1. Description

▶ 모든 데이터로 모델을 훈련시키면 과소적합인지, 적정적합인지, 과적합인지 가늠하기가 힘들다. 훈련을 시키면 시킬수록 error rate 가 계속 줄어드는 경향이 있으므로 결국 과적합으로 이어지기 쉽다. 과적합을 피하기 위해 validation set 을 활용하는 법이 cross-validation 이라고 한다.

▶ 50%의 데이터는 검증을 위해 남기고, 다른 50%는 모델 학습을 위해 남겨 학습 시킨 후, 모델 검증은 validation data 로 하는 방법은 validation set approach 이라고 한다. 간단하고 빠르게 동작할 수 있지만 가장 큰 단점으로는 매번 다른 random set 을 뽑을 때마다 결과가 달라질 수 있다.

▶ Leave-One-Out CV(LOOCV) 방식은 총 N 번의 model 을 만들고, 각 모델을 만들 때 하나의 샘플만 제외하면서 그 제외한 샘플로 test set performance 를 계산하여 N 개의 performance 에 대해서 평균을 내는 방법이다. 거의 모든 데이터로 모델링해서 bias 가 적고 안정적인 결과를 얻을 수 있지만, 그만큼 많은 수의 model 을 만들고 test 해야 하기 때문에 시간이 오래 걸리는 단점이 있다.

▶ k-fold cross validation 은 전체 데이터를 K 개의 부분집합으로 나눈후, K-1 개 집합으로 학습하여 회귀분석 모형을 만들고 K 번째 집합으로 교차검증을 한다. 이 과정을 K 번 실시하고 K 개의 교차검증 성능을 평균하여 최종 교차검증 성능을 계산한다. K 가 작을수록 모델의 평가는 편중되고, K 가 높을수록 bias 는 낮아지지만 결과의 분산이 높을 수 있다.

2. Implement

Lab : Cross-Validation and the Bootstrap

- ISLR library안에 내장 데이터인 Auto를 이용하여 test 데이터의 error rates로 검정하는 절차를 거쳤다. sample() 함수를 이용하여 train, test 데이터로 나누고 train 데이터로 model을 fitting시킨 후, test 데이터로 mse를 구하였다.
- 다항식을 표현하고 싶을 때는 모형 문자열에 poly()함수를 이용하여 표현한다.
- LOOCV를 추정할 때는 glm(), cv.glm() 함수를 쓰는데 family를 설정해주지 않으면 lm()과 같은 결과를 낸다.
- k-fold CV를 적용하고 싶을 때는 cv.glm(model, K)을 사용한다.
- boot()함수를 이용하여 데이터를 복원추출하여 반복적으로 샘플링하여 bootstrap의 결과를 낼 수 있다.

5. In Chapter 4, we used logistic regression to predict the probability of default using income and balance on the Default data set. We will now estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.

(a) Fit a logistic regression model that uses income and balance to predict default.

Logistic regresssion에 fitting 시킨 결과, income과 balance 모두 유의미한 변수로 나타났다.

glm(default ~ income + balance, family=binomial, data=Default)					
Coefficients :					
	Estimate	Std.Error	t-value	p-value	
(intercept)	-1.154e+01	4.348e-01	-26.545	<2e-16	***
Income	2.081e-05	4.985e-06	4.174	2.99e-05	***
balance	5.647e-03	2.274e-04	24.836	<2e-16	***
AIC : 1585					

(b) Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:

i. Split the sample set into a training set and a validation set.

ii. Fit a multiple logistic regression model using only the training observations.

iii. Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the default category if the posterior probability is greater than 0.5.

iv. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

	No	Yes
No	2890	68
Yes	12	30
Test error : 0.0266		

Default 데이터를 train:test = 7:3으로 나누었다. Train 데이터로 모델링 후, test 데이터를 fitting시켜 확률이 0.5보다 크면 "Yes", 작으면 "No"라 하였을 때, confusion matrix는 다음과 같다. 실제 "No"인 2902 중에 "No"라고 예측한 것이 2890개, 실제 "Yes"인 98개 중에 "Yes"라고 예측한 것이 68개이다. 오분류율은 0.0266이다.

(c) Repeat the process in (b) three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.

