02450 Introduction to Machine Learning and Data Mining

Report 2 - Supervised Learning: Classification and Regression

s153355 Leif Førland Schill
s205972 Jens Waaben
s202915 Angelos Daglaroglou

	Leif	Angelos	Jens
Section			
Regression part A:	30	30	40
Regression part B:	30	40	30
Classification:	40	30	30
Discussion:	30	40	30

1 Regression part A

This assignment aims to apply machine learning techniques to data obtained from Stanford.edu.. For a more in detail explanation of the data set and its associated variables we refer to our previous assignment. This first section utilizes linear regression to make a prediction of systolic blood pressure - a crucial parameter for cardiac health.

1.1 Variable selection for linear regression predicting systolic blood pressure:

For the linear regression problem we consider only continuous variables (**ldl**, **tobacco**, **adiposity**, **typeA**, **obesity**, **alcohol**, and **age**). To further subset the important variables we perform two-level 10-fold cross validation, with the outer split dividing the data set into 10 partitions of training and test data to validate the linear regression, and 10 inner splits of the training sets - for each of these splits the loss for each variable is calculated and the one with the smallest loss function is chosen. The average loss function (overthe inner splits) is then calculated for the remaining variables (sequential feature selection) until the loss function does not increase anymore. The following variables were selected for each outer split:

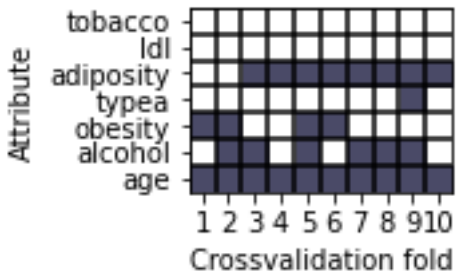

Figure 1: Attributes selected across 10 outer folds

The 4 attributes selected for the final model is therefore: Age, Obesity, Alcohol and Adiposity.

1.2 Optimizing the regularization parameter λ :

The next step in optimizing our linear model was to introduce the regularization parameter λ. To optimize the size of λ, a number of different λ values were chosen ($\lambda \in (0, 300)$) and 10-fold cross validation was performed on each λ, estimating the average generalization error of the linear model with a λ-term of the given size. The value of λ can be seen in figure 2.

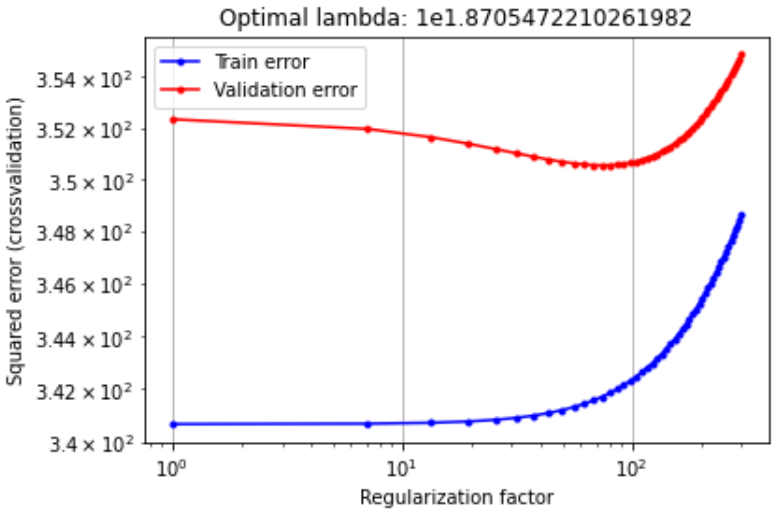

Figure 2: Average generalization error of each λ value across 10-fold cross validation

As it can be seen from the figure, the optimal λ value is app. 74.22. At smaller λ values the test-set error increases, and likewise for bigger values.

1.3 Predicting systolic blood pressure based on lowest generalization error estimate:

The finished model uses the optimal lambda to optimize weights to minimize the error function (squared error).

$$f(x) = \boldsymbol{x}^T \lambda \boldsymbol{w} + w_0$$

The actual estimated numbers are (with λ weights applied):

$$f(x) = \begin{pmatrix} x_{Adiposity} & x_{Obesity} & x_{Alcohol} & x_{Age} \end{pmatrix} \begin{pmatrix} 2.73 \\ 1.19 \\ 1.78 \\ 5.01 \end{pmatrix} + 138.33$$

Where x are the data attributes: adiposity, obesity, alcohol and age. The lambda parameter is 1.91 (found above) and w_0 is the offset value. When the SBP of a new patient is predicted, the attribute values are plugged into the equation above, and the result noted.

