Technical University of Denmark

02450
Introduction to Machine Learning and Data Mining

Assignment 2

Work Distribution

Celia Sullca Cailloux (s132534) **50%**

Oseze Esther Iyore (s121922) **50%**

Technical University of Denmark

April 10th, 2018

Contents

Regression

In this section the data presented in assignment 1 is used and two types of regression are performed: linear regression with and without feature forward selection (FS) and artificial neural network (ANN). The objective of the collection of WDBC data is classification of malignant and benignant breast lumps and based on our knowledge, none of the collected attributes are directly correlated to the classification, thus the following regression example is an artificial construction.

The desired input feature for prediction is chosen to the *Mean Area*. The input features used for predicting the are are listed in Tab. 1. From the original data presented in assignment 1, mean radius and mean perimeter are removed given their obvious mathematical relation to the mean area.

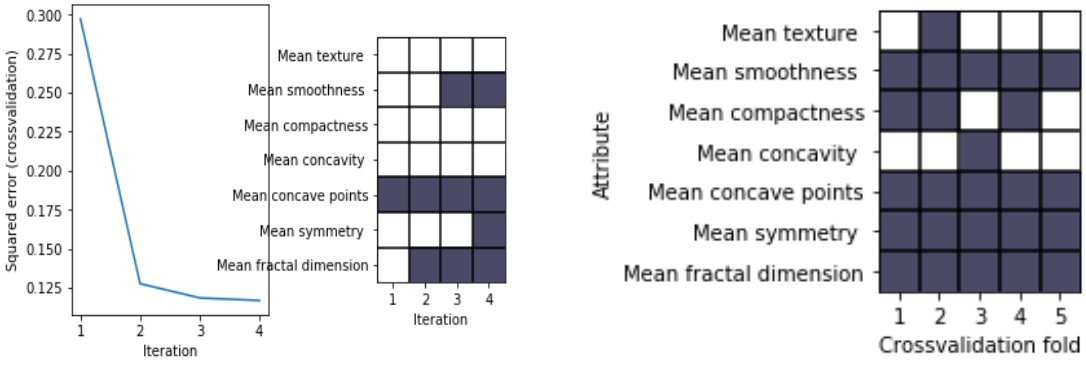
x_1	x_2	x_3	x_4	x_5	x_6	x_7
Texture	Smoothness	Compactness	Concave Points	Concave Points	Symmetry	Fractal Dimension

Table 1: Attributes (mean values) for prediction of Mean Area

Linear Regression using Forward Feature Selection

In linear regression the objective is to model the output feature $\boldsymbol{y}$ as a linear combination of the input features $\boldsymbol{X}$, $\boldsymbol{y}_i = \boldsymbol{x}_i^T \boldsymbol{w}$ where $\boldsymbol{w}$ are the parameters to be optimized to obtain the optimal model. Feature forward selection is used as a method to split the input features into relevant and irrelevant input features for the predicted output feature. However, for the WDBC data significance of a feature selection is not decisive as the number of input features ($M = 7$) is not large compared to the number of observations ($N = 569$).

To find the optimal model for linear regression (avoiding overfitting) two-layer cross validation is performed: 10-fold inner layer and a 5-fold outer layer. The inner layer select the attributes for prediction using forward selection, $\tilde{\boldsymbol{X}}$. The outer layer selects the best model. The generalization error for each model is thus determined by the weighted average of the test errors.



(a) (b)

Figure 1: Linear regression of WDBC using feature forward selection and two-layer cross-validation. (a) illustrates the steps of the forward selection. The selection stops when the generalization error has reached a minimum. (b) summarizes the selected input features in each outer cross-validation fold and shows which input features that repeatedly appear.

Fig. 1 shows the results of the double layer cross validation. All input features and out feature are normalized prior to the linear regression. Fig. 1a shows the forward selection of the best found model (fold 5). Fig. 1b summarizes the selected input features found in each outer layer cross validation fold and shows which input features repeatedly appear. The left graph in Fig. 1a shows how the generalization error reduces when including more and more input features by forward selection. The selected input features per iteration are shown to the right in Fig. 1. However, at some point the generalization error reaches a minimum and the feature selection stops.