	No	Yes
No	2888	70
Yes	11	31
Test error : 0.027		

	No	Yes
No	2901	64
Yes	12	23
Test error : 0.0253		

	No	Yes
No	2877	79
Yes	9	35
Test error : 0.0293		

Sample로 train, test 데이터를 나누기 때문에, 나누어진 데이터에 따라서 fitting된 모델도 달라지므로 결과는 달라질 수 있다. 같은 비율로 데이터를 나누어 3번 시행했을 때, 예측한 값은 정확하게 같지 않지만 오분류율을 보면 비슷하다는 것을 확인 할 수 있다.

(d) Now consider a logistic regression model that predicts the probability of default using income, balance, and a dummy variable for student. Estimate the test error for this model using the validation set approach. Comment on whether or not including a dummy variable for student leads to a reduction in the test error rate.

	No	Yes
No	2889	71
Yes	12	28
Test error : 0.0276		

student 변수를 추가하였을 때, 오분류율은 0.0276으로 변수가 없을 때와 비교하였을 때, test error가 낮아지지 않았다. 따라서 dummy 변수인 student가 test error를 줄이는데 영향을 주는 변수가 아니라고 할 수 있다.

7. In Sections 5.3.2 and 5.3.3, we saw that the cv.glm() function can be used in order to compute the LOOCV test error estimate. Alternatively, one could compute those quantities using just the glm() and predict.glm() functions, and a for loop. You will now take this approach in order to compute the LOOCV error for a simple logistic regression model on the Weekly data set. Recall that in the context of classification problems, the LOOCV error is

given in (5.4).

(a) Fit a logistic regression model that predicts Direction using Lag1 and Lag2.

glm(formula = Direction ~ Lag1 + Lag2, family = binomial, data=Weekly)

Coefficients :

	Estimate	Std.Error	t-value	p-value	
(intercept)	0.2212	0.0614	3.599	0.0003	***
Lag1	-0.0387	0.0262	-1.477	0.1396	
Lag2	0.0602	0.0265	2.270	0.0232	*

AIC : 1494.2

Lag1과 Lag2를 이용하여 Direction을 예측하는 logistic regression 모델에서 Lag2만이 약간 유의미한 변수로 나왔다.

(b) Fit a logistic regression model that predicts Direction using Lag1 and Lag2 *using all but the first observation*.

glm(formula = Direction ~ Lag1 + Lag2, family = binomial, data=Weekly)

Coefficients :

	Estimate	Std.Error	t-value	p-value	
(intercept)	0.2232	0.0615	3.630	0.0002	***
Lag1	-0.0384	0.0262	-1.466	0.1426	
Lag2	0.0608	0.0265	2.291	0.0219	*

AIC : 1492.5

첫번째 관측치를 제외하고 (a)와 같이 모델을 fitting시켰을 때 결과는 다음과 같다.

(c) Use the model from (b) to predict the direction of the first observation. You can do this by predicting that the first observation will go up if P(Direction="Up"|Lag1, Lag2) > 0.5. Was this observation correctly classified?

(b)에서 만들어진 모델에 첫번째 관측치의 결과를 예측하였을 때, 확률은 0.5713으로 0.5보다 크므로 Direction="Up"으로 할당되었지만 실제 값은 "Down"이었다.

(d) Write a for loop from i = 1 to i = n, where n is the number of observations in the data set, that performs each of the following steps:

i. Fit a logistic regression model using all but the ith observation to predict Direction using Lag1 and Lag2.

ii. Compute the posterior probability of the market moving up for the ith observation.

iii. Use the posterior probability for the ith observation in order to predict whether or not the market moves up.

iv. Determine whether or not an error was made in predicting the direction for the ith observation. If an error was made, then indicate this as a 1, and otherwise indicate it as a 0.

0	1
605	484

i번째 관측치를 제외하고 모델을 fitting시킨 후, i번째 관측치로 만들어진 모델에서의 확률을 구하여 Direction이 "Up"인지 "Down"인지 할당하였다. 전체 관측치의 개수만큼 반복시켜 실제 값과 같으면 0, 다르면 1이라고 하였다. 전체 1089개의 관측치 중, 맞게 예측한 값은 605개였다.

(e) Take the average of the n numbers obtained in (d)iv in order to obtain the LOOCV estimate for the test error. Comment on the results.

LOOCV 를 이용하였을 때 test error 는 0.444 로 44.4%이다.

9. 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 of medv. Call this estimate $\hat{\mu}$

변수 medv의 평균은 22.5328로 $\hat{\mu} = 22.5328$이라고 하자.

(b) Provide an estimate of the standard error of $\hat{\mu}$. Interpret this result.

