Examination

Linköping University, Department of Computer and Information Science, Statistics

Course code and name	TDDE01 Machine Learning
Date and time	2019-01-16, 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:	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 **Influenza.csv** contains contains the number of registered cases of influenza and mortality. The variables in the data are:

- Year, Week
- Mortality: number of mortality cases
- Influenza: number of influenza cases
- Temperature_deficit: a temperature measurement
- Influenza_lag, Temp_lag: measurements at a previous time point (t-1)
- Influenza_lag2, Temp_lag2: measurements at a time point which is two units back (t-2)

Import data to R.

1. Assume that mortality y is Poisson distributed, i.e. $p(y=Y|\lambda)=\frac{e^{-\lambda}\lambda^{Y}}{Y!}$, where $Y!=1\cdot 2\cdot \ldots \cdot Y$. Write an R code computing the minus-loglikelihood of Mortality values for a given λ (use only basic R functions, do not use implemented Poisson distribution in R). Compute the minus log-likelihood values for $\lambda = 10{,}110{,}210, \ldots, 2910$ and produce a plot showing the dependence of the minus log-likelihood on the value of λ. Define the optimal value of λ by means of visual inspection (i.e. approximately**). (2p)**
2. Scale all variables except of Mortality. Divide the data randomly (50/50) into training and test sets and fit a LASSO regression with Mortality as a Poisson distributed target and all other variables as features. Select the optimal parameters in the LASSO regression by the cross-validation and report the optimal LASSO penalization parameter and also the test MSE. Is the MSE actually the best way of measuring the error rate in this case? Report also the optimal LASSO coefficients and report which variables seem to have the biggest impact on the target. Check the value of intercept α, compute $\exp(\alpha)$ and compare it with the optimal λ in step 1. Are these quantities similar and should they be? **(3p)**
3. Fit a regression tree with Mortality as a target and all variables as a features and select the optimal size of the tree by cross-validation. Report the test MSE and compare it with the MSE of the LASSO regression in step 2. Why is it not reasonable to do variable selection by applying LASSO penalization also to the tree models? **(2p)**
4. Perform principal component analysis using all the variables in the training data except of Mortality and report how many principal components are needed to capture more than 90% of the variation in the data. Use the coordinates of the data in the principal component space as features and fit a LASSO regression with Mortality as a Poisson distributed target by cross-validation, check penalty factors $\lambda = 0, 0.1, 0.2, \ldots, 50$. Provide a plot that shows the dependence of the cross-validation error on $\log(\lambda)$. Does complexity of the model increase when λ increases? How many features are selected by the LASSO regression? Report a probabilistic model corresponding to the optimal LASSO model. **(3p)**

Assignment 2 (10p)

