Mikkel Sinkjær - s164548
Iben Fjord Kjærsgaard - s164529
Danmarks Tekniske Universitet

10. april 2018

Regression

The main objective of this project is to see if the chosen data set can predict diabetes in people with Pima India heritage. This main objective can not be examine with a regression problem because it is a classification problem. Therefore the following regression problem have been chosen instead:

$$Glucose = f(Pregnant, bloodPressure, Skinthickness, BMI, Pedigreefunction, Age) \tag{1}$$

The above problem have been chosen based on our PCA, which we made in Project 1, as it stated that glucose was the attribute which explained most in relation to the data set. In the next sections we will try to solve the above stated regression problem with the use of Linear regression model and ANN. Throughout the report there will be used normalized data, which is calculated with a function in python.

Linear regression

Table 1 below show which coefficient there are most likely to predict the glucose level. The attributes have been evaluated by a two layer cross validation where the inner layer was a 10-fold cross validation forward features selections to select which attributes will give the lowest squared error in a linear regression model. Thereafter the linear regression model with the lowest squared error have been through a 5 fold outer cross-validation to test the model and calculate the mean squared error by dividing the sum of the squared error with the number of observations.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.146
Model:	OLS	Adj. R-squared:	0.141
Method:	Least Squares	F-statistic:	30.65
Date:	Thu, 05 Apr 2018	Prob (F-statistic):	5.10e-13
Time:	16:00:13	Log-Likelihood:	-485.19
No. Observations:	362	AIC:	974.4
Df Residuals:	360	BIC:	982.2
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.1620	0.049	3.322	0.001	0.066	0.258
x2	0.3377	0.049	6.924	0.000	0.242	0.434

Omnibus:	12.265	Durbin-Watson:	1.888
Prob(Omnibus):	0.002	Jarque-Bera (JB):	13.014
Skew:	0.461	Prob(JB):	0.00149
Kurtosis:	2.887	Cond. No.	1.05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Tabel 1: Linear regression summary.

The P-value in the above table is respectively 0.001 and 0 which indicates that there is a statistically significant relationship between the term and the response. Furthermore is the 95% confidence interval for the to coefficient not between zero which supports that there are a statistically significant relationship. If there is however looked at the squared error it can be seen that the model is not good at predicting the correct values because it is close to zero.

The following figure shows the squared errors vs. attributes for each 5-fold cross validation which means these attributes were the best to predict the glucose level in regards to the linear regression model.

