Examination

Linköping University, Department of Computer and Information Science, Statistics

Course code and name	732A52 Introduction to Machine Learning
Date and time	2015-01-18, 14.00-19.00
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.

Assignment 1 (7p)

The data file "**glass.csv**" contains information about the contents of chemicals in various samples of the glass. In the tasks below, you are assumed to investigate how the contents of Aluminium (Al) can be explained by the other chemicals. Therefore, consider Al as target variable and the remaining variables as thefeatures in the models below.

1. Partition data into training, validation and test sets (50/25/25) by using seed 12345. Use training and validation trees to fit the regression trees of different sizes and estimate the predictive error. Provide the plot showing the dependence of the training and validation errors on the tree size and comment which tree is optimal and why. Interpret this graph also in terms of bias-viariance tradeoff. **(2p)**
2. Investigate the optimal tree from step 1 and report which variables were chosen. Report also the test error (assume that the target is normally distributed). **(1p)**

3. Fit a PLS regression model in which the amount of variables are chosen by cross validation. Answer the following questions **(3p)**:
    a. How many variables are enough to explain at least 90% of variation in the feature space?
    b. How many variables are enough to explain at least 90% of variation of the target?
    c. What is the optimal amount of variables according to the cross-validation?
    d. Which variables contribute mostly to the first principle component?
    e. What is the equation of the Target in the coordinates of the principle components?
    f. What is the prediction error of the optimal PLS model for the test data?
4. Compare the test errors of the optimal tree and PLS models and answer which model has a better predictive power. Comment on why choosing the holdout principle for these data is much less reliable than using cross-validation. **(1p)**

Assignment 2 (8p)

In this assignment, you are going to analyze dataset **mtcars** available in basic R. Scale predictors *qsec* and *hp* for further analysis.

1. Plot the data in the coordinates *hp* versus *qsec* where the data are colored by *am*. Does it seem that the Linear Discriminant analysis will be able to separate these data perfectly if class priors are chosen appropriately? Do these data seem to fulfill assumptions of LDA? **(1p)**
2. Perform LDA with response *am* and predictors *qsec* and hp and
    a) Equal priors
    b) Proportional priors

    Plot the classified data and compare the results obtained by a) and b). Which method seems to classify the data better? How has the parameters of the decision boundary (intercept and slope) changed from case a) to case b) and why? **(3p)**

3. Implement kernel density estimation with Epanechnikov kernel that uses matrices X, Xtest and a scalar λ to estimate density from X and predict it at Xtest (observations in the matrices are given in the rows). Estimate the kernel density for λ =0.2 and for
    a. X is matrix with two columns: column 1 are *qsec* values such that *am=0,* column 2 are *hp* values such that *am=0,* Xtest is matrix with two columns: column 1 are all *qsec* values, column 2 are all *hp* values
    b. X is matrix with two columns: column 1 are *qsec* values such that *am=1,* column 2 are *hp* values such that *am=1,* Xtest is matrix with two columns: column 1 are all *qsec* values, column 2 are all *hp* values

    Use the estimated densities and the Bayesian rule to classify your data (assume equal class priors) and plot the classified data. Comment on the quality of fit. How does the performance of this classifier change when λ is set as a very small or very large value and why? **(4p)**

Assignment 3 (5p)

The file **wine.csv** contain 130 bottles of wine of two different types and some measurements of levels within these bottles.

1. First change the class “2” to “-1” in order to model a classification problem using the *neuralnet* package.
2. Separate the data set into 70 percent training and 30 percent test using seed 12345.
3. Fit a neural network with seed 12345, 0 hidden nodes and *tanh* as the activation function of the output layer. Present the weights and state which variable is deemed most and least important. **(1p)**
4. Compute the misclassification rate of the training and test set and interpret the results. (Note that the model predictions need to be converted to the -1, 1 classes, this is easily done by the sign-function in R.) (**1p)**
5. Fit a new neural network with the same seed as in step 3, 1 hidden node and *tanh* as the activation function of the hidden layer. The output layer should have no activation function. Repeat step 4 and compare with the model from step 3. **(2p)**
6. Reflect on what type of model the neuralnet-function in step 3 and 5 is trying to fit, specifically the error and activation functions. Given the data set and the response variable, is this type of architecture the proper one to use? Motivate your answer. **(1p)**