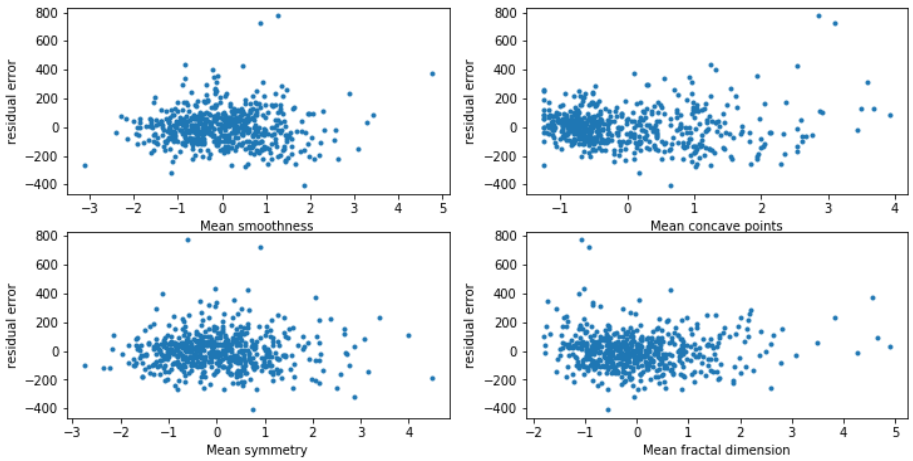


Figure 2: The residual error from the fitted model for cross-validation fold 1 as a function of the input feature. All residual errors appear randomly distributed why input feature transformation is not required.

Fig. 2 shows the residual error of the best model M^* retrieved from cross-validation fold 5 as a function of each of the selected input features. The residuals are retrieved from the predicting on entire data set $\boldsymbol{X}$. The random distribution of the residual error meet the requirement of a residual error ε that is normally distributed around $\overline{\varepsilon} = 0$ and independent of the input feature. Consequently, no input feature transformations were performed.

$\tilde{\boldsymbol{x}}_i$	$\boldsymbol{w}_i$	weight
	-1.05	
Mean smoothness	-0.130	∼2
Mean concave points	0.984	∼14
Mean symmetry	-0.068	1
Mean fractal dimension	-0.337	∼5

Table 2: The selected input features $\tilde{\boldsymbol{x}}$ and the respective parameters of the best model M^* and the weight of each parameter. Mean concave points is clearly the most dominant input feature.

The best model M^* is selected on the basis of the model with the smallest generalization error. The best model is the model found in cross-validation fold 5. The selected attributes are listed in Tab. 2 together with the optimized regression parameters $\boldsymbol{w}$. The dominating input features agree well with the repeatedly occurring input features in Fig. 1b.The sign and magnitude determine the weight of each input feature in the model. The dominating input feature is clearly mean concave point.

	Test error	R^2
No feature selection	0.133	0.865
Forward feature selection	0.134	0.866

Table 3: Comparison of the best models w/o forward feature selection.

As mentioned, for the WDBC data set feature selection is not a necessity. Thus to evaluate the method of feature forward selection, the mean of test errors from the five outer cross-validation fold are compared to the mean test error of linear regression without feature selection. The mean test error and the coefficient of determination R^2 are listed in Tab. 3. The mean test errors are very similar and the R^2 too.

Artificial Neural Network

Artificial Neural Network (ANN) is a method that is inspired by the biological neural network. An ANN consist of a collection of neurons that are connected to other neurons by weighted connections. The neurons are organized in subsequent layers where one layer feeds the next layer etc. First the input features x_i are presented to the input layer where a neuron is given an activation which is then further propagated through the hidden units and finally reaches the output layer.

To find the optimal ANN, the parameter to optimize is the number of hidden units n. For this purpose, two-layer cross-validation is used. Similarly to the procedure used for the linear regression, a 10-fold inner layer is used to select the optimal number of hidden units n^* while the 5-fold outer layer then selects the best of the inner layer models.

Number of hidden units	2	4	6	8	10
Fold 1	0.29	0.27	0.28	0.24	0.17
Fold 2	0.26	0.3	0.32	0.31	0.33
Fold 3	0.15	0.16	0.15	0.17	0.16
Fold 4	0.19	0.14	0.18	0.15	0.13
Fold 5	0.26	0.21	0.21	0.23	0.22
Fold 6	0.27	0.29	0.3	0.32	0.3
Fold 7	0.18	0.2	0.21	0.21	0.19
Fold 8	0.29	0.31	0.29	0.3	0.32
Fold 9	0.25	0.27	0.3	0.22	0.25
Fold 10	0.30	0.28	0.32	0.36	0.34
Average	0.25	0.25	0.26	0.25	0.24

Table 4: Example of the computed test errors in one outer layer with a 10-fold inner cross-validation. In each inner fold five different number of hidden units are tested. The number of hidden units giving the smallest test error average is selected as the optimal number of hidden units n^*_{inner}. I.e. for this outer layer $n^*_{inner} = 10$.