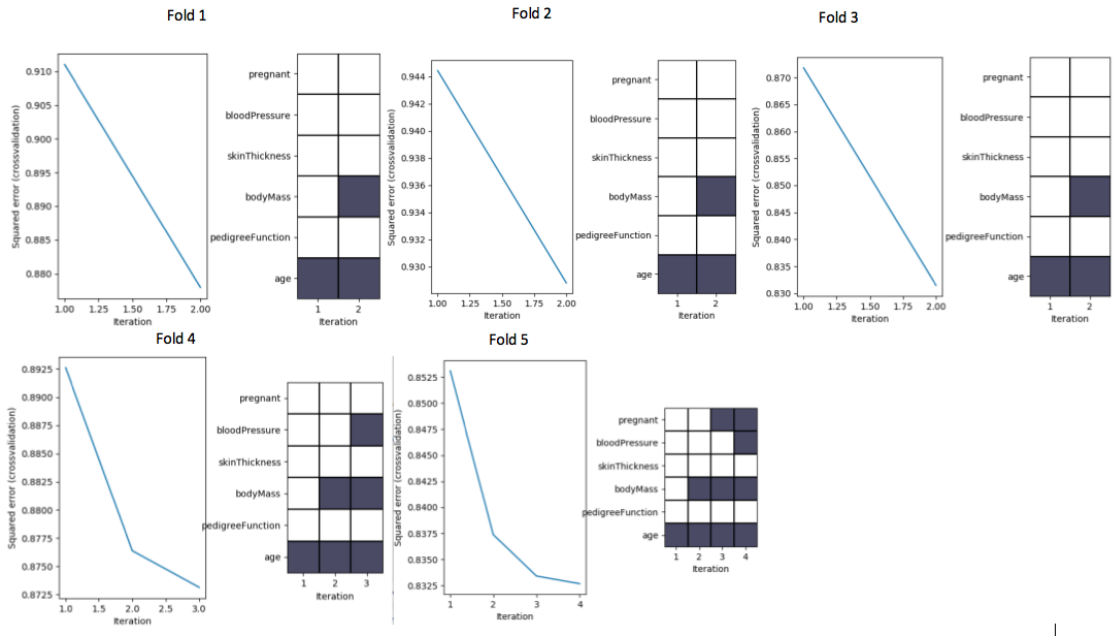

Figure 1: Squared errors vs. Attributes for features selected.

The attribute combination in cross validation fold 3 is the one with the lowest squared error. Therefore it is the one that have been used to make the linear regression model. Thus coefficient x1 is multiplied with the attribute Age and coefficient x2 is multiplied with the attribute BMI.

To investigate whether a transformation of one of the variable can improve the model, the residual error is plotted against the attributes which gives the following result:

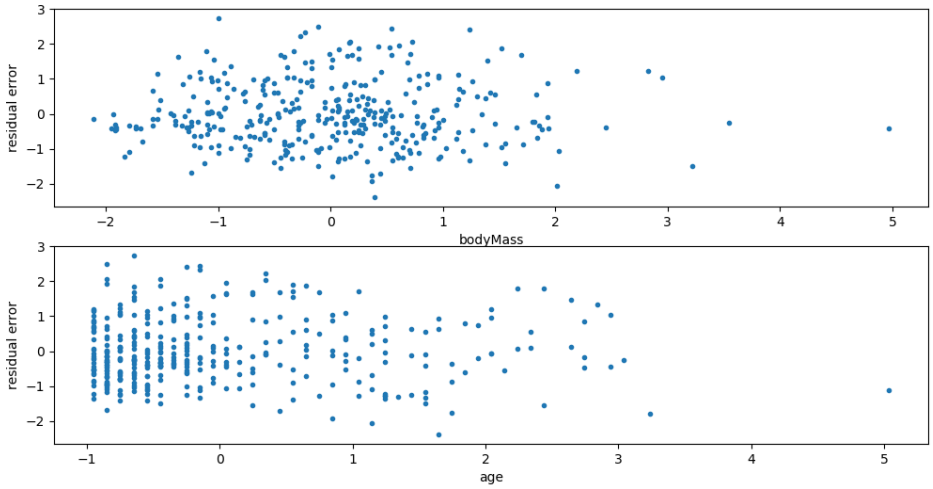

Figure 2: Residual error vs. Attributes: BodyMass and Age.

From the above figure it can be seen that there is no visible correlation between residual error and the attribute, which means that none of the variable will be transformed.

The linear regression model will therefore be:

$$Glucose = 0.337 * Age + 0.162 * BMI \tag{2}$$

ANN Regression

The ANN regression model which is best at predicting the glucose level is found by a two layer 5-fold cross validation. The 5-fold cross validation inner layer is a feature selection to find the optimal hidden units which is saved in each fold and used in the outer layer. Additionally the model is tested in a 5-fold cross validation outer layer where the mean squared error is calculated by dividing the sum of the squared error with the number of observations. The lowest mean squared error was calculated to be 0.88 which can be seen in the below figure:

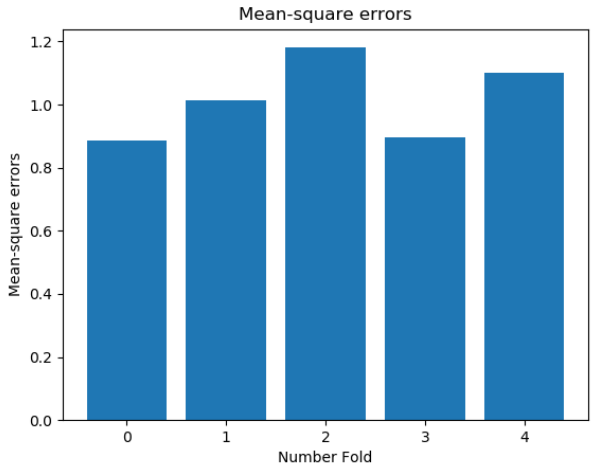

Figure 3: Square errors ANN.

