Homework 2 - Solution

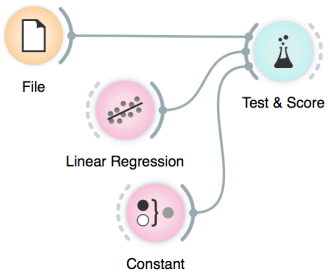
Task

Download the data about body fat measurements from http://file.biolab.si/files/fat.xlsx. The data is described 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 with a Scatter plot to show the relation between the actual and predicted values.
3. Which variables have the highest coefficients 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, pass is through a widget called Continuize and select "Normalize by standard deviation" under "Numeric features". This will subtract the mean from all columns and divide them by their standard deviations. Does it affect the coefficients? Does it affect the 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 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 for the relation between the strongest predictor and the outcome and comment the figure?

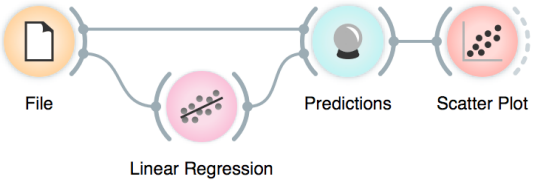
Solution

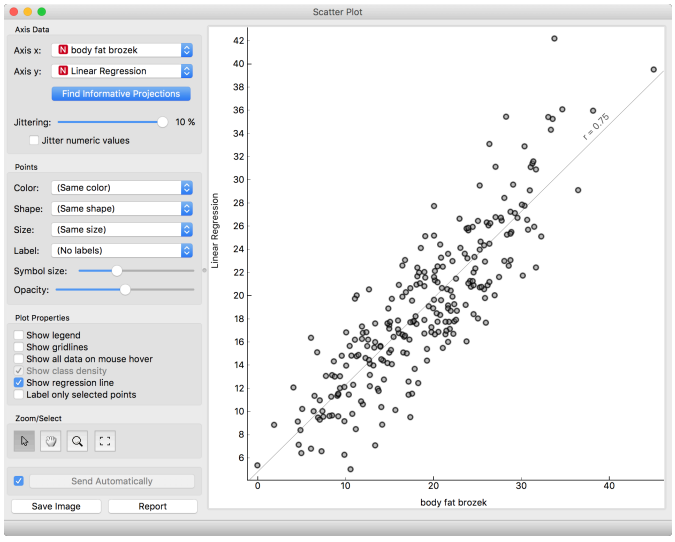
1. We already know accuracy needs to be assessed on the test data sets, and that procedures such as cross-validation are the most suitable ones to estimating the accuracy of a particular method. On our data set, linear regression is substantially more accurate than the baseline classifier (RMSE of 4.40 vs. 7.76, and R2 of 0.676). Note: we took



care that linear regression models were not regularized (corresponding setting in the Linear Regression widget).

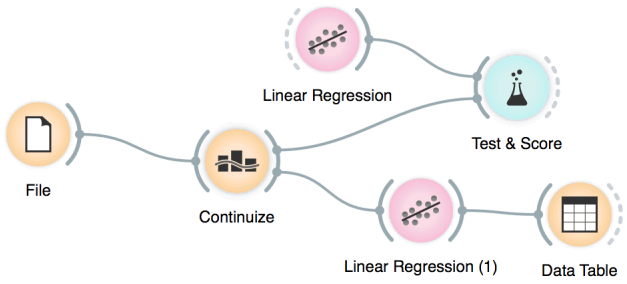
2. This time we train the model on the entire data set and show model-estimated values vs. true class values in the scatterplot. Note that if we compute correlation between these two measurements, the R2=0.75 is higher than the one estimated by cross-validation (0.67). Why?





3. Abdomen (0.878).

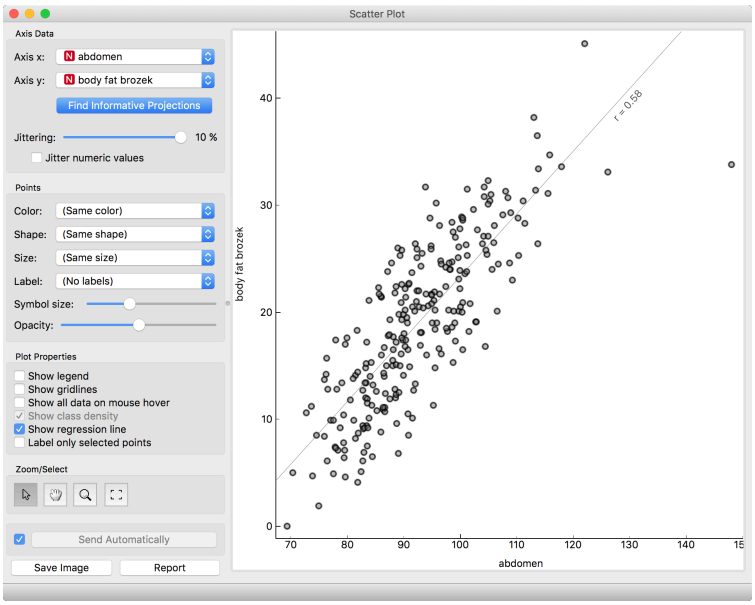
4. Coefficients of linear regression model change. Abdomen is still the feature with highest coefficient, but but the order of the features has changed. No effect on accuracy; this was expected, as linear model should be agnostic on linear transformations of individual features (scale the feature by 4.2, and the corresponding coefficient in the linear regression should decrease by the same factor).



5. Coefficients from (3) are useful when computing the body fat. Those from (4) tell us about the importance of each feature in the model.

6. Same workflow as in (4) was used. First, we used Linear Regression (1) - the lower branch of the workflow, to find the level of regularization that leaves us with four non-zero coefficients (disregarding the intercept). This regularization strength is at about 0.5. Now we set the same strength of regularization in the other Linear Regression widget, and observe very similar cross-validated performance (R2 at about 0.67). Great! We can actually build a simpler model that is as accurate as the one that includes all features.
   Note: workflow from (4) can actually use a single Linear Regression widget, where we use both of its output channels, one for model from the training data set, and the other one for passing the learning method to the Test & Score.

7. Strongest predictor is abdomen. People with bigger abdomen size tend to have more body fat. The relation with body fat is almost linear, with some outliers. Explained variance is at R2=0.58; we need other features to achieve best accuracy with linear model.


Scatter Plot
Axis Data
Axis x: abdomen
Axis y: body fat brozek
Find Informative Projections
Jittering: 10 %
Jitter numeric values
Points
Color: (Same color)
Shape: (Same shape)
Size: (Same size)
Label: (No labels)
Symbol size:
Opacity:
Plot Properties
Show legend
Show gridlines
Show all data on mouse hover
Show class density
Show regression line
Label only selected points
Zoom/Select
Send Automatically
Save Image
Report
body fat brozek
abdomen
r = 0.53
0
10
20
30
40
70
80
90
100
110
120
130
140