Tab. 4 shows the computed test errors for an outer cross-validation fold. In each outer layer a 10-fold cross validation is run with for different number of hidden units, $n = 2, 4, 6, 8$ and 10, respectively.

Outer cross-validation fold	1	2	3	4	5
Optimal number of hidden units n^*_{inner}	10	10	6	6	10
Test error using n^*_{inner}	0.19	0.24	0.37	0.28	0.14

Table 5: This table shows the test errors using the $_{inner}$ to train the ANN model.

Tab. 5 shows the test errors when using n^*_{inner} in the ANN model. Thus based on the results in Tab. 5 of the two-layer cross-validation, the best number of hidden units given the lowest test error is $n^* = 10$.

Method Comparison

The performance of the regression classifiers will now be compared. The fitted ANN, linear regression without feature selection and the linear regression with feature selection are evaluated statistically to determine whether there is a significant performance difference between them. For linear regression with forward feature selection the best model $M*$ found previously is used. For ANN the optimal number of hidden units is $n^* = 10$.

	Linear reg. (no FS) vs. Linear reg. with FS	Linear reg. (no FS) vs. ANN	Linear reg. with FS vs. ANN
Significantly different	no	yes	yes

Table 6: Significant difference between models. In the three columns: 1) Linear regression without feature selection vs. Linear regression with feature selection. 2) Linear regression without feature selection vs. artificial neural network 3) Linear regression with feature selection vs. artificial neural network.

Table 6 shows whether there is a significance difference between the classifiers or not. The test error for each of the three models is calculated and compared two by two using a paired t-test where the credibility interval ($\alpha = 0.05.$) is computed and used to determine whether the classifiers are significantly different or not. The two linear regression (w/o forward feature selection) are not significantly different. However, the two are significantly different from ANN. Figure 3 shows box-plots for the cross validation error for each of the three classifiers as well as the average of the training data output.

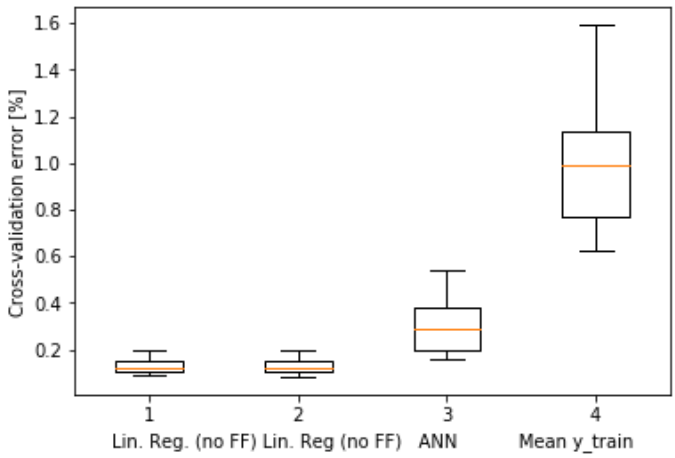


Figure 3: Box-plot showing the cross-validation error for each of the regression classifiers.

The boxplots illustrate the conclusions made in Tab. 6. The figure shows that there is not significant difference between the error for the linear regression classifiers with and without feature selection and also that these two performs better than the artificial neural network classifier. It confirms that there is a significant difference between the linear regression classifiers and the artificial neural network. In addition, the classification error for a model using the average of training data included. This *Mean* y_{train} model is expected to be highly bias which agrees with the significantly higher classification error than the three regression methods.

Classification

In this section a classification problem for the data set is solved. The goal is to classify the instances of measurement of a cell nuclei in a breast mass in two classes: benign (B) or malignant (M). Thus, we wish to predict the output y given a data set $\boldsymbol{X}$. In this case all of the given attributes are used to run the classification. The three methods chosen are:

1. Logistic regression
2. Artificial Neural Network
3. K-Nearest Neighbours

Relevant parameters for the three methods are selected and determined through cross validation. In cross validation the data set is randomly divided into K equal parts and each of these are used as a test set for the created classifier. The data set as well as the training samples are fairly large which result in high accuracy.

