STA 3180 Statistical Modelling: Resampling

Resampling

Resampling is a statistical technique used to estimate the accuracy of a given sample statistic by using the same data multiple times. It is a way of evaluating the performance of a model or estimator on a dataset without having to use a separate validation set. It can also be used to compare different models or estimators on the same dataset.

Definitions

* **Bootstrapping**: Bootstrapping is a resampling technique in which a sample is drawn with replacement from the original dataset. This means that some observations may be repeated in the bootstrap sample.

* **Cross-validation**: Cross-validation is a resampling technique in which the dataset is split into two or more parts, and each part is used as a test set for the model or estimator being evaluated.

* **Jackknife**: The jackknife is a resampling technique in which the dataset is divided into n parts, and each part is used as a test set for the model or estimator being evaluated.

Coding Examples

Bootstrapping

```
Start of Code
import numpy as np
# Generate a random sample of size 10
sample = np.random.randint(0, 10, 10)
# Create a bootstrap sample of size 10
bootstrap_sample = np.random.choice(sample, 10, replace=True)
End of Code
```

Cross-validation

```
Start of Code
from sklearn.model_selection import KFold
# Generate a random sample of size 10
sample = np.random.randint(0, 10, 10)
# Create a KFold object with 5 folds
kf = KFold(n_splits=5)
# Iterate through the folds and use each fold as a test set
for train_index, test_index in kf.split(sample):
```

```
        X_train, X_test = sample[train_index], sample[test_index]
End of Code
```

Practice Questions

1. What is resampling?

A. Resampling is a statistical technique used to estimate the accuracy of a given sample statistic by using the same data multiple times.

2. What is the difference between bootstrapping and cross-validation?

A. Bootstrapping is a resampling technique in which a sample is drawn with replacement from the original dataset, while cross-validation is a resampling technique in which the dataset is split into two or more parts, and each part is used as a test set for the model or estimator being evaluated.