Chapter 8 lecture questions

Q1: After you completed the 5-fold cross-validation, some new data have become available. How would you make predictions with the new data?

Answer:
With 5-fold cross-validation, you built 5 models, one for each validation period, so there are 5 models. Now you have new data and you want to make predictions. The best way is to use all 5 models to predict, then ensemble average the predictions from the 5 models to get a single prediction value.

Q2: For bootstrap resampling applied to a dataset with N observations, derive an expression for the fraction of data in the original dataset drawn in an average bootstrap sample. What is this fraction as $N \to \infty$? [Hint: $\left(1 - \frac{1}{N}\right)^N \to \mathrm{e}^{-1}\,, \quad \text{as } N \to \infty$]

Answer: $1 - \mathrm{e}^{-1} = 1 - 0.368 = 0.632$. On average, 63.2% of the original data is drawn, while 36.8% is not drawn into a training dataset under bootstrap resampling.

Solution:
When selecting a data point from the data set, the probability for a given data point to be selected is $1/N$, hence the probability of not being selected is $1 - 1/N = (N-1)/N$. A bootstrap sample involves making N independent selections, hence the probability of not being selected for N times is $[(N-1)/N]^N$, which is also the fraction of data from the original data set not selected in a bootstrap sample. The fraction of original data selected in a bootstrap sample is then $1 - [(N-1)/N]^N$.

Next, we look at the limiting situation as $N \to \infty$. Fraction of data not selected is

$$\left(\frac{N-1}{N}\right)^N = \left(1 - \frac{1}{N}\right)^N \to \mathrm{e}^{-1}\,, \quad \text{as } N \to \infty.$$

Thus the fraction of original data selected in a bootstrap sample is $1 - \mathrm{e}^{-1} = 1 - 0.368 = 0.632$.

Q3: Prove Cauchy's inequality for the special case $M = 2$, i.e. prove that

$$(\epsilon_1 + \epsilon_2)^2 \leq 2(\epsilon_1^2 + \epsilon_2^2).$$

Proof:
Let $a = \epsilon_1$ and $b = \epsilon_2$. Need to prove that

$$(a + b)^2 \leq 2(a^2 + b^2).$$

For any number $a - b$, we have $(a - b)^2 \geq 0$.
So $(a - b)^2 = a^2 - 2ab + b^2 \geq 0$, i.e. $a^2 + b^2 \geq 2ab$.
Adding $a^2 + b^2$ to both sides of this inequality gives $2a^2 + 2b^2 \geq a^2 + 2ab + b^2 = (a + b)^2$.
Hence, $2(a^2 + b^2) \geq (a + b)^2$.

Q4: You have 50 years of data, and you want to do a 25-fold cross-validation of your MLP NN model. You also want to run an ensemble of 30 runs with random initial weights for the NN model. How many NN model runs do you have to perform?

Answer:
You will need to do $25 \times 30 = 750$ NN model runs.