2 Regression part B

In the second part of Regression, we will implement two-level cross-validation to compare the models with 10-folds ($K_1 = K_2 = 10$). The models we will use is the linear model used in the previous section, a baseline model which is a linear regression model with no features and a Artificial Neural Network (ANN). We will use the same splits for the folds during training to allow statistical comparison between them, in the next sections. As complexity parameters we used $\lambda \in (0, 200)$ for the linear regression model (where, as we noticed, the estimation error first drops and then increases) and for the ANN, $h \in [1, 10]$ hidden units.

2.1 Comparing baseline, regularized linear model, and ANN using two-fold cross validation:

Outer fold	ANN		Linear Regression		Baseline
i	h^*	E_i^{Test}	λ_i^*	E_i^{Test}	E_i^{Test}
1	6	330.376	66.6667	316.055	368.912
2	6	382.008	44.4444	374.448	439.705
3	7	386.678	44.4444	323.562	417.374
4	7	520.757	44.4444	475.042	595.098
5	5	390.540	66.6667	391.740	414.351
6	6	417.597	44.4444	334.035	405.105
7	7	415.096	66.6667	387.828	485.878
8	5	195.830	66.6667	198.597	238.794
9	6	459.506	66.6667	394.426	457.416
10	7	317.213	88.8889	303.334	384.232

Table 1: Comparison of three, two-Level cross-validation regression models

After running the three models, we can initially compare them and then evaluate them. In order to compare them, we produced a table which shows the optimal value of λ and h for each outer fold as found after each inner loop and then gives us the error measure as estimated by the function:

$$E^{Test} = \frac{1}{N^{test}} \sum_{(\imath=1)}^{N^{test}} (y_1 - \hat{y}_i)^2.$$

With a quick glance at the table, we see that the ANN gives us the least error for 6 or 7 hidden units, while the linear regression model for values of 44.444 and 66.667. In general we see that, although the error being always high, the Linear regression model gives better prediction while the baseline gives the highest error . Finally, taking under consideration the values of λ that we took for this part, we see that we get similar optimal values for it as in part A.

2.2 Statistical evaluation of performance differences

Using a paired t-test, we can analyze the differences in loss

$$z_i = z_i^A - z_i^B$$

where z_i^A and z_i^B are the squared losses for each observation for model A and model B.

$$\hat{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$$

$$\tilde{\sigma}^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(z_i - \hat{z})^2$$

The estimated difference in generalization error, computed using the L2 loss, is given as $\hat{z}$, and since n is large we compute a confidence interval based on the student-t distribution with parameters $\mu = \hat{z}$ and variance $\sigma^2 = \tilde{\sigma}^2$.

Based on this approximation, 95% confidence intervals for $[z_L, z_U]$ are constructed for each pair of models, along with a p-value for the hypothesis

$$H_0 = \text{Model } \mathcal{M}_A \text{ and } \mathcal{M}_B \text{ have the same performance.}$$

Using this method, these conclusions are valid only for this specific training set.

	z_L	$\hat{z}$	z_U	p
RLR - baseline	-95.67	-70.57	-45.47	2.76e-08
ANN - baseline	-56.85	-39.70	-22.56	3.42e-06
RLR - ANN	-49.08	-30.87	-12.66	4.66e-04

Table 2: Pairwise comparisons of the three regression models.

The results show that both models are significantly better than the baseline, and regularized linear regression is significantly better than the artifical neural network.

3 Classification

In classification, our objective is to identify to which set of subcategories a new observation belongs on the basis of a training set of data containing observations (or instances), whose category membership is known. In short, a classification model attempts to draw some conclusion from observed values and tries to predict the value of one or more outcomes.

3.1 Overall classification problem:

In this section, we will try to create a model that will be able to predict Coronary Heart Disease (CHD) occurrence, based on all risk factors listed in paragraph 1.1. This is a binary classification problem, where $y = 0$ corresponds to the negative class (CHD - Absent) and $y = 1$ to the positive class (CHD - Present).

3.2 Log. regression, ANN and baseline models used to predict CHD:

We will use three methods in our study. The first is baseline model which computes the largest class on the training data, and predict everything in the test-data as belonging to that class. The second is a logistic regression model, using ten values of $\lambda \in (0, 200)$ as a complexity-controlling parameter. The final model we will use is a KNN (k-Nearest Neighbours) model, with complexity controller parameter $k = 1, 2, ..., 10$. The range of numbers for λ and k are based on trial runs, occurred prior to this report.

