MANIPAL
ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

Linear Regression Models

Segment 4 – Model Diagnostics

Topic 2 – Cross Validation

Sudarsan N.S. Acharya (sudarsan.acharya@manipal.edu)

Topics

1. Basic Idea of Cross Validation
2. K-fold Cross Validation
3. Leave-one-out Cross Validation
4. Train-Test-Validation Split vs. Cross Validation

Basic Idea of Cross Validation

Basic Idea of Cross Validation

- Cross validation is a technique for estimating prediction error of a model when the amount of available data is *small*.

Basic Idea of Cross Validation

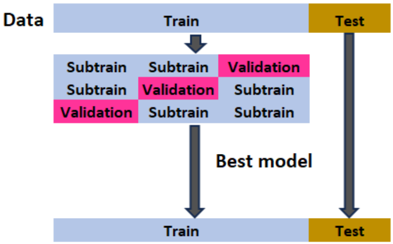
- Cross validation is a technique for estimating prediction error of a model when the amount of available data is *small*.
- Basic idea is to split the data first into training and test parts,

Basic Idea of Cross Validation

- Cross validation is a technique for estimating prediction error of a model when the amount of available data is *small*.
- Basic idea is to split the data first into training and test parts, followed by subsetting of the training part into subtrain and validation parts and cycling through those subsets in a round-robin fashion for model-building:

Basic Idea of Cross Validation

- Cross validation is a technique for estimating prediction error of a model when the amount of available data is *small*.
- Basic idea is to split the data first into training and test parts, followed by subsetting of the training part into subtrain and validation parts and cycling through those subsets in a round-robin fashion for model-building:



Data

Train | Test

Subtrain	Subtrain	Validation
Subtrain	Validation	Subtrain
Validation	Subtrain	Subtrain

Best model

Train | Test

Basic Idea of Cross Validation–Continued

Basic Idea of Cross Validation–Continued

- Each time, a model is built using the subtrain part, and its prediction error is estimated using the validation part.

Basic Idea of Cross Validation–Continued

- Each time, a model is built using the subtrain part, and its prediction error is estimated using the validation part.
- The best model is identified from the average prediction error on the validation subsets.

Basic Idea of Cross Validation–Continued

- Each time, a model is built using the subtrain part, and its prediction error is estimated using the validation part.
- The best model is identified from the average prediction error on the validation subsets.
- Finally, the best model is trained on the entire training data followed by an application of it on the test data to estimate its generalization error on unseen data.

Basic Idea of Cross Validation–Continued

- Each time, a model is built using the subtrain part, and its prediction error is estimated using the validation part.
- The best model is identified from the average prediction error on the validation subsets.
- Finally, the best model is trained on the entire training data followed by an application of it on the test data to estimate its generalization error on unseen data.
- This is an example of a 3-fold cross validation as the training data is split into 3 subsets.

K-fold Cross Validation

K-fold Cross Validation

- The number of subsets K into which the training data is split results in different types of cross validation approaches.

K-fold Cross Validation

- The number of subsets K into which the training data is split results in different types of cross validation approaches.
- Each subset of the training data is referred to as a *fold*.

K-fold Cross Validation

- The number of subsets K into which the training data is split results in different types of cross validation approaches.
- Each subset of the training data is referred to as a *fold*.
- When K is much smaller than the number of training samples, the models are built using subtrain-folds with a small number of samples.

K-fold Cross Validation

- The number of subsets K into which the training data is split results in different types of cross validation approaches.
- Each subset of the training data is referred to as a *fold*.
- When K is much smaller than the number of training samples, the models are built using subtrain-folds with a small number of samples.
- This means, the K-fold cross validation procedure will result in an overestimation of the true prediction error of the model obtained using the entire training data.

K-fold Cross Validation

- The number of subsets K into which the training data is split results in different types of cross validation approaches.
- Each subset of the training data is referred to as a *fold*.
- When K is much smaller than the number of training samples, the models are built using subtrain-folds with a small number of samples.
- This means, the K-fold cross validation procedure will result in an overestimation of the true prediction error of the model obtained using the entire training data.
- On the other hand, the K-fold cross validation procedure will result in an estimate of the true prediction error (of the model obtained using the entire training data) which will not be too sensitive to the subtrain data used to train the models.

Leave-one-out Cross Validation

Leave-one-out Cross Validation

- When $K = n$, the resulting cross validation approach is called the leave-one-out cross validation procedure.

Leave-one-out Cross Validation

- When $K = n$, the resulting cross validation approach is called the leave-one-out cross validation procedure.
- The leave-one-out cross validation procedure makes use of almost the entire training data for model-building which results in an almost unbiased estimate of the true prediction error.

Leave-one-out Cross Validation

- When $K = n$, the resulting cross validation approach is called the leave-one-out cross validation procedure.
- The leave-one-out cross validation procedure makes use of almost the entire training data for model-building which results in an almost unbiased estimate of the true prediction error.
- On the other hand, the leave-one-out cross validation procedure will result in an estimate of the true prediction error that is very sensitive to the subtrain data used to train the models.

Leave-one-out Cross Validation

- When $K = n$, the resulting cross validation approach is called the leave-one-out cross validation procedure.
- The leave-one-out cross validation procedure makes use of almost the entire training data for model-building which results in an almost unbiased estimate of the true prediction error.
- On the other hand, the leave-one-out cross validation procedure will result in an estimate of the true prediction error that is very sensitive to the subtrain data used to train the models.
- K is typically chosen to be 5 or 10 to strike a balance between computational efficiency and a reliable estimate of the true prediction error.

Train-Test-Validation Split vs. Cross Validation

Train-Test-Validation Split vs. Cross Validation

- Suppose we are given a dataset;

Train-Test-Validation Split vs. Cross Validation

- Suppose we are given a dataset; how would we choose between a train-validation-test procedure and a cross validation procedure for estimating the generalization error of the model on unseen data?

Train-Test-Validation Split vs. Cross Validation

- Suppose we are given a dataset; how would we choose between a train-validation-test procedure and a cross validation procedure for estimating the generalization error of the model on unseen data?
- In production systems, such as recommendation systems, the model will be trained on all available data.

Train-Test-Validation Split vs. Cross Validation

- Suppose we are given a dataset; how would we choose between a train-validation-test procedure and a cross validation procedure for estimating the generalization error of the model on unseen data?
- In production systems, such as recommendation systems, the model will be trained on all available data.
- In this case, if the computational cost is not too high, it would be better to use a cross validation procedure as it would yield a reliable (low-variance) estimate of the generalization error.

Train-Test-Validation Split vs. Cross Validation

- Suppose we are given a dataset; how would we choose between a train-validation-test procedure and a cross validation procedure for estimating the generalization error of the model on unseen data?
- In production systems, such as recommendation systems, the model will be trained on all available data.
- In this case, if the computational cost is not too high, it would be better to use a cross validation procedure as it would yield a reliable (low-variance) estimate of the generalization error.
- On the other hand, if there is a separate test set with a substantial number of samples that will be used for model evaluation, it would be better to use a train-validation split procedure which will lead to an unbiased estimate of the generalization error.

Summary

Summary

- Describe cross validation and its necessity.

Summary

- Describe cross validation and its necessity.
- Compare and contrast different cross validation approaches.