Resampling Methods Exercises

Slides on *Introduction to Statistical Learning*, Chapter 5

Calvin Khor

January 2024

Exercise 1

Using basic statistical properties of the variance, as well as single- variable calculus, derive (5.6). In other words, prove that α given by (5.6) does indeed minimize $\mathrm{Var}(\alpha X + (1-\alpha)Y)$. (5.6) is:

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}.$$

Answer Simply note that

$$\begin{aligned}
\mathrm{Var}(\alpha X + (1-\alpha)Y) &= \alpha^2\sigma_X^2 + (1-\alpha)^2\sigma_Y^2 + 2\alpha(1-\alpha)\sigma_{XY} \\
&= A\alpha^2 + B\alpha + C \\
&= A\left(\alpha + \frac{B}{2A}\right)^2 + \tilde{C}
\end{aligned}$$

where $A = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}$ and $B = -2\sigma_Y^2 + 2\sigma_{XY}$, QED.

Exercise 2

We will now derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap sample from a set of n observations.

(a) What is the probability that the first bootstrap observation is not the jth observation from the original sample? Justify your answer.

(b) What is the probability that the second bootstrap observation is not the jth observation from the original sample?

(c) Argue that the probability that the jth observation is not in the bootstrap sample is $\left(1 - \frac{1}{n}\right)^n$.

Answer

(a) $1 - \frac{1}{n}$ since the bootstrap observation is uniformly randomly chosen.

(b) Its sampling with replacement, so $1 - \frac{1}{n}$.

(c) $\mathbb{P}(j\text{th obs not in boot sample})$ is the product as each bootstrap sample is independent.

Exercise 2 cont. 1

(d) When $n = 5$, what is the probability that the jth observation is in the bootstrap sample?

(e) When $n = 100$, what is the probability that the jth observation is in the bootstrap sample?

(f) When $n = 10\,000$, what is the probability that the jth observation is in the bootstrap sample?

Answer

(d) $\left(1 - \frac{1}{5}\right)^5 \approx 0.32768000000000002.$

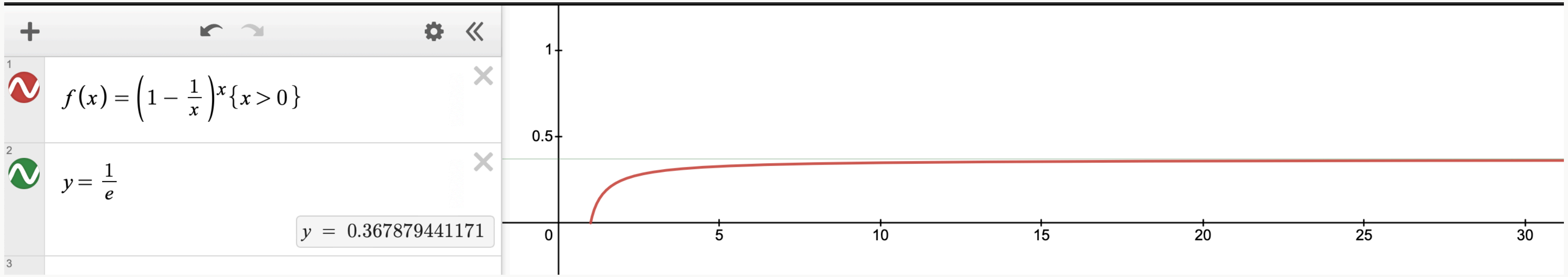
(e) $\left(1 - \frac{1}{100}\right)^{100} \approx 0.3660323412732289.$

(f) $\left(1 - \frac{1}{10000}\right)^{10000} \approx 0.36786104643302414.$

Exercise 2 cont. 2

(g) Create a plot that displays, for each integer value of n from 1 to $100\,000$, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.

Answer

(g) 

(see Desmos plot.) I observe convergence to $\frac{1}{e}$. Of course, this is because

$$\left(1 - \frac{1}{n}\right)^n = e^{n \log\left(1 - \frac{1}{n}\right)} = e^{n\left(-\frac{1}{n} - \frac{1}{2n^2} + O\left(\frac{1}{n^3}\right)\right)} = e^{-1 - \frac{1}{2n} + O\left(\frac{1}{n^2}\right)} = \frac{1}{e} - \frac{1}{2en} + O\left(\frac{1}{n^2}\right).$$

Exercise 2 cont. 3

(h) We will now investigate numerically the probability that a bootstrap sample of size $n = 100$ contains the jth observation. Here $j = 4$. We first create an array store with values that will subsequently be overwritten using the function `np.empty()`. We then repeatedly create bootstrap samples, and each time we record whether or not the fifth observation is contained in the bootstrap sample.

```
rng = np.random.default_rng(10)
store = np.empty(10000)
for i in range(10000):
    store[i] = np.sum(rng.choice(100, size=100, replace=True) == 4) > 0
np.mean(store)
```

(NB typo corrected) Comment on the results obtained.

Answer We get `0.6362`. The bootstrap sample size is $10\,000$, so the true probability is $1 - 0.36786104643302414 = 0.6321389535669759$ which is consistent.

Exercise 3

We now review k-fold cross-validation.

(a) Explain how k-fold cross-validation is implemented.

(b) What are the advantages and disadvantages of k-fold cross-validation relative to:

  i. The validation set approach?

  ii. LOOCV?

Answer

(a) Split the data randomly into k bins. For each bin, train on the other $k-1$ bins and then test on the kth bin. Then report the average test error across bins $\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i$.

(b) i. **Pros** Helps prevent fitting to the particular split

  Cons not repeatable (unless seed fixed), more compute needed

  ii. **Pros** usually less compute needed, less variance (for small k)

  Cons not repeatable, higher bias

Exercise 4

Suppose that we use some statistical learning method to make a prediction for the response Y for a particular value of the predictor X. Carefully describe how we might estimate the standard deviation of our prediction.

Answer We use the bootstrap. First we use our method to make some small number n of predictions Y_i. Then we use these to create $N \gg n$ bootstrap samples $Y_i^* = (Y_{i1}^*, ..., Y_{in}^*)$. Finally, our estimate for the standard deviation is the average of the bootstrap standard deviations,

$$\sigma_Y \approx \frac{1}{N}\sum_{i=1}^{N} \sigma_{Y_i^*} = \frac{1}{N}\sum_{i=1}^{N} \sqrt{\frac{1}{n}\sum_{j=1}^{n} Y_{ij}^{*2} - \left(\frac{1}{n}\sum_{j=1}^{n} Y_{ij}^*\right)^2}.$$