Examination

Linköping University, Department of Computer and Information Science, Statistics

Course code and name	TDDE01 Machine Learning
Date and time	2020-03-18, 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 (10p)

The data file optdigits.csv contains information about normalized bitmaps of handwritten digits from a preprinted form from a total of 43 people. The data were first derived as 32x32 bitmaps which were then divided into nonoverlapping blocks of 4x4 and the number of on pixels are counted in each block. This has generated the resulting image of size 8x8 where each element is an integer in the range 0..16. Accordingly, each row in the data file is a sequence corresponding to 8x8 matrix, and the last element shows the actual digit from 0 to 9.

1. Make a Principal Component Analysis of this data without scaling the data and compute how many principal components are needed to explain more than 95% of the variation in the data. Make a trace plot and comment how many features have almost zero contribution to this component. Finally, keep only this amount of principal components needed to obtain 95% of

variation, discard the remaining variation, and compute the coordinates of the first observation in the original space after the PCA compression. Make a heatmap of the first observation before and after the PCA compression and make a comment whether much information was lost by the compression. **(4p)**

   a. **Hint:** Remember that PCA operates by first subtracting the variable means so when restoring to the original space these means should be added up.

2. Split the original data randomly into training and test sets (50%/50%), and fit a sequence of classification trees to the training data with settings mindev=0 and minsize=3,4,5,…,50. Estimate training and test misclassification errors for these models, and present dependences of the training and test errors on the minsize parameter into one plot. Comment how the complexity of the model changes with minsize parameter. Explain the dependence observed in terms of bias-variance tradeoff. **(3p)**
3. Create variable Y which is equal to 1 if the observed digit is "0" and zero otherwise. Fit a logistic regression model to the original data with target Y and one feature equal to the first principal component. Compute also a parametric bootstrap to estimate the confidence bands for the probability of Y=1 given the feature values. Create a plot showing the estimated probabilities of Y=1 versus the feature (from the logistic model) together with the confidence band and comment on whether the probability model estimated by the logistic regression seems to be reliable. **(3p)**

   a. **Hint**: Remember that Bernoulli distribution is equivalent to Binomial with number of trials equal to 1.

Assignment 2 (10p)

KERNEL METHODS – 4 POINTS

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. 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.

(3 p) Divide the data into training (1500 points), validation (500 points) and test (500 points). Use the training and validation data to select among $h = 0.5, 1, 5, 10$.

(1 p) Use the test data to estimate the accuracy of the model selected. Explain your code.

NEURAL NETWORKS – 6 POINTS

You are asked to implement the backpropagation algorithm for fitting the parameters of a neural network (NN) for regression. The NN has one input unit, 10 hidden units, and one output unit. Use the rectifier linear unit (ReLU) activation function, i.e. $h(a_j) = \max(a_j, 0)$. Recall that you have an example of the backpropagation algorithm in Bishop's book as well as in the course slides. Use the template in the file `templateNN.R`. Use only basic R functions. Comment your code. Run your code and report the result.

Tip 1: The derivative of ReLU is different if $a_j > 0$ or $a_j < 0$, and it is undefined at $a_j = 0$. In the latter case, the convention is to use the same derivative as when $a_j < 0$.

Tip 2: Do not forget to update the bias terms in the backward propagation step. They can be updated as the rest of the weights if you assume that all of them are associated to a dummy input $x_0 = 1$ (this is the meaning of the blue node in the slides and Bishop's book).