Øving 4

2022-04-29

Problem 1

Explain how k-fold cross validation is implemented

b) Specify what is done, and in particular how the results from each fold are aggregated.

We split the dataset into K sets, with equal amounts of samples n_k in each. One is chosen as the validation set and the rest is used for training. The combination that makes the best fit is chosen. We calculate the MSE for each fold, and the cross-validation error is given by

$$CV_k = \frac{1}{n} \sum_{j=1}^{k} n_j \cdot MSE_j$$

Where n_j is the size of the n´th fold. And the MSE is given by

$$MSE_j = \frac{1}{n_j} \sum_{i \in V_j} (y_i - \hat{y}_i)^2$$

c) Relate to one example from regression.

Finding the optimal number of neighbours k in KNN regression. Finding the optimal degree in polynomial regression.

d) Relate to one example in classification.

K in KNN classification. Test/train split in logistic regression.

Problem 2

What are the advantages and disadvantages of k-fold cross-validation relative to

a) The validation set approach

This method is essentially just doing a split into test/train. The advantage is that k-fold CV optimizes the problem, we can gind the split that gives us the most optimal model, we get a model with less variance and less bias. However, for large datasets this is much more computational expensive than just a random split.

b) Leave one out cross-validation (LOOCV)

The LOOCV is the k-fold where k = n, which is an extreme version of normal k-fold CV and has the maximum computational cost. This results in a reliable and unbiased estimate of model performance. The disadvantage of this approach holds for large datasets and computationally expensive models.

k-fold has less variance than LOOCV (the variance in model selection), because in LOOCV we are averaging from n fitted models that are trained on nearly the same data, therefore we have positively correlated data.

k-fold has more bias, as in LOOCV we use a larger data set to fit the model, which gived a less biased version of the test error.

c) What are recommended values for k and why?

Experimental simulations has shown that K = 5 or K = 10 is optimal.

Very large k will lead to the estimator of test error having high variance and low bias, and is also very computational expensive.

For small k the estimator will have larger bias but lower variance.

Problem 4

We will calculate the probability that a given observation in our original sample is part of a bootstrap sample.

Our sample size is n.

a) We draw one observation from our sample. What is the probability of drawing observation i? And of not drawing observation i?

The probability of drawing observation i is simply 1 to n.

$$P(x = x_i) = \frac{1}{n}$$

The probability of not drawing observation i is

$$P(x \neq x_i) = 1 - P(x = x_i) = 1 - \frac{1}{n} = \frac{n-1}{n}$$

b) We make n independent drawing with replacement. What is the probability of not drawing observation i in any of the n drawings? What is then the probability that data point i is in our bootstrap sample?

The probability of not drawing observation i in any of the n drawing is

$$P(x \neq x_i) = (1 - \frac{1}{n})^n$$

The probability of the data point i being in our bootstrap sample is

$$P(X \ in \ sample) = 1 - P(x \neq x_i) = 1 - (1 - \frac{1}{n})^n$$

c) When n is large, we have that $(1 - \frac{1}{n})^n \approx \frac{1}{e}$. Use this to give a numerical value for the probability that a specific observation i is in our bootstrap.

$$P(X \ in \ sample) = 1 - P(x \neq x_i) = 1 - (1 - \frac{1}{n})^n \approx 1 - \frac{1}{e} = 0.632$$

d) Write a short R code to check your results.

```
n = 100
B = 10000

j = 1

res = rep(NA, B)

for(b in 1:B){
  res[b] = (sum(sample(1:n, replace=TRUE)== j)> 0)
}
mean(res)
```

```
## [1] 0.63
```

Problem 5

Explain with words and an algorithm how you would proceed to use bootstrapping to estimate the standard deviation and the 95% CI of one of the regression parameters in multiple linear regression. Comment on which assumptions you make for your regression model.

I would first choose the number of boots B. Then I would proceed to draw with replacement a bootstrap sample, fit my model and calculate the estimator $\hat{\beta}_b$ and store this. Then I would proceed for every b up to B and store all of the $\hat{\beta}_1.....\hat{\beta}_B$. I would then calculate $\hat{SD}(\hat{\beta})by$:

$$\hat{SD}(\hat{\beta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\beta}_b - \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b)^2}$$

For the 95% CI, we can calculate the 0.025 and 0.975 quantiles of the sample $\hat{\beta}_b$.

For the multiple linear model we assume that we have a linear relationship, multivariate normality, no or little multicollinearity, no auto-correlation and homoscedasticity.

Problem 6

Implement your algorithm from task 5 both using for-loop and using the boot function. Use out SLID data set and provide standard errors for the coefficient for age. Compare wih the theoretical value $(X^TX)^{-1}\hat{\sigma}^2$.

```
#install.packages("car")
library(car)
library(boot)

SLID = na.omit(SLID)
n = dim(SLID)[1]

SLID.lm = lm(wages ~ ., data = SLID)
summary(SLID.lm)$coeff["age",]
```

```
##      Estimate    Std. Error       t value      Pr(>|t|)
##  2.551368e-01  8.714144e-03  2.927847e+01 7.822688e-171
```

```
confint(SLID.lm)
```

```
##                     2.5 %     97.5 %
## (Intercept)    -9.0891576 -6.6884008
## education       0.8484610  0.9847670
## age             0.2380522  0.2722214
## sexMale         3.0452719  3.8655493
## languageFrench -0.8518577  0.8214111
## languageOther  -0.4946904  0.7798996
```

```
B = 100

beta_hat = rep(NA, B)

#IKKE ferdig
```