Big Data Methods for Economists
Seminar MOEC0482:
Exercises*

Massimo Mannino Lin Xu

March 5, 2021

These exercises use data from the `ISLR` R package. The package is available from CRAN. For each exercise, we specify the data set it is based on. For instance, in Exercise 1 you should use the data set `Auto`. This can be accessed by first installing and then loading the package with `library(ISLR)`. The `Auto` data is loaded by writing `data(Auto)`. The data is then contained in the `Auto` variable. In the case that you want to use Python, you may need to import data from R at first.

Each group just needs to finish **one** exercise, one-to-one correspondence so that a group assigned to Topic N has to solve Exercise N.

*Following closely the exercises in the textbook ISLR and online resources. Thank Prof. Rainer Winkelmann and Carlo Zanella for organizing these exercises in previous years.

Exercise 1

This question involves the use of multiple linear regression on the `Auto` data set.

Important: The variable `origin` is a factor variable. Make sure you create dummy variables for all the regressions.

(a) Produce a scatterplot matrix which includes all of the variables in the data set.

(b) Compute the matrix of correlations between the variables using the function `cor()`. You will need to exclude the `name` variable, which is qualitative.

(c) Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:

   i. Is there a relationship between the predictors and the response?

   ii. Which predictors appear to have a statistically significant relationship to the response?

   iii. What does the coefficient for the `year` variable suggest?

(d) Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

(e) Try a few different transformations of the variables, such as $\log(X)$, $\sqrt{X}$, X^2. Comment on your findings.

(f) Now you are interested in selecting a model. Perform *forward* and *backward* model selection based on AIC using all variables (except `name`). Report the model specification you obtain for both methods. How do they differ?

Exercise 2

This question should be answered using the Weekly data set, which is part of the `ISLR` package. This data is similar in nature to the `Smarket` data from this chapters lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

(a) Produce some numerical and graphical summaries of the `Weekly` data. Do there appear to be any patterns?

(b) Use the full data set to perform a logistic regression with `Direction` as the response and the five lag variables plus `Volume` as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

(d) Now fit the logistic regression model using a training data period from 1990 to 2008, with `Lag2` as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

(e) Next Repeat (c) and (d) using a Linear Probability Model. Compare the models.

Exercise 3

We will now perform cross-validation on a simulated data set.

(a) Generate a simulated data set as follows:

```
>   set.seed ( 1 )
>   y <- rnorm (100)
>   x <- rnorm (100)
>   y <- x - 2 * x ^ 2 + rnorm (100)
```

In this data set, what is n and what is p? Write out the model used to generate the data in equation form

(b) Create a scatterplot of X against Y. Comment on what you find.

(c) Set a seed of your choice, and then compute the LOOCV errors that result from fitting the following four models using least squares:

i. $Y = \beta_0 + \beta_1 X + E$

ii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + E$

iii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + E$

iv. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + E.$

Note you may find it helpful to use the `data.frame()` function to create a single data set containing both X and Y .

(d) Repeat (c) using another random seed, and report your results. Are your results the same as what you got in (c)? Why?

(e) Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.

(f) Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?

(g) Now perform k-fold Cross-Validation and compute the error for all four models from equation 4. Use $k = 10$. Comment on your results. What happens when you change the seed?

Exercise 4

In this exercise, we will predict the number of applications received using the other variables in the `College` data set.

(a) Split the data set into a training set and a test set.

(b) Fit a linear model using least squares on the training set, and report the test error obtained.

(c) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.

(d) Fit a lasso model on the training set, with λ chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

(e) Plot the mean squared errors obtained for different values of λ for the

   i. Ridge regression,

   ii. Lasso regression.

   Comment on the figures. How do they differ qualitatively? Why?

(f) Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these three approaches?

Exercise 5

In this exercise, you will further analyze the `Wage` data set considered throughout this chapter.

(a) Perform polynomial regression to predict `wage` using `age`. Use cross-validation to select the optimal degree d for the polynomial. Plot the CV error as a function of the degree. What degree do you choose, and how does this compare to the results of hypothesis testing using ANOVA? Make a plot of the resulting polynomial fit to the data.

(b) Fit a step function to predict `wage` using `age`, and perform cross-validation to choose the optimal number of cuts. Plot the CV error as a function of the number of cuts. What number of cuts do you choose? Make a plot of the fit obtained.

(c) Perform a cubic spline regression using cross-validation to select the optimal number of knots. Plot the CV error as a function of the number of knots. Make a plot of the fit obtained.

Exercise 6

We will seek to predict `Sales` using regression trees and related approaches, treating the response as a quantitative variable.

(a) Split the data set into a training set and a test set.

(b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?

(c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

(d) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the `importance()` function to determine which variables are most important.

(e) Use random forests to analyze this data. What test MSE do you obtain? Use the `importance()` function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.

(f) Fit a linear model to the training set. What test MSE do you obtain?

(g) Compare the regression tree methods to the simple linear least squares estimation. Comment.

Exercise 7

This problem involves the OJ data set which is part of the ISLR package.

(a) Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.

(b) Fit a support vector classifier to the training data using cost=0.01, with Purchase as the response and the other variables as predictors. Use the summary() function to produce summary statistics, and describe the results obtained.

(c) What are the training and test error rates?

(d) Use the tune() function to select an optimal cost. Consider values in the range 0.01 to 10.

(e) Compute the training and test error rates using this new value for cost.

(f) Repeat parts (b) through (e) using a support vector machine with a radial kernel. Use the default value for gamma.

(g) Repeat parts (b) through (e) using a support vector machine with a polynomial kernel. Set degree=2.

(h) Overall, which approach seems to give the best results on this data?

Exercise 8 [1]

(a) Create the data set on which we want to do a simple regression. Set the seed to 42, generate 200 random points between -10 and 10 and store them in a vector named X. Then, create a vector named Y containing the value of sin(x).

(b) Use a feed-forward neural network and the logistic activation which are the defaults for the package nnet. We take one number as input of our neural network and we want one number as the output so the size of the input and output layer are both of one. For the hidden layer, well start with three neurons. Its good practice to randomize the initial weights, so create a vector of 10 random values, picked in the interval [-1,1].

(c) Split data to a training set containing 75% of the values in your initial data set and a test set containing the rest of your data.

(d) Load the nnet package and use the function of the same name to create your model. Pass your weights via the Wts argument and set the maxit argument to 50. We want to fit a function which can have for output multiple possible values. To do so, set the linout argument to true. Finally, take the time to look at the structure of your model.

(e) Predict the output for the test set and compute the RMSE of your predictions. Plot the function sin(x) and then plot your predictions.

(f) The number of neurons in the hidden layer, as well as the number of hidden layer used, has a great influence on the effectiveness of your model. Repeat the exercises (c) to (e), but this time use a hidden layer with seven neurons and initiate randomly 22 weights.

[1] based on online exercises