Comparing Linear regression and ANN regression

When evaluating the significance of the difference in performance between the linear regression model with the ANN regression we use the 5-fold cross validation errors for the outer layer.

This is calculated by a t-test which test the following null hypothesis and alternative hypothesis:

The Linear regression model is not significantly different from the ANN regression model.
The Linear regression model is significantly different from the ANN regression model.

The t-test calculate a p-values equal to 0.19 which is higher then the significance level on 0.05, meaning that we can not reject the null hypothesis:

The to regression models are not significantly different!

There is not one of the models there are significantly better then the other. The to models 5-fold cross validation errors for the outer layer is shown in below boxplot:

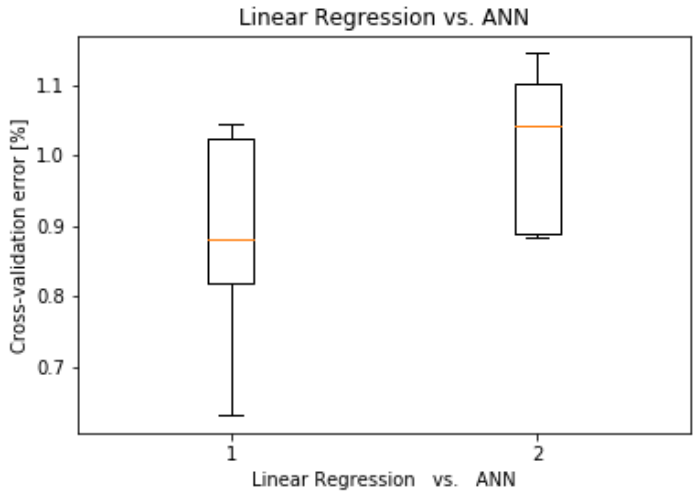

Figure 4: Linear Regression vs. ANN.

In addition to the above comparing of the to model, we will compare whether the models are better than simply predicting the output to be the average of the training data output. We will again use the t-test to evaluated if the models are significance difference to be the average of the training data output.

The null hypothesis and alternative hypothesis for comparing Linear regression with subtracting the mean is:

The Linear regression model is not significantly different from subtracting the mean.
The Linear regression model is significantly different from subtracting the mean.

The above t-test calculate a p-values equal to 0.15 which again is higher then significance level on 0.05 meaning that the null hypothesis can not be rejected:

The Linear regression model are not significantly different to subtracting the mean!

The null hypothesis and alternative hypothesis for comparing ANN regression with subtracting the mean is:

The ANN regression model is not significantly different from subtracting the mean.
The ANN regression model is significantly different from subtracting the mean.

The t-test calculate a p-values equal to 0.8, which means the null hypothesis can not be rejected:

The ANN model are not significantly different to subtracting the mean!

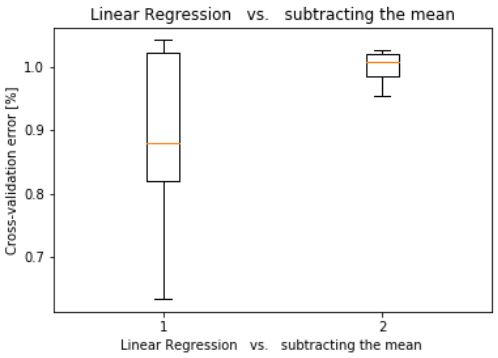

Figure 5: Linear Regression vs. ANN.

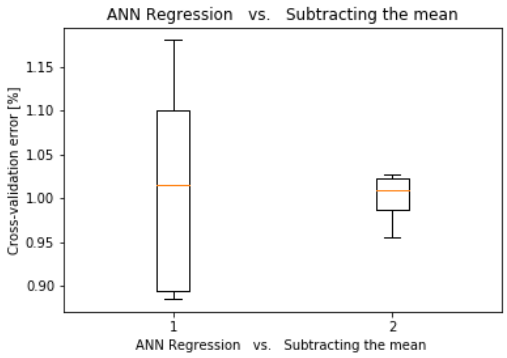

Figure 6: ANN Regression vs. Subtracting the mean.