Previously in the literature classification procedure decision trees [1],[2] has been used to classify the data. In addition artificial neural networks [3] has also been used to classify the data.

[1]W. Nick Street, William H. Wolberg "Breast Cancer Diagnosis and Prognosis Via Linear Programming"
[2]K. P. Bennett. Decision tree construction via linear programming. In M. Evans, editor, Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society Conference, pages 97101, 1992.
[3]O. L. Mangasarian. Mathematical programming in neural networks. ORSA Journal on Computing, 5:349 360, 1993

Logistic Regression

Logistic regression is a way to use regression for classification by use of probabilities. A train and test procedure is carried out using a single layer 10-fold cross validation to determine the best classifier. In each fold the models are evaluated and compared to determine the best one. Table 7 shows the error measure for each of the folds.

Fold	1	2	3	4	5	6	7	8	9	10
Test error	0.053	0.018	0.053	0.036	0.	0.018	0.018	0.070	0.089	0.071

Table 7: Test error for each fold.

It can be seen from table 7 that the lowest test error appears in the fifth fold which would then be the fold which model would be choosen.

Artificial Neural Network

Artificial neural network (ANN) is described in the previous section in this case it is used for a discrete output instead of a continuous output. Two layer cross validation is performed where the inner layer is used to determine the number of hidden units and the outer is used to evaluate the model. Both the inner and the outer layer has 4 folds. The number of hidden units in each fold are: $n = 2, 4, 6, 8$ and 10, respectively. The error measure for each of the folds in the layer is given in table 8.

Outer cross-validation fold	1	2	3	4
Optimal number of hidden layers n^*	6	10	8	6
Test error using n^*	0.62	0.28	0.34	0.36

Table 8: Test error and optimal number of hidden units for each fold.

Regarding table 8 it is seen that the lowest test error is in fold 2 where the number of hidden units is 10. The highest test error is in fold 1 where the number of hidden units is 6.

K-Nearest Neighbours

In this subsection the K-Nearest neighbours method is used. The goal of this method is to determine the classification of an object based on the class of the surroundings neighbours. The object is assigned to the class of which the most of the its neighbours belong.

In this case the measure used to determine distance is the euclidean distance. With this defined the first thing is to use cross validation to determine the number of nearest neighbors which results in the best classification. This is done using the leave-one-out method and figure 4 shows the classification error rate in % as a function of the number of nearest neighbours and it is determined that the optimal number of neighbours in this case is [4] since this is the value where the error rate is the the lowest.

