Case 1

02582 Computational Data Analysis

Andri Bergsson

s150843

March 24, 2019

Code

The code for the case can be found in this repository on github:
`https://github.com/andberger/computational-data-analysis-case-1`

All the code is in one file, called `script.py`:
`https://github.com/andberger/computational-data-analysis-case-1/blob/master/work/script.py`

Introduction/Data description

We are given a set of data consisting of 100 observations (Y, X), where Y is the response (univariate) and X are the features (100-dimensional). We are additionally given 1000 observations of the features.

The problem is then to build a predictive model, so that we can predict a new value of Y based on the 100 features X.

Y is a continuous variable, as are the feature variables $X1$ up to $X95$. $X96$ up to $X100$ are categorical variables, holding values of A, B, C and so forth. Since we do not know the real meaning behind those categorical values, we cannot assume that their order has any meaning, so they must be thought of as nominal.

Model and method

Since we are predicting a single continuous variable, a regression model seems like it would give us a good predictor. The approach is then to try out different regression methods (regularization techniques), namely linear regression, ridge regression, lasso regression and elasticNet regression, and see which one can give us the best predictor for our data.

In short we load the data, handle missing values (see how in section below), one-hot encode the categorical data (see how in section below), standardize and normalize the data and then do K-fold cross-validation ($K = 10$) for each of the regression methods. For each regression method, the cross-validation tries out different hyper parameters and picks the one that gives the model that has the best score (we decided on using the coefficient of determination, R^2, as a scoring metric). We now have the best model we can find for each regression method, and out of these models we choose the model we think best fits our data (see how in section below).

Missing values

Missing numerical values are replaced by the mean of their variable (so for example a missing value in $X1$ is replaced by the mean of all the values in $X1$). Missing categorical values are replaced by the most common value of their variable (so for example, missing values in $X96$ are replaced by the most common value in $X96$, which is B). These are common practices which ensure that replacing missing values doesn't alter the underlying structure of the data set.

Factor handling

Factor variables, also called categorical variables, are handled by using one-hot encoding. One-hot encoding is a way to work with categorical data when doing regression. One-hot encoding transforms each categorical variable into multiple variables, 1 variable for each categorical value, and holding values of either 0 or 1 (1 if the categorical variable had said value in a particular row, 0 otherwise). One-hot encoding is illustrated in Figure 1.

X96	X97	X98	X99	X100	one-hot	X96_A	X96_B	X97_A	X97_B	X97_C	...
B	A	C	C	H	=====>>>>>	0	1	1	0	0	
B	B	D	B	B		0	1	0	1	0	
B	B	B	D	E		0	1	0	1	0	
B	B	A	B	D		0	1	0	1	0	
A	C	D	D	D		1	0	0	0	1	
.	.	.	.	.		.	.	.	.	.	
.	.	.	.	.		.	.	.	.	.	
.	.	.	.	.		.	.	.	.	.	

Figure 1: One-hot encoding explained

Model selection

Model selection is performed using cross-validated grid-search over a parameter grid. (`https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`).

Model validation

Model validation is done by K-fold cross-validation ($K = 10$). The scoring metric used for evaluating each model for each of the regression methods is the coefficient of determination R^2. R^2 is an evaluation metric that represents the ratio between how good a model is versus a model that always predicts the expected value of the dependent variable, disregarding the input features. The model with the highest R^2 (best possible score is 1.0) will be the best predictor for our data.

Results

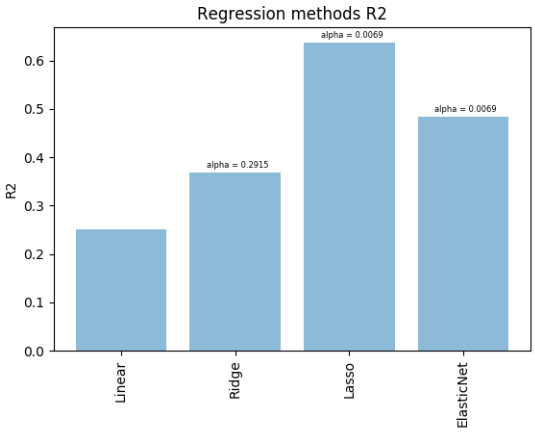

Figure 2: R^2 for the different regression methods tried

As can be seen by Figure 2, Lasso regression with hyperparameter α value 0.0069 yields the highest value for R^2 and is therefore the best and our chosen model. We will use this model to generate our predictions, $\hat{y}_{new}$.

To estimate $RMSE$, we use K-fold cross-validation again ($K = 10$), but now using our chosen model as an estimator and using the mean squared error as a scoring metric. We average over the K errors to obtain the cross-validation error (in our case the average MSE from the K cross-validation folds), and finally take the square root of that average to obtain an estimate of the *root mean squared error*, $\widehat{RMSE}$. We get that

$$\widehat{RMSE} = 0.5526577194231526$$