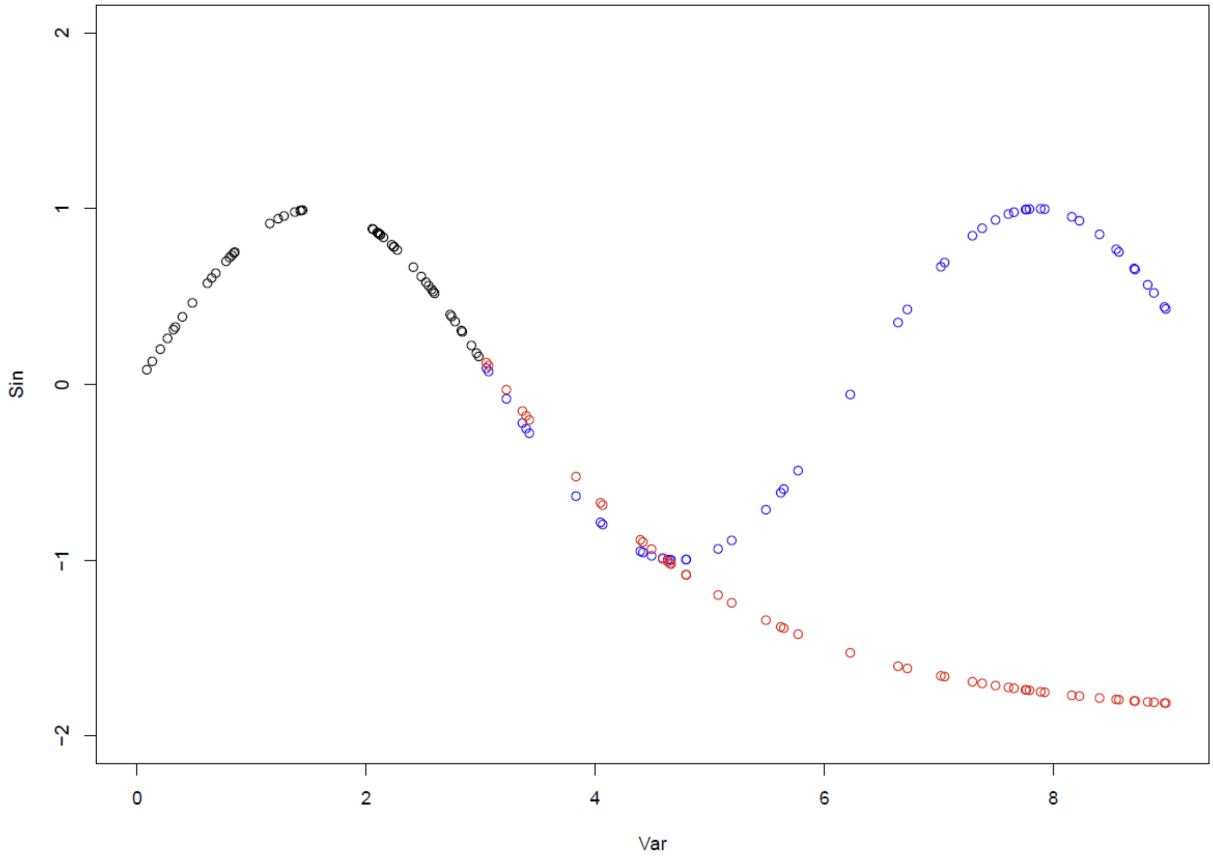
NEURAL NETWORKS - 4 POINTS

You are asked to use the function neuralnet of the R package of the same name to train a neural network (NN) to mimic the trigonometric sine function. You should run the following code to obtain the training and test data.

```
install.packages("neuralnet")
library(neuralnet)
set.seed(1234567890)
Var <- runif(50, 0, 3)
tr <- data.frame(Var, Sin=sin(Var))
Var <- runif(50, 3, 9)
te <- data.frame(Var, Sin=sin(Var))
```

(2 p) Produce the code to train the NN on the training data tr and test it on the data te. Use a single hidden layer with three units. Initialize the weights at random in the interval [-1,1]. Use the default values for the rest of parameters in the function neuralnet. You may need to use the function compute. Confirm that you get results similar to the following figure. The black dots are the training data. The blue dots are the test data. The red dots are the NN predictions for the test data.

(2 p) In the previous figure, it is not surprising the poor performance on the range [3,9] because no training point falls in that interval. However, it seems that the predictions converge to -2 as the value of Var increases. Why do they converge to that particular value ? To answer this question, you may want to look into the weights of the NN learned.



Support Vector Machines - 6 Points

You are asked to use the function ksvm from the R package kernlab to learn a support vector machine (SVM) to classify the spam dataset that is included with the package. You should use the radial basis function kernel (also known as Gaussian) with a width of 0.05.

(2 p) You should select the most appropriate value for the C parameter, i.e. you should perform model selection. For this task, you can use any method that you deem appropriate.

(1 p) In the previous question, you may have obtained an error message “no support vectors found” for C = 0. Can you give a plausible explanation for this error ?

(1 p) Estimate the generalization error of the SVM with the C value selected above. Use any method of your choice.

(2 p) Once a SVM has been fitted, a new point is essentially classified according to the sign of a linear combination of support vectors. You are asked to produce the pseudocode (no implementation is required) for computing this linear combination. Your pseudocode should make use of the functions alphaindex, coef and b. See the help of ksvm for information about these functions.