The Linear regression model and the ANN regression model is not significantly better the simply subtracting the mean from the output, which indicates that the models is not very good at predicting glucose concentration at Pima Indian women.

Classification

Our data set has a attribute called class variable which explain if a person is diagnosed with diabetes or not. It is a great attribute to try to predict using classification, as it is a binary feature and remaining attributes only has the purpose to explain if the is a correlation between them and diabetes patients. It is by then the obvious choice response to predict given the values of the other attributes. We want to solve the classification problem as stated below.

$$Diabetes = f(Pregnant, glucose, bloodPressure, Skinthickness, BMI, Pedigreefunction, Age) \tag{3}$$

We have chosen to use the following three classification methods: decision tree, Naïve Bayes and ANN.

Decision Tree

The decision tree has as goal to minimize impurities and by then increase the purity gain. We have chosen to measure impurity by the Gini functionen. The decision tree is found by using a two layer 5-fold cross validation. in the inner layer of the cross validation we do a feature selection to find the optimal depth of the tree. The best depth is found by making a tree in each inner fold and monitor the error at each depth when bieng tested. By meaning the error for each depth and choosing the depth with the lowest error, we can expect to have found the optimal depth. every time the inner layer is run we find 3 the best depth to use in the outer layer.

In the outer layer a decision tree is made in each iteration. The error is calculated as times the decision tree make a wrong guess divided by amount of observations in the test data. the best error is found to be 20.5% in the second decision tree. The mean proformance for the five decision trees is plotted beneath

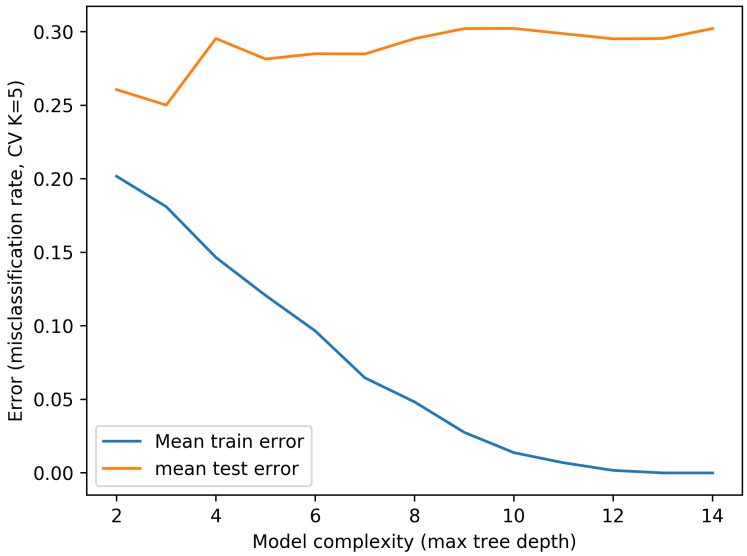

Figure 7: Mean Performance of decision tree. Numbers on the bars represent value of hidden layer

The plot shows that the decision tree keep performing better and better on the train. It is not the same case for the training data where it only improve its performance until depth reach 3. When this depth is reached the error go up a bit and stabilizes.

The following diagram show the decision tree made from the entire data set with a depth = 3

