Elements of Statistics and Econometrics - 2021

Assignment 4

Problem 6: Regression techniques

The overall objective is to explain the sale prices by a set of variables using the data set from the previous assignments. The variable `SalePrice` is taken as the dependent variable and the remaining variables as explanatory.

1. The ridge and lasso regressions are regularisation techniques used to stabilize the estimation.

   (a) Explain in your own words the idea of the lasso regression. Sketch a situation when a simple linear regression fails, but the lasso regression still can be estimated. Why do we need to maximize the objective function numerically (in contrary to ridge).

   (b) For the usual regression model the variables are rarely normalized/standardized. However, in the case of the lasso regression the scaling becomes crucial. Why? Scale your data. Can/should the binary variables be scaled in the same fashion? What about factor variables (for example, `FireplaceQu`)?

   (c) Run a lasso regression with $\lambda > 0$. Plot the estimated parameters as functions of λ. Which value of λ would you recommend? If it is easy to implement, then determine the optimal λ by cross-validation.

   (d) Consider the ridge regression. Estimate it for a fixed (and optimal if possible) value of λ. Compare the estimated parameters and their variances for ridge and for the multiple linear regression. What would you expect from such a comparison. Remember to fit the linear regression to scaled variables.

2. In the next step we model `SalePrice` using regression trees.

   (a) Let `GrLivArea` be the first variable used for splitting. Write down the corresponding optimization problem and explain how the optimization works.

   (b) (**optional**) Implement the optimization manually. Order the values of `GrLivArea`. Make a grid for potential split points in such a way, that every element of the grid is in the middle of two subsequent `GrLivArea` observations (i.e. $(x_{(i)} + x_{(i+1)})/2$). Compute the objective function for every element of the grid and plot it as a function of the split point. Compare your results with R/Python results (use this feature as a single variable and set number of splits to 1).

(c) Consider now the variable `OverallQual`. It takes 10 different ordered values from "very poor" to "very excellent". How would you determine the optimal split for this type of a variable? Provide details of your ideas. Implementation is not needed.

(d) Obviously you can get very long trees. Tree prunning helps to get trees of a reasonable size. Fit a CART to the data and prune it to have at most 10 splits.

(e) Find the value of the complexity parameter for the trained tree in the last question. The complexity parameter uniquely determines the tree. What does it imply? What happens to the tree if we sufficiently increase or decrease the complexity parameter? Check it for the given tree and data.

(f) Check the value of the improvement in the first split. Explain the idea of improvement and provide numerical expression how this improvement is computed for the first split.

(g) Compute the importance of the variables and compare the results to the importances from lasso (remaining variables) and linear regression.

(h) (**optional:**) Here we apply the bootstrap technique to estimate the distribution of the parameters of a linear regression. Consider the first 10 features only. For every step of the bootstrap draw with replacement a sample of the same size as the original sample. Estimate the parameters by OLS and store the parameters for `LotArea` and `Street`. Repeat this 1000 times. Subsequently plot the histogram (or KDE) for these two parameters and compare them with the usual normal density (with mean $\hat{\beta}$ and variance given by the diagonal element of $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$.) Is using bootstrap a reasonable approach here?

(i) (**optional**) Having the sample of the parameters compute the confidence interval defined by 2.5% and 97.5% quantiles of the sample. Compare to the confidence intervals based on the assumption of normality.

(j) Use bagging and random forest to predict the `SalePrice`. Plot the error as a function of the number of trees.

(k) Check computational times for both approaches. Why the random forest is typically faster and needs less trees? Check which approach is implemented in your software for variable importance. Compute these and compare to the results for lasso, single tree and LR.

3. Next we consider classify the houses by defining

$$Y_i = \begin{cases} 1, & \text{if } SalePrice > 13000 \\ 0, & \text{else} \end{cases}.$$

The aim is classify houses to these two classes.

(a) Estimate the logistic regression using a backward model selection with AIC as the selection criterion. To speed up the computation consider only numeric features in the initial model. Further work only with the final model.

(b) Consider the explanatory variable `LotArea`. Obviously its parameter cannot be interpreted in the same way as for a linear regression. Provide the correct interpretation using the estimated parameter and using odds.

(c) Select several variables that increase the probability of $Y = 1$ and several variables that decrease this probability? Is this consistent with economic intuition?

(d) Randomly pick up five houses. Determine their probabilities of having a price above 130000. Provide for the first of them the formula for the probability with inserted values of the parameters and variables. If you want to predict the membership in one of the two groups for a particular house, what is the simplest way to proceed using these probabilities?

(e) Compute the classification table (confusion matrix) and calculate the specificity and sensitivity. Provide verbal interpretation for the elements of the classification table and the performance measures (specificity, sensitivity, accuracy).

(f) It makes sense to change the threshold used for classification to improve the performance. This can be done using the ROC curve. Plot this curve and determine the optimal threshold. Explain the idea of the curve and of the AUC measure.

(g) Recompute the classification table, sensitivity and specificity for the new threshold. Provide interpretation of the obtained values. Compare the results with the original values. Is the procedure now more conservative or less conservative?