3.3 Using two-level cross validation to compare model performance:

Outer fold	KNN		Logistic Regression		Baseline
i	k*	E_i^{Test}	λ_i^*	E_i^{Test}	E_i^{Test}
1	5	0.340426	23.1111	0.255319	0.319149
2	8	0.425532	1	0.276596	0.382979
3	7	0.434783	1	0.326087	0.369565
4	9	0.326087	1	0.26087	0.391304
5	8	0.347826	1	0.26087	0.369565
6	8	0.347826	1	0.391304	0.369565
7	8	0.391304	1	0.282609	0.347826
8	5	0.347826	67.3333	0.326087	0.347826
9	9	0.282609	1	0.326087	0.23913
10	6	0.23913	1	0.152174	0.326087

Table 3: Comparison of three, two-Level cross-validation classification models

We have used two-level cross-validation to train our data and now we can compare them and evaluate them accordingly. Once again we can see the table above and have an overview of how our models performed. We implemented a 10-fold ($K_1 = K_2 = 10$) cross-validation for the outer and the inner folds, we extracted the inner-folds that gave us the smallest error for each outer fold and their corresponding error as measured by the error rate:

$$E^{Test} = \frac{\text{Number of misclassifed observations}}{N^{test}}.$$

We see that, in general, the KNN model gave us the least error for high values of k, while the Logistic Regression for $\lambda = 1$. Generally, it seems that the Logistic Regression worked slightly better than both the KNN model and the Baseline that we set.

3.4 Evaluating models statistically:

To compare the classifier models we analyze their binary predictions for each observation. For models A and B these are denoted $\hat{y}_1^A, ..., \hat{y}_n^A$ and $\hat{y}_1^B, ..., \hat{y}_n^B$. Let n_{12} be the number of observations model A predicted correctly, and B incorrectly. Let n_{21} be the number of observations model A predicted incorrectly, and B correctly. The estimated difference in accuracy of the models is given as

$$\hat{\theta} = \frac{n_{12} - n_{21}}{n}$$

Using McNemar's test for comparing classifiers, 95% confidence intervals $[\theta_L, \theta_U]$ for θ can be calculated, along with p-values for the hypothesis

$$H_0 = \text{Model } \mathcal{M}_A \text{ and } \mathcal{M}_B \text{ have the same performance.}$$

	θ_L	$\hat{\theta}$	θ_U	p
LR - baseline	0.0037	0.050	0.096	0.043
KNN - baseline	-0.022	0.017	0.056	0.46
LR - KNN	-0.0056	0.032	0.070	0.12

Table 4: Pairwise comparisons of the three classification models.

The tests show that logistic regression is significantly better than baseline, but only just. KNN was not significantly better than baseline, and logistic regression was not significantly better than KNN.

These tests are valid for the specific data set the models are trained on.

3.5 Training logistic regression with optimal λ

The most common optimal λ from the two-level cross validation (Table 3) is $\lambda = 1$. Using this value, the logistic regression is trained on the entire data set. The resulting weights are shown below.

intercept	-0.837
sbp	0.116
tobacco	0.333
ldl	0.392
adiposity	0.134
typea	0.388
obesity	-0.237
alcohol	0.035
age	0.723

The prediction is made by rounding

$$\hat{y}_i = \sigma(\tilde{\boldsymbol{x}}^T \boldsymbol{w})$$

where σ is the logistic sigmoid function.

In terms of making a comparison with the regression for sbp in Part A, we are now predicting CHD; on top of that we are using all the continuous variables instead of just the four (adiposity, obesity, alcohol, age) we used in regression.

It seems that alcohol is not as important in predicting coronary heart disease as blood pressure. Obesity had a positive weight in predicting blood pressure, but it has a negative weight here for the coronary heart disease classification.

4 Discussion:

This study applied different regression and classification methods in an effort to predict sbp and CHD. This next section provides comparison and evaluation of the models chosen to do both the regression analysis and the classification.

4.1 Evaluation of overall machine learning aim:

The main machine learning aim was to attempt to predict CHD. In our previous assignment we stated that, based individual attribute analysis, it could be difficult to answer the main machine learning task. The results presented in this assignment seem to back up these claims. It proved hard to predict CHD with great certainty. However, we did manage to improve decision making compared to baseline by using logistic regression (although slightly). KNN showed a less precise CHD estimation than a baseline. To improve the models a bigger sample or diversifying the population could be attempted. For actual real-world model application this would be unavoidable, since we cannot expect general trends in CHD to be displayed solely by a small south African population.

For the regression analysis part, we chose SBP as it is known to be associated with CHD. Both our ANN and our linear regression model outperformed our baseline model, although the estimation error remained relatively high for all models. SBP is a measure which is highly biologically variable, which makes it reasonable that it can be hard to predict. This along with the fact that the attributes might not be the best in predicting SBP (this is also the case for CHD) could another explanation.

4.2 Comparison of results to other studies:

We have not been able to find a fulfilling peer reviewed study using the data, since the original paper where the data was first published. The initial study took place in 1983 and does not include advanced modeling techniques, but rather looks at the incidence elevated levels of the data sets attributes (e.g. SBP above 160). This type of data analysis is not comparable to what we have performed in this assignment. The data is however very well represented as teaching material for regression and classification. The conclusions (as well as their questionable quality in a lot of cases) from these "studies" have not yielded significantly different results from ours, and have therefore been omitted. However, the need for a reliable way to predict CHD would seem extremely useful for doctors, since there is no cure for coronary heart disease.