Chapter 5 Concept Test

Due Apr 22 at 12pm **Points** 1 **Questions** 4 **Time Limit** None
Allowed Attempts Unlimited

Instructions

Take this concept test after completing your pre-class preparation for week 5

Take the Quiz Again

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	26 minutes	1 out of 1

Score for this attempt: **1** out of 1
Submitted Apr 22 at 11:32am
This attempt took 26 minutes.

Question 1

0 / 0 pts

Examine figure 5.11 on page 190. The dataset from which the bootstrap samples will be generated contains 3 unique observations.

Consider a bootstrap sampling process which has *n* samples of the original data. How many possible unique bootstrap samples could be generated from the original data? Express your answer in terms of a function of *n,* and then compute the number of unique bootstrap samples that could occur given the original number of data samples is 3.

○ $2^n = 2^3 = 8$

○ $n^2 = 3^2 = 9$

○ $n! = 3! = 6$

Correct!

● $n^n = 3^3 = 27$

There will be *n* samples drawn during the bootstrap, and each of those samples is drawn with replacement, resulting in *n* options for each sample. Thus the correct answer is n^n

Question 2

0 / 0 pts

What are the valid uses of cross-validation and test? Select all that apply

Correct!

☑

Cross-validation can be used to estimate the comparative performance of models using multiple subsets of the training data. Cross validation is often used to find good settings for hyperparameters. After ML decisions made during cross validation are complete, the best model's test set performance can inform the expected performance on future unseen data.

Making machine learning decisions during validation and reporting expected performance on the test set is a standard practice.

□

Exploring the data is an important step in the machine learning pipeline. After exploring all the data to obtain insights for the machine learning approach, before building any models, a subset of the data should be identified as the test set and sequestered.

'ou Answered

☑

The boss wants to know which of two models (A and B) are better, and how well the chosen model will perform on unseen future data. If the performance of model A and B are each computed on the test set and model A performs better, select model A and report its performance on the test set since the test set is the best estimate for performance on future unseen data.

In a strict sense, deciding to use model A instead of model B is a machine learning decision. The test set should never be used to make machine learning decisions. Model-selection decisions such as choosing model A or B should be accomplished based on comparative performance during validation phase, using a process such as validation set, cross-validation or LOOCV. Only after all ML decisions have been made should the test set be used to estimate performance on future unseen data.

Correct!

☑

While one use of cross-validation is to estimate performance of multiple options to inform machine learning decisions, a valid alternative is to use cross-validation to estimate performance of a model on unseen future data as long as the cross-validation process was not used to make any other model decisions (such as hyperparameter or model selection).

Using cross-validation to estimate a model's performance on unseen future data is ok as long as none of the cross-validation data was used to make machine learning decisions.

Question 3

0 / 0 pts

There are two methods you can use for validation: Validation set, and cross validation. Define each of these two approaches (to include an explanation of differences in how they are implemented). What are the benefits and drawbacks of using the validation set approach vs. the cross validation approach.

Your Answer:

Validation set approach involved splitting the dataset into two groups, a training set and a validation set. All of the model fitting is then performed only on the training set and the performance is measured on the never before seen training set.

Cross-validation approach splits the dataset into a specific number of groups (k for k-fold or n groups for LOOCV). This approach then leaves out one group and performs model fitting on the remaining groups and then performance is measured on the group that is left out. This approach is repeated until all groups have been left out at least once. The total estimate is then averaged between the number of groups to determine the actual reported estimated error.

One benefit of the validation set approach is that it is easy to implement, but a downside is that it is highly dependent on the selection of the test and training set

A benefit of cross-validation is that it utilizes all of the data and therefore doesn't suffer the selection issues of the validation set approach. However, depending on the groupings, there may be higher bias or variance.

The validation set approach partitions the data into two datasets. The first dataset is used to fit the model. The second set (known as the validation set) is used to determine how well the fitting process is working, so that further machine learning decisions can be made. The validation set approach is computationally simple, but is subject to chance in the choice of the partition. It is possible that one set or the other is not representative of the original phenomenon. As a result, the validation set approach is highly variable on its estimation of parameters.

In cross validation, partitions of the dataset are repeatedly created. In each repetition, a new set is chosen for training and validation. The cross validation approach mitigates the influence of chance on the validation approach. By repeated partitioning and gathering of all the results, the influence of chance on the partition are reduced. This allows for a better estimation of model performance. Cross validation is more computationally expensive than the validation-set approach, but the increase in complexity depends on the number of repetitions. Leave One Out Cross validation minimizes variance, but maximizes computation, but k-fold cross validation with a reasonable value for k (5 to 10) yields a balance between computation and reduction of variance

Question 4

1 / 1 pts

Please answer the following question in text form. Be specific - wherever possible, include page numbers, filenames, concept names to help your instructor understand what you are referring to:

What was the most confusing aspect of the material you reviewed?

Your Answer:

How bootstrapping works was the most confusing aspect.

Quiz Score: **1** out of 1