$\hat{\mu}$의 표준오차는 변수 medv의 표준오차에서 Boston의 관측치 수를 나눈 값인 0.4088611이다.

(c) Now estimate the standard error of $\hat{\mu}$ using the bootstrap. How does this compare to your answer from (b)?

R=1000일 때, bootstrap으로 추정된 $\hat{\mu}$의 표준오차는 0.4119로 (b)에서 추정한 0.4089과 매우 근접하다.

(d) Based on your bootstrap estimate from (c), provide a 95% confidence interval for the mean of medv. Compare it to the results obtained using t.test(Boston$medv).

```
##  One Sample t-test
## data:  medv
## t = 55.1111, df = 505, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  21.72953 23.33608
## sample estimates:
## mean of x
##  22.53281
```

t.test()함수를 이용한 medv 평균의 95% 신뢰구간은 위와 같다. bootstrap으로 추정된 신뢰구간은 21.7062 23.3538으로 t.test()함수에서 얻어진 결과와 거의 근접하다.

(e) Based on this data set, provide an estimate, $\hat{\mu}_{med}$, for the median value of medv in the population.

변수 medv의 중앙값은 21.2로, $\hat{\mu}_{med} = 21.2$라 하자.

(f) We now would like to estimate the standard error of $\hat{\mu}_{med}$. Unfortunately, there is no simple formula for computing the standard error of the median. Instead, estimate the standard error of the median using the bootstrap. Comment on your findings.

median의 standard error를 구하는 식은 없지만 bootstrap을 이용하여 가정 없이 표본에서 데이터를 반복적으로 추출하여 median을 구해 median의 평균의 분포와 standard error인 0.368을 알 수 있게 된다.

(g) Based on this data set, provide an estimate for the tenth percentile of medv in Boston suburbs. Call this quantity $\hat{\mu}_{0.1}$.

변수 medv의 하위 10% 값은 12.75로, $\hat{\mu}_{0.1} = 12.75$라 하자.

(h) Use the bootstrap to estimate the standard error of $\hat{\mu}_{0.1}$. Comment on your findings.

10% 값의 standard error를 구하는 식은 없지만 bootstrap을 이용하여 가정 없이 표본에서 데이터를 반복적으로 추출하여 10% 값을 구해 평균의 분포와 standard error인 0.499을 알 수 있게 된다.

3. Discussion

▶ 4.2에서는 주어진 관측치가 bootstrap의 샘플에 포함될 확률을 구해보았다. j번째 관측치가 bootstrap 샘플에 포함되지 않을 확률은 $(1-1/n)^n$ 이고, 포함될 확률은 1에서 포함되지 않을 확률을 뺀 값이므로 1-$(1-1/n)^n$ 이다. 이는 데이터의 사이즈가 커지면 bootstrap의 샘플에 포함될 확률이 작아진다는 것을 의미하고 확률 값은 0.632에 근사함을 배웠다.

▶ 4.5에서는 모델 검증 방법 중 하나인 *validation set approach*에 대해 예제와 함께 학습해보았다. train과 test를 7:3으로 나누어 train 데이터로 모델을 학습시킨 후, 모델을 test 데이터를 이용하여 사후 분포를 구해 0.5를 기준으로 class를 할당하였다. 실제 값과 비교하여 오분류율을 구하고, 같은 방법을 반복시켜 오분류율을 비교해보았다.

▶ 4.7에서는 cross validation의 방법 중 하나인 *LOOCV*에 대한 예시에 대해 학습해보았다. Weekly데이터를 이용하여 한 개의 관측치를 뺀 데이터로 학습시킨 후, 제외시킨 하나의 데이터로 test에 사용하여 예측치를 확인하였다. 예측치가 실제 값과 같으면 0, 다르면 1로 할당하여 그 값들의 평균이 LOOCV의 test error가 되었다.

▶ 4.9에서는 변수 medv의 평균, 중앙값, 하위 10%의 값의 분포를 알아보았다. 평균의 standard error는 쉽게 구할 수 있지만 중앙값, 하위 10%의 값의 분포는 쉽게 알 수 없으므로 *bootstrap*을 이용하였다. 확률변수의 분포를 모르는 경우, bootstrapping을 사용하여 중복 허용하여 데이터를 뽑고, 그들의 평균을 여러 번 구하여 평균의 분포를 구할 수 있게 되고, 이로부터 95% 확률로 sample mean이 사이의 구간의 위치하는 신뢰구간을 알 수 있게 되었다.