p)



In the slides 11 and 12 of the lecture on kernel methods, you can see how to produce a probabilistic classifier by using kernel density estimation and Bayes theorem. You are asked to implement such a classifier and estimate its generalization error. The learning data (2500 1-D points with their corresponding class labels) can be obtained via `read.table("dataKernel.txt")`. You should use the Gaussian kernel as implemented by the R function `dnorm`, i.e. the standard deviation in the function plays the role of kernel width h. Note that you want to estimate the generalization error while optimizing the hyperparameter h. One solution to this problem is to use 2×2 nested cross-validation:

```
 1  Divide the learning data into approximately equal sized folds $D_1$ and $D_2$
 2  Divide the fold $D_1$ into approximately equal sized folds $D_{11}$ and $D_{12}$
 3  For each hyperparameter value $h = 0.5, 1, 5, 10$ do
 4      Train the classifier on $D_{11}$ and validate it on $D_{12}$
 5      Train the classifier on $D_{12}$ and validate it on $D_{11}$
 6      Compute the average error on the validation folds
 7  Select the hyperparameter value with the lowest average validation error
 8  Train the classifier on $D_1$ and test it on $D_2$
 9  Repeat the steps 2-8 above swapping the roles of $D_1$ and $D_2$
10  Report the average error on the test folds
```

(4 p) Implement the pseudocode above.
(2 p) Select the classifier to return to the user, i.e. the value of the hyperparameter value h. Use any method that you deem appropriate.

Part 2: Neural networks (4p)

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 $[0, 10]$ 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 $[-1, 1]$. Stop the training when the partial derivatives of the error function are below a threshold value of 0.001. Recall that cross-validation works as follows:

1 Divide the learning data into approximately equal sized folds D_1 and D_2
2 Train the regressor on D_1 and test it on D_2
3 Train the regressor on D_2 and test it on D_1
4 Report the average error on the test folds

Hint: Check the argument `threshold` in the documentation. Use the function `predict()` 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
```