Stat 427/627, Statistical Machine Learning

Homework 3

Due: Friday, June 7, 2024

Contents

- This assignment covers cross-validation (LOOCV, K-fold), Jackknife and Bootstraping.
- Finish Q.1 - Q.3 after Tuesday's class, and the rest after Thursday's class.
- 46 Points

Question	1	2	3	4	5	6	Total
427	6	10	12	2	6	10	46
627	6	10	12	2	6	10	46

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

Do by hand. 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, using Friends 2, 3, 4, 5 as training and Friends 1, 6 as 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:

```
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) Use function `glm()` fits Normal linear regression when you set `family=gaussian`. (3). Use function `cv.glm() in pacakge`boot‘.)

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

(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)

Refer the R example handouts. Find the example of cross-validation in LDA and QDA. Is it LOOCV or K-fold CV?

5 Basses and sopranos (Jackknife) (6 pts)

An acoustic studio needs to estimate the range of voice fundamental frequencies that an adult singer can produce. A sample of $n = 10$ recordings contains frequencies 102, 115, 127, 127, 162, 180, 184, 205, 239, 240.

(a) Manually compute (by hand) the jackknife estimator of the population lowest fundamental frequency of a human voice. Compare your results with the natural range of human voice frequencies. (Use Google or Wiki.)

(b) Use software to confirm your result.

(c) (Stat-627 only) Generalize the results. Assume a sample $X_1, \ldots, X_n$ of size n, where X_1, X_2 are the smallest two observations. Derive equations for the jackknife estimators of the population minimum. Use your result in (a) to verify your formula.

6 Ex.5.4.9. Bootstrap the mean of median house values in the `Boston` dataset. (p. 223, 10 pts)

We will now consider the `Boston` housing data set from the {`MASS`} library.

(a) Based on this data set, provide an estimate for the population mean μ of `medv`, which is the median value of owner-occupied homes in \$1,000s. Call this estimate $\hat{\mu}$.

(b) Estimate the standard error of $\hat{\mu}$ (as we know, $\text{StdError}(\bar{x}) = s/\sqrt{n}$, where s is the sample standard deviation. R function `sd()`.)

(c) Estimate the standard error of $\hat{\mu}$ using the bootstrap method. How does this compare to your answer from (b)?. Remember to set your seed to be reproducible.

(d) Based on your bootstrap estimate from (c), provide a 95% confidence interval for μ. A popular approximation is $\hat{\mu} \pm 2\text{StdError}(\hat{\mu})$. Compare it to the results obtained using an R command `t.test(Boston$medv)`.

(e) Now, estimate M, the population median of `medv` with the sample median $\hat{M}$.

(f) We now would like to estimate the standard error of $\hat{M}$, but unfortunately, there is no simple formula for computing the standard error of a sample median. Instead, estimate this standard error using the bootstrap method.

—— **This is the end of HW 3.** ——