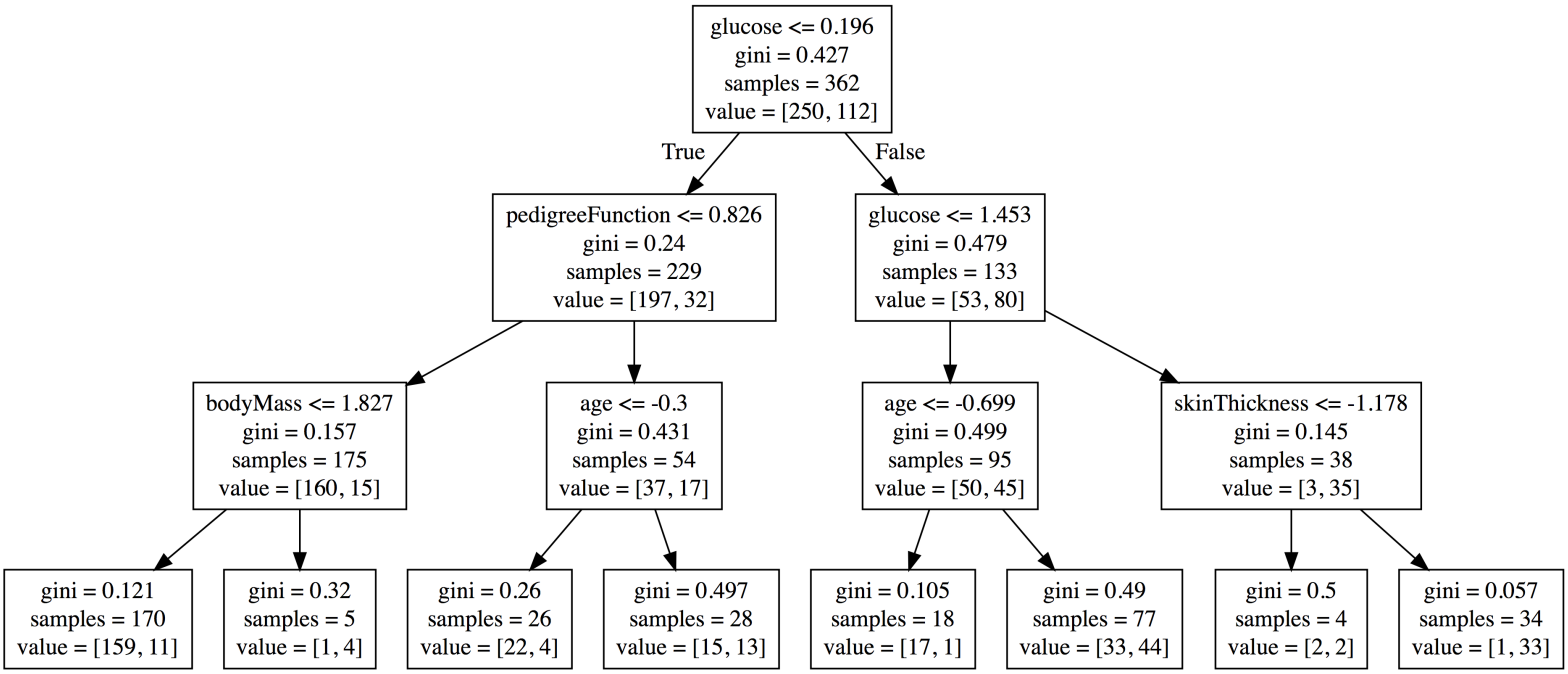

Figure 8:Decision tree.

The first question the decision tree is rather a person has a glucose <= 0.196(standardized value).

The first question asked is the one which has the largest purity gain and by the the best question to ask. The purity gain can be calculated by using equation (8.1) and (8.3) in the text book, the purity gain for the first question is calculated as follows:

$$\Delta = I(r) - \sum_{k=1}^{C} I(v_k)$$
$$I(parent) = 1 - (\frac{250}{362})^2 - (\frac{112}{362})^2 = 0.42$$
$$I(True) = 1 - (\frac{197}{229})^2 - (\frac{32}{229})^2 = 0.24$$
$$I(False) = 1 - (\frac{53}{133})^2 - (\frac{80}{133})^2 = 0.47$$
$$\Delta = 0.42 - \frac{229}{362}0.24 - \frac{53}{362}0.47 = 0.20$$

The First question in the decision tree has a purity gain = 0.20.

After the first split the tree either ask about a persons pedigreeFunction value or glucose value again. In the last split the tree as for bodymass, age or skinThicknesss. The four attribute the tree ask about seem to have bigger impact in deciding reather a person has diabetes or not. If we ar to make some new observations and trust the decision tree we will only have to measure then four attributes to decide if a person is sick or not. Tree has as mentioned an average error rate = 25 %, based on this you properly wouldn't trust the answer very much.

