Statistical Learning, Tutorato #4

Veronica Vinciotti, Marco Chierici

March 29, 2021

Exercise 1

In this exercise, we go back to the `Default` dataset from the first tutorial and we use the validation set approach to estimate the test error of two logistic regression models .

- Use the validation set approach to estimate the test error of a logistic regression model that predicts `default` from `income` and `balance`. Follow these steps:
  1. Split the data into training and validation;
  2. Fit a model using only the training observations;
  3. Obtain predictions on the validation set (ie compute the posterior probabilities and threshold them using a chosen threshold);
  4. Compute the validation set error.
  5. Repeat 1-3 a number of times, using different training/validation splits. If you can, try to write this as a for loop. Inspect the different test errors and discuss the results.
- Now consider a logistic regression model that predicts the probability of `default` from `income`, `balance`, and a dummy variable for `student`. Estimate the average test error using the validation set approach as before. Is there an advantage in including `student`?

Exercise 2

In this exercise, you will explore cross-validation in a regression context.

A) Simulate a data set using randomly generated numbers from a normal distribution with mean 0 and variance 1, as follows:

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

Write down the model used to generate data in equation form. What is n and what is p?

B) Create a scatterplot of x vs y. Comment on the relationship that you see.

C) Set the random seed and compute the leave-one-out cross-validation (LOOCV) errors that result from fitting the following four models using least squares:

1. $y = \beta_0 + \beta_1 x + \epsilon$
2. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$
3. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$
4. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \epsilon$

D) Repeat the above changing the seed and discuss the results, comparing with the earlier seed. Do you get something different? Why?

E) Which of the models evaluated in C had the smallest LOOCV error? Did you expect this? Why?

F) Inspect model (4) and the significance of the coefficient estimates resulting from fitting the model using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?

Hints

- You can store y and x in a data frame to ease model creation in points C-D.
- LOOCV can be used in combination with `glm()` using the function `cv.glm(data, glm.fit)` (library `boot`) on a `glm` object. In general, the function can be used for K-fold cross-validation: by default, the functions performs a LOOCV ($K = n$).
- Optional: you can also try to implement the function to calculate the LOOCV error by yourself by:
  1. Writing a loop over all the observations (1:n)
  2. Fitting the model on all data minus the ith observation
  3. Predicting the observation left out.
  4. Do you get the same result as the cv.glm function?

Exercise 3

This is an exercise on subset selection in a regression context.

- Similarly to Exercise 2, generate a predictor x of length $n = 100$ and a noise vector ϵ of the same size, using the `rnorm()` function.
- Generate a response vector y of length $n = 100$ from the model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

  where you can freely choose the values for the constants β_i.
- We now pretend that we do not know the true model. On your simulated data, we try models with degrees varying between 1 and 10.
- Apply the "best subset selection" approach to select the best subset of the 10 predictors. The method is implemented by the `regsubsets()` function in the library `leaps`. This function has the same syntax as `lm()` and it returns the "best" model containing a subset of the given number of variables (input `nvmax`). What is the best model according to C_p, BIC, and adjusted R^2? You can find these values in the summary of the fitted `regsubsets` object, with the names `cp`, `bic`, and `adjr2`, respectively. Choose the subset that optimizes these statistics (max or min depending on the statistic). Show some plots to support your answer and report the coefficients of the best model.
- Repeat the previous task using forward stepwise selection and backwards stepwise selection, ie more efficient methods that do not search through the whole space of models. Both methods are implemented by `regsubsets()`, using the input `method="forward"` or `method="backward"`, respectively. Compare the answer from the stepwise methods with best subset selection.