Homework #2: Linear Regression

Download the data about body fat measurements from http://file.biolab.si/files/fat.xlsx. You can find the description of this data set at http://ww2.amstat.org/publications/jse/datasets/fat.txt. We use "Percent body fat using Brozek's equation" as the target, and we have removed its near-duplicate "Percent body fat using Siri's equation". From the original features, we have also removed the feature "Density", which cannot be routinely measured by GPs.

1. Build a linear regression model to predict the body fat from the given measurements. Report on its accuracy and compare it with the baseline model.
2. Use the Predictions widget and a Scatter Plot to show the relation between the actual and predicted values.
3. Which variables have the highest coefficients (parameters) in the linear regression model? Consider their absolute values, ignoring the direction of influence.
4. Instead of feeding the raw data to the linear regression model, develop the model on normalized data. To normalize the data, feed it to the Preprocess widget, and use the preprocessing by "Normalize Feature" (see the screenshot of this widget with the proper setting below). Normalization will subtract the mean from all columns and divide them by their standard deviations. Then train the linear model on the output data from Preprocess widget. Does normalization affect the coefficients of the linear model? Does normalization affect the predictive performance of the model? Explain!
5. Both sets of coefficients - those for items (3) and (4) - are useful for something. What can you read from the former (coefficients in question 3) that you can't from the latter (coefficients in question 4) and vice-versa?
6. Use Lasso regularization with a suitable strength to identify a subset of 3-5 most important independent variables. How well does this model perform in comparison with the one on all features? Explain the workflow that you used in sufficient detail, including the relevant settings in the widgets and the way the widgets are connected.
7. Show a scatter plot that relates the value of dependent variable and predictions for your most accurate predictor, and discuss the results in the figure. Rather than using Predict widget, use the outcome of Test & Score widget, which outputs the predicted values when a data instance was used in the test set. Use 10-fold cross-validation in Test & Score. In this way, you will avoid reporting results that would be the outcome of overfitting.

Submit your homework as a short report in PDF (not Word!) where you answer the above questions in the same order as the questions above. Please number your answers (1 to 7). Use 11 pt Arial or non-serif font. Limit the number of pages in the report to two. All figures should be captioned and referred to in the text. Make sure the figures are readable. Include only the visualizations, not the screenshots that include images of windows, programs, or desktops. Do not skew the images by changes the aspect ratio. Name your PDF document as lastname-firstname-2.pdf (like zupan-blaz-2.pdf; notice there are no spaces in the name, all letters are lowercase, and the dash is used to separate the first and last name) where the last name is your last name, and the firstname is your first name. Email the report to bzupan@gmail.com with subject DM-HW2 (copy the subject title and then paste it into the email title field; notice there are no spaces in the subject title).

The deadline for this homework is 8:00 am on Wednesday, December 23.