Homework 5. Cross-Validation

Stat 427/627, Statistical Machine Learning

See due date on Canvas.

Contents

This assignment covers cross-validation (LOOCV, K-fold). Since cross-validation involves more computation, the running (and knitting) time may be much longer than usual.

Question	1	2	3	4	Total
427/627	6	10	12	2	30

1 Predicting a grade (CV in KNN) (6 pts)

Answer sub-questions (a) and (b) by hand calculation. A student wants to predict their grade for the Statistical Machine Learning course, using the KNN algorithm with $K = 3$. Six friends who took the course last year had the following mid-term test scores and grades.

Friend	1	2	3	4	5	6
Midterm	90	88	83	78	85	84
Course Grade	A	A	A	B	B	B

Estimate the prediction error rate of the algorithm, by means of:

(a) The validation-set method. Use Friends 2, 3, 4, 5 as the training data, and use Friends 1, 6 as the testing data.

(b) The leave-one-out cross-validation method.

(c) (Stat 627) Use `knn.cv()` function in package `class` to confirm your computation in (b). (You can start with reading the help file of `knn.cv()`.)

2 Ex.5.4.8. Cross-validation in linear regression on simulated data (p.222, 223, 10 pts.)

(a) Generate a simulated data set as follows:

```r
set.seed(1)
x <- rnorm (100)
y <- x - 2*x^2 + rnorm(100)
sim.df <- data.frame(x, y)  # data.frame will be helpful in part (b), (c)
```

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

(b) Compute the LOOCV estimates of prediction error that result from fitting each of the following four regression models: (Hints: (1) LOOCV is the same as K-fold CV with K=(sample size). (2) Function `glm( ..., family=gaussian, ...)` fits Normal linear regression (just like `lm()`). (3). Function `cv.glm()` in package `boot\` conducts cross-validation for the output objects from `glm()`.)

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X + \varepsilon \\
Y &= \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon \\
Y &= \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon \\
Y &= \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \varepsilon
\end{aligned}
$$

(c) Which of these models have the smallest adjusted prediction mean squared error as estimated by LOOCV? Is this what you expected? Explain your answer.

(d) Repeat step b but with $K = 10$-Fold validation and compare the prediction mean squared error. What do you notice compared to LOOCV?

- Use `set.seed(12)` before **each** `cv.glm()` call.

(e) (Stat-627) In part (c) (LOOCV), we did not use `set.seed()`. In part (d) (10-fold CV), we used `set.seed()`. The random seed is set so that we can get replicate the the same results for our homework practice. Why is the random seed relevant in the 10-fold CV but not in LOOCV?

3 Ex.5.4.5. Predicting defaults on loans (p.220, 221, 12 pts)

Use the `Default` data set in {ISLR2} package to create a logistic regression model for predicting the probability of variable `default` based on predictors `income`, `balance`, and `student`.

Use each of the following methods to estimate the *test error rate* of the logistic regression model and decide whether it will be improved if the dummy variable `student` is excluded from the prediction.

- Use a seed of 123 and a threshold of .5 where appropriate.

(a) The validation set approach with a 60% split. I.e. split the data set only once, 60% of the observation will be used for training, and the the remaining 40% will be used for validation/testing.

(b) Leave-one-out cross-validation. (Your computer may take a really long time to run the code on LOOCCV due to the large sample size. Considering using a chunk option for cache, e.g., set `{r, cache=TRUE}`.)

(c) K-fold cross-validation for $K = 100$ and $K = 1000$.

4 Cross-validation in LDA and QDA. (2 pts)

There is an example of cross-validation in LDA/QDA in one of the R handouts. Locate that handout, and find the example. Using the explanations in the handout, or the function's help file, or online resources, determine whether the cross-validation used in the example is K-fold cross-validation or Leave-One-Out cross-validation (LOOCV).

—— This is the end of HW 5. ——