Examination

Linköping University, Department of Computer and Information Science, Statistics

Course code and name	TDDE01 Machine Learning
Date and time	2018-01-11, 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:	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.

Note: seed 12345 should be used in all codes that assumes randomness unless stated otherwise!

Assignment 1 (10p)

The data file **video.csv** contains characteristics of a sample of Youtube videos. Import data to R and divide it randomly (50/50) into training and test sets.

1. Perform principal component analysis using the numeric variables in the training data except of "utime" variable. Do this analysis with and without scaling of the features. How many components are necessary to explain more than 95% variation of the data in both cases? Explain why so few components are needed when scaling is not done. **(2p)**
2. Write a code that fits a principle component regression ("utime" as response and all scaled numerical variables as features) with M components to the training data and estimates the training and test errors, do this for all feasible M values. Plot dependence of the training and test errors on M and explain this plot in terms of bias-variance tradeoff. (**Hint**: prediction function for principal component regression has some peculiarities, see predict.mvr) **(2p)**
3. Use PCR model with $M = 8$ and report a fitted probabilistic model that shows the connection between the target and the principal components. **(1p)**

4. Use original data to create variable “class” that shows “mpeg” if variable “codec” is equal to “mpeg4”, and “other” for all other values of “codec”. Create a plot of “duration” versus “frames” where cases are colored by “class”. Do you think that the classes are easily separable by a linear decision boundary? **(1p)**
5. Fit a Linear Discriminant Analysis model with “class” as target and “frames” and “duration” as features to the entire dataset (scale features first). Produce the plot showing the classified data and report the training error. Explain why LDA was unable to achieve perfect (or nearly perfect) classification in this case. **(2p)**
6. Fit a decision tree model with “class” as target and “frames” and “duration” as features to the entire dataset, choose an appropriate tree size by cross-validation. Report the training error. How many leaves are there in the final tree? Explain why such a complicated tree is needed to describe such a simple decision boundary. **(2p)**

Assignment 2 (10p)

Support Vector Machines

You are asked to use the function ksvm from the R package kernlab to learn a support vector machine (SVM) 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 0.5, 1 and 5. This implies that you have to consider three models.

(2p) Perform model selection, i.e. select the most promising of the three models (use any method of your choice except cross-validation or nested cross-validation).

(1p) Estimate the generalization error of the SVM selected above (use any method of your choice except cross-validation or nested cross-validation).

(1p) Produce the SVM that will be returned to the user, i.e. show the code.

(1p) What is the purpose of the parameter C ?

Neural Networks

(3p) Train a neural network (NN) to learn the trigonometric sine function. To do so, sample 50 points uniformly at random in the interval $[0, 10]$. Apply the sine function to each point. The resulting pairs are the data available to you. Use 25 of the 50 points for training and the rest for validation. The validation set is used for early stop of the gradient descent. Consider threshold values $i/1000$ with $i = 1, \ldots, 10$. Initialize the weights of the neural network to random values in the interval $[-1, 1]$. Consider two NN architectures: A single hidden layer of 10 units, and two hidden layers with 3 units each. Choose the most appropriate NN architecture and threshold value. Motivate your choice. Feel free to reuse the code of the corresponding lab.

(1p) Estimate the generalization error of the NN selected above (use any method of your choice).

(1p) In the light of the results above, would you say that the more layers the better ? Motivate your answer.