

TESTING UNSUPERVISED LEARNING

This slide provides an overview of the techniques used to test for overfitting in customer segmentation models.

CROSS-VALIDATION

Adapting Cross-Validation for Unsupervised Learning

Training on the Training Set

Applying to the Validation Set

Evaluating Cluster Similarity

While cross-validation is typically used in supervised learning, it can be adapted for unsupervised scenarios as well. The idea is to randomly split the data into training and 'validation' sets, train the clustering algorithm on the training set, and then apply it to the validation set to see if similar cluster structures emerge.

The first step is to train the clustering algorithm on the training set. This allows the model to learn the underlying patterns and structure in the data.

After training on the training set, the next step is to apply the clustering algorithm to the validation set. This allows you to assess how well the model generalizes to new, unseen data.

observed in the validation set are similar to those in the training set, it suggests that the model has learned the true underlying patterns in the data and is not overfitting.

Significant differences in the cluster structures may indicate overfitting.

SCORE ANALYSIS

• What are Silhouette Scores?

Silhouette scores measure how similar an object is to its own cluster compared to other clusters. A high silhouette score indicates the object is well-matched to its own cluster and poorly matched to neighboring clusters.

Identifying Overfitting with Silhouette Scores

If the training clusters have very high silhouette scores but these scores drastically drop when new data is clustered, it might indicate that the model has overfit to the training data and is not generalizing well to new samples.