ANN for classification

The ANN classification model is made in a two layer 5-fold cross validation. The inner layer is used for selecting optimal model. The parameter to be selected is number of hidden units. Each fold of the inner layer is trained with hidden units equal 1 to 8. The value for hidden units with the smallest error is saved in each fold and used in the outer layer. In the outer layer a new ANN is made with using the best hidden layer performing the best. The error for the outer layer ANN is saved an used for evaluating the performance in each outer fold. The following histogram show the performance of the ANN's

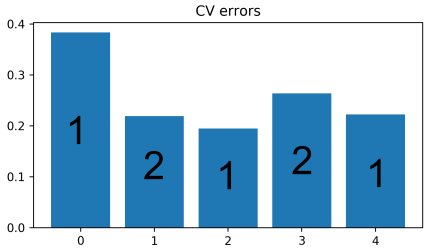

Figure 9:Mean prediction error og ANN found in outer layer.

The histogram that one and two layer are found to be the two best value for hidden units. by looking at the histogram it seems that one hidden unit has the ability to have the lowest prediction error but have a big variance. Two unit has a lower variance and perform from the limited error better on average.

The following grapth show the estimated class vs the true class

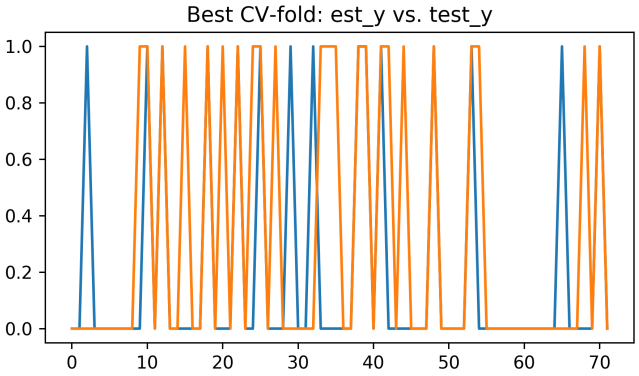

Figure 10:Last CV-fold: est_y vs. test_y.

the blue line represent the estimated calls and the orange the true class. It is clear to see that our model guest that person is non-diabetick most of the time. Model model showen has a prediction error equal 22.2 % and the avereage proformance of all five ANNs is equal 25.7 %

Naïve Bayes classifier

In Naïve Bayes classifier a to layer cross validation is used to find the optimal model. In the inner the parameter alpha is selected. The best alpha is found by making a model for alpha range from 1 to 1000 in the inner layer. The The alpha from the best model is selected and used in the outer layer. The alpha value smooth out the model. An alpha value equal zeros makes no smoothing on the model and the higher the alpha the more smoothing. The alpha values used in the outer layer is 93, 11, 88, 8 and 342. The alpha value.

In the outer layer five model are found. They perform on average with a error rate equal to 27.6% with a standard division equal to 3%

Comparing classification models

The two classification models which perform the best is ANN (error = 25.7 %) and decision tree (error = 25.0 %), chosen in addition to lowest mean error. we will test the two data set statistical by testing the null hypothesis using a t-test. Running the t-test the result of the p-value is 0.93. We can not reject the null hypothesis and can't conclude if one model is better than the other.

Both models are now tested against error rate if we always guess on the biggest class, non-diabetes. Testing the decision tree and biggest class guess the p-value is found to be 9.2e-08. It can be concluded that the decision tree is significantly better than just guessing on the biggest class. The same is done testing ANN instead of decision tree. The p-value is found to be 6.7e-05, again significant better performance than just guessing the class to be the biggest class all the time.

It is not possible to say if ANN perform better than decision tree and vice versa. both model do seem to perform better than just guessing on the biggest class as the models in both t-test showed significant better performance.

Distribution of assignments:

Regression	**Navn**
1	Iben
2	Mikkel
3	Iben
4	Mikkel
5	Iben

Classification	**Navn**
1	Mikkel
2	Iben
3	Mikkel
4	Iben