Data Mining Machine Learning

Md. Abdullah-Al-Kafi
Lecturer
Daffodil International University

1 Cross-Validation and Hold-Out

Cross-Validation is a technique used to evaluate a model's performance by partitioning data into subsets. The model is trained on some subsets (training sets) and validated on others (validation sets). This process is repeated multiple times, helping to reduce bias and improve accuracy.

Hold-Out Validation is a simpler approach where the dataset is split into two parts: a training set and a test (or validation) set. The model is trained on the training set and tested on the hold-out test set.

Example Scenario: Suppose you have a dataset of 500 reviews for a sentiment analysis model.

- **Hold-Out**: Split the dataset into 80% (400 reviews) for training and 20% (100 reviews) for testing. Train the model on the 400 reviews and evaluate it on the 100 remaining reviews.
- **k-Fold Cross-Validation** (e.g., 5-fold): Split the dataset into 5 equal parts, each containing 100 reviews. Train the model 5 times, using 4 parts (400 reviews) for training and 1 part (100 reviews) for testing. Every review is used for testing exactly once, resulting in 5 evaluation scores that can be averaged.

2 Rule Support

Support measures the frequency with which an item or itemset appears in the dataset. In association rule mining, it represents the proportion of records that contain a particular item or set of items.

Example Calculation: Consider a grocery dataset where:

- 100 transactions contain "Milk and Bread."
- The total number of transactions is 1000.

The support for the rule "If a customer buys Milk, they also buy Bread" is given by:

$$\text{Support} = \frac{\text{Transactions with Milk and Bread}}{\text{Total Transactions}} = \frac{100}{1000} = 0.1 \text{ (or 10\%)}$$

3 Rule Confidence

Confidence measures the likelihood that a consequent item is bought if a preceding item (antecedent) is already bought. It is the conditional probability of the rule.

Example Calculation: Continuing with the example above, if 150 transactions contain "Milk" in total (regardless of whether Bread is also present), the confidence of the rule "If a customer buys Milk, they also buy Bread" is:

$$\text{Confidence} = \frac{\text{Transactions with Milk and Bread}}{\text{Transactions with Milk}} = \frac{100}{150} = 0.67 \text{ (or 67\%)}$$

So, there is a 67% chance that if a customer buys Milk, they will also buy Bread.

4 Rule Coverage

Coverage of a rule is the number or proportion of instances in the dataset that match the antecedent of the rule.

Example Calculation: If 150 transactions out of 1000 contain "Milk," the coverage of the rule "If a customer buys Milk, they also buy Bread" is:

$$\text{Coverage} = \frac{\text{Transactions with Milk}}{\text{Total Transactions}} = \frac{150}{1000} = 0.15 \text{ (or 15\%)}$$

This means that the rule applies to 15% of all transactions in the dataset.

Example

Md. Abdullah-Al-Kafi
Lecturer
Daffodil International University

5 Scenario: Grocery Store Analysis

Imagine you are a data analyst at a grocery store, working with a dataset of customer transactions. Each transaction records the items that customers bought together, such as "Milk," "Bread," and "Butter." The goal is to identify common purchasing patterns that can help the store optimize product placements, design promotional offers, and increase cross-selling opportunities.

6 Data Splitting with Cross-Validation and Hold-Out

To predict which items a customer might buy together, a machine learning model can be trained. To evaluate its performance, we use **Cross-Validation** and **Hold-Out Validation** techniques.

6.1 Hold-Out Validation

In Hold-Out Validation, the data is split into two sets: 80% training and 20% testing. The model is trained on the training set (80% of the transactions) and then evaluated on the test set (20% of the transactions). This approach is straightforward but may not fully represent the entire dataset.

6.2 5-Fold Cross-Validation

In k-Fold Cross-Validation (e.g., 5-fold), the dataset is divided into 5 equal parts. For each of the 5 runs, 4 parts are used for training, and 1 part is used for testing. This process is repeated 5 times, so each part serves as the test set once. By averaging the results, we get a more reliable estimate of model performance across different subsets.

7 Association Rules: Support, Confidence, and Coverage

Now, let's explore association rules to identify which items are frequently bought together. Suppose we have the following data:

- Total transactions: 1000
- Transactions containing both "Milk and Bread": 100
- Transactions containing "Milk" (regardless of Bread): 150

Using these values, we can calculate **Support**, **Confidence**, and **Coverage** for the rule:

"If a customer buys Milk, they are likely to buy Bread."

7.1 Support

Support measures how frequently "Milk and Bread" are bought together across all transactions. High support indicates that the item combination is commonly purchased.

$$\text{Support} = \frac{\text{Transactions with Milk and Bread}}{\text{Total Transactions}} = \frac{100}{1000} = 0.1 \text{ (or 10\%)} \tag{1}$$

Interpretation: 10% of all transactions contain both "Milk" and "Bread," indicating a popular item combination.

7.2 Confidence

Confidence measures the likelihood of Bread being bought when Milk is already purchased. It represents the conditional probability of a customer buying Bread given that they bought Milk.

$$\text{Confidence} = \frac{\text{Transactions with Milk and Bread}}{\text{Transactions with Milk}} = \frac{100}{150} = 0.67 \text{ (or 67\%)} \tag{2}$$

Interpretation: If a customer buys Milk, there is a 67% chance they will also buy Bread. High confidence means the rule is fairly reliable.

7.3 Coverage

Coverage measures the proportion of total transactions that contain the antecedent of the rule (in this case, "Milk"), regardless of whether Bread was bought as well. Coverage shows how broadly the rule applies.

$$\text{Coverage} = \frac{\text{Transactions with Milk}}{\text{Total Transactions}} = \frac{150}{1000} = 0.15 \text{ (or 15\%)} \tag{3}$$

Interpretation: The rule "If a customer buys Milk, they also buy Bread" applies to 15% of all transactions. This indicates the general applicability of the rule in the dataset.

8 Why These Metrics Matter

These metrics are essential for association rule mining in retail scenarios like this grocery store example:

- **Support** helps determine if an item combination (like Milk and Bread) is common enough to be valuable for promotional offers.
- **Confidence** indicates the strength of the association between items. A high confidence suggests that placing Milk near Bread in the store might increase Bread sales when Milk is bought.
- **Coverage** shows the rule's applicability across all transactions. High coverage implies that the rule is relevant to a significant portion of customers, making it impactful for marketing strategies.

For instance, knowing that Milk and Bread are frequently bought together, the store could offer a discount on Bread when Milk is purchased, potentially boosting overall sales.