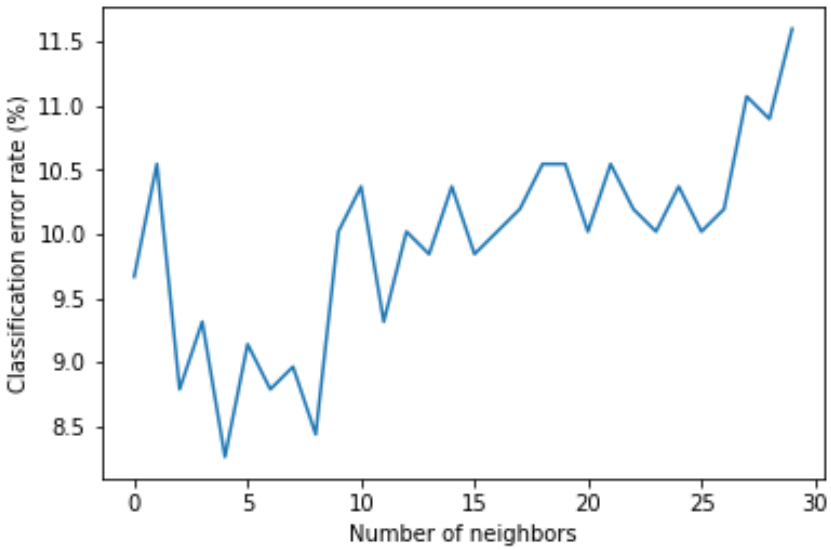


Figure 4: Classification rate as a function of number of nearest neighbors

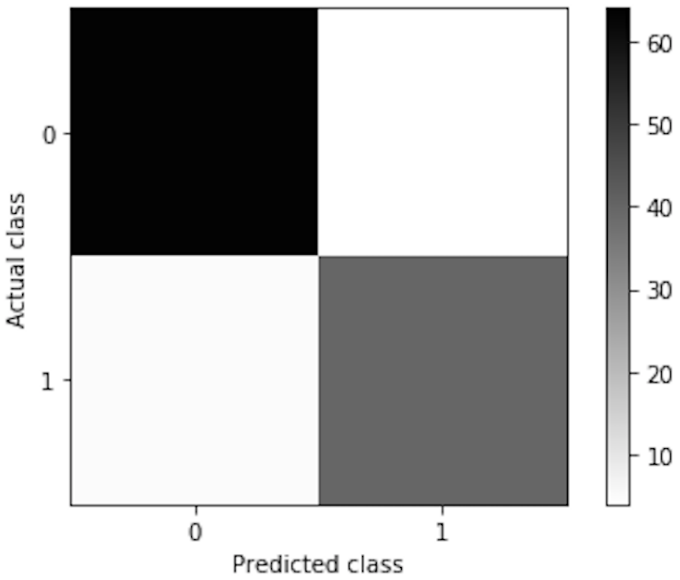


Figure 5: Confusion matrix

This value along with the euclidean distance is now used as inputs for the K-Nearest Neighbours method to classify the objects. Figure 5 shows the confusion matrix of the problem. A confusion matrix or error matrix visualizes the performance of a certain classifier. As seen on the figure the diagonal represents the correct classification and all other positions represent wrong classification. In this case it is seen that there is close to no instances of wrong classification. The accuracy is given as 92.04 % which leads to an error rate of 7.96 %.

	Logistic reg. vs. K Nearest Neighbours	Logistic reg. vs. ANN	K Nearest Neighbours vs. ANN
Significantly different	yes	yes	yes

Table 9: Significant difference between logistic regression, KNN and ANN.

Method Comparison Tab. 9 shows whether there is a significance difference between the classifiers or not. The test error for each of the three models is calculated and compared two by two using a paired t-test where the credibility interval is computed and used to determine whether the classifiers are significantly different or not.

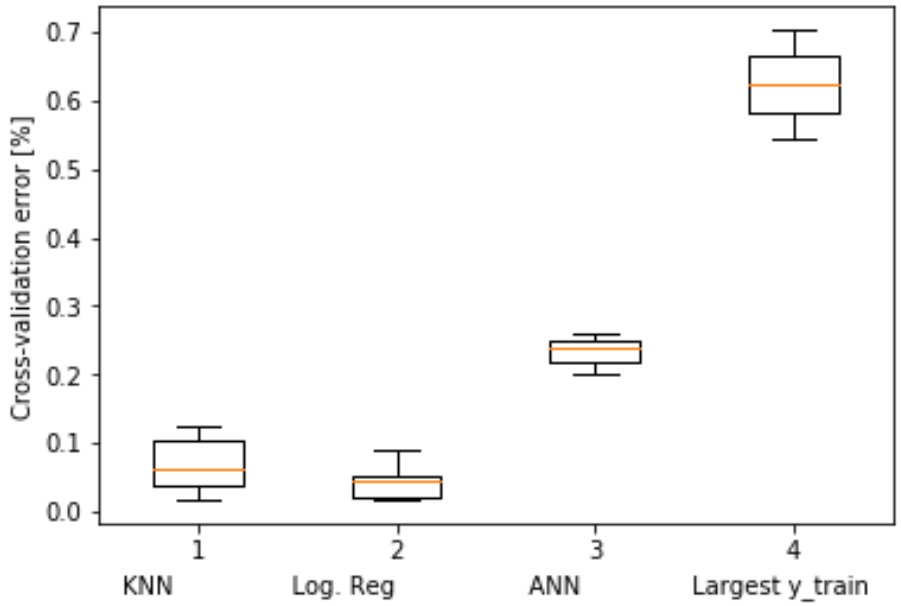


Figure 6: Box-plot showing the cross-validation error for each of the classification classifiers

In figure 6 the box-plot of the cross-validation error for KNN, logistic regression, ANN and the largest value for ytrain. The figure indicates that there is a significant difference between all of the three classifiers where the logistic regression methods seems to be the one performing the best and the artificial neural networks seems the be the one performing the worst with the highest cross-validation error. Still the method seems to be performing significantly better than the largest ytrain value. In this case logistic regression would be the best choice due to the low error and the scope of the analysis.

Appendix

Work distribution

We worked in collaboration for both the regression and classification part of the assignment. However Celia (s132534) had the main responsibility the for regression part while Oseze (s121922) had the main responsibility for the classification part.