Introduction to Machine Learning and Dataset Handling

Feature Sets, Dataset Division, Cross Validation with Examples

What is Machine Learning?

- • Machine Learning (ML) is a branch of artificial intelligence (AI) that enables machines to learn from data without being explicitly programmed.
- • ML algorithms identify patterns, make decisions, and predictions based on historical data.
- • Types of ML:
- - Supervised Learning
- - Unsupervised Learning
- - Reinforcement Learning

Usage of Datasets in Machine Learning

- • A dataset is a collection of data used to train and evaluate machine learning models.
- • Datasets consist of multiple features (inputs) and target labels (outputs).
- • Important dataset types include:
-   - Structured data (tables)
-   - Unstructured data (text, images, etc.)
- • Example: Predicting house prices based on features like size, location, and number of rooms.

Handling Datasets for Machine Learning

- • Dataset Preprocessing: Cleaning and transforming raw data into a usable form.
- - Handle missing values, outliers.
- - Normalize or standardize features.
- - Feature engineering: Create new features from existing ones.
- • Example: Convert categorical data (e.g., 'Red', 'Blue') into numerical form using one-hot encoding.

Feature Sets in Machine Learning

- • A feature set consists of all the attributes or columns in a dataset that influence the model's prediction.
- • Features represent the independent variables, while the target label represents the dependent variable.
- • Example: In a dataset predicting house prices, features include size, location, and number of rooms, while the target label is the price.

Dataset Division: Train, Test, Validation Sets

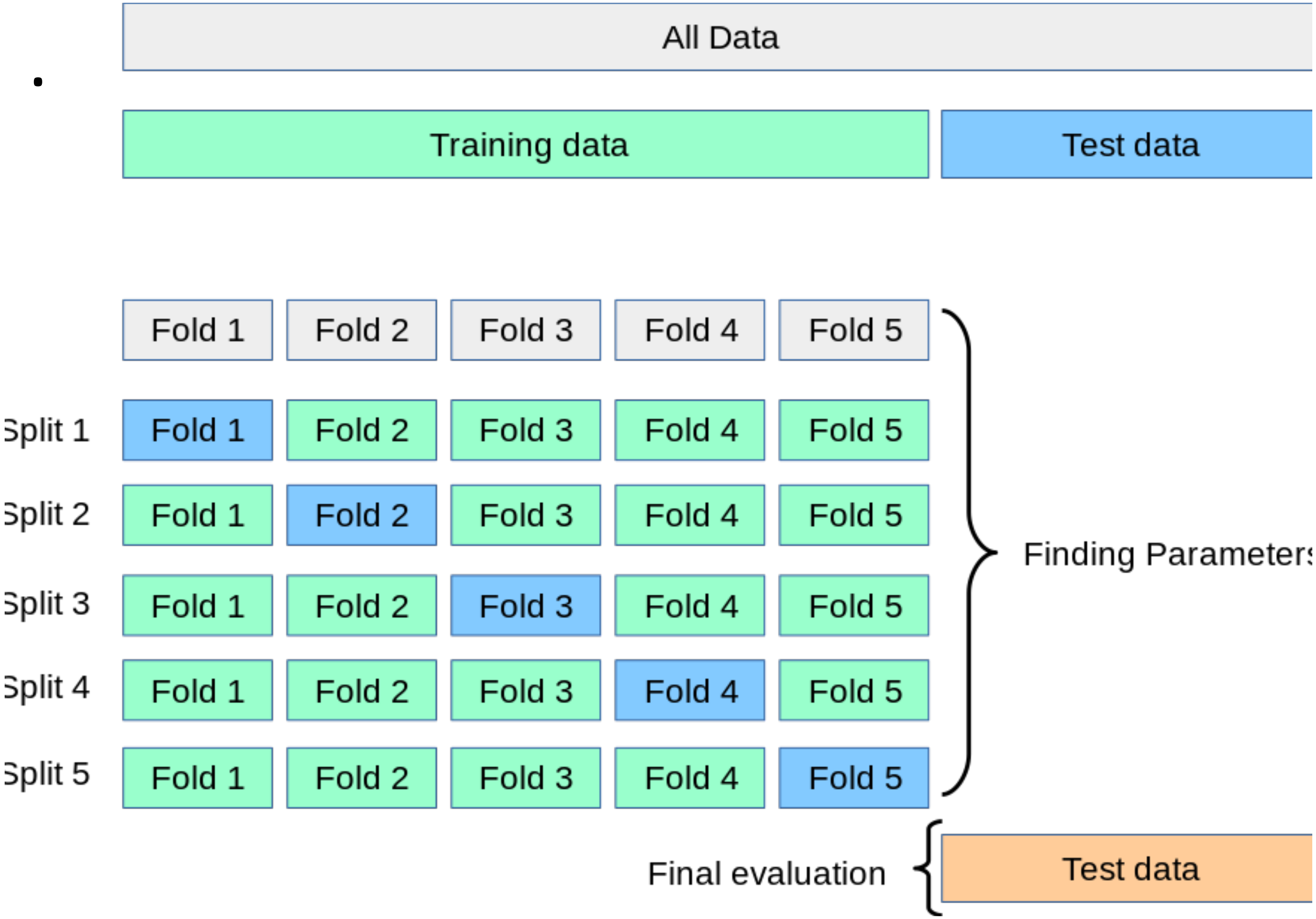
- • Train Set: Used to train the model and fit the parameters.
- • Test Set: Used to evaluate model performance on unseen data.
- • Validation Set: Used to fine-tune model parameters and prevent overfitting.
- • Common Split: 70% training, 20% testing, and 10% validation.

Cross Validation

- • Cross-validation is a technique to assess model performance on multiple subsets of data.
- • Example: K-Fold Cross Validation:
- - Split the data into K subsets.
- - Train on K-1 subsets and test on the remaining one.
- - Rotate the test set and average the results for a more accurate estimate of model performance.

5-cross Validation

- .



All Data

Training data Test data

Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 1 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 2 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 3 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 4 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 5 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Finding Parameters

Final evaluation Test data

.

Example For 5-cross validation

We will do 5 iteration:

1. First Iteration:
   - **Training Set: Folds 2, 3, 4, 5 (Samples 3, 4, 5, 6, 7, 8, 9, 10)**
   - **Test Set: Fold 1 (Samples 1, 2)**
   - **Train the model on Folds 2, 3, 4, and 5.**
   - **Test the model on Fold 1.**

2. Second Iteration:

- **Training Set: Folds 1, 3, 4, 5 (Samples 1, 2, 5, 6, 7, 8, 9, 10)**
- **Test Set: Fold 2 (Samples 3, 4)**
- **Train the model on Folds 1, 3, 4, and 5.**
- **Test the model on Fold 2.**

3. Third Iteration:

- **Training Set: Folds 1, 2, 4, 5 (Samples 1, 2, 3, 4, 7, 8, 9, 10)**
- **Test Set: Fold 3 (Samples 5, 6)**
- **Train the model on Folds 1, 2, 4, and 5.**
- **Test the model on Fold 3.**

4. Fourth Iteration:

- **Training Set: Folds 1, 2, 3, 5 (Samples 1, 2, 3, 4, 5, 6, 9, 10)**
- **Test Set: Fold 4 (Samples 7, 8)**
- **Train the model on Folds 1, 2, 3, and 5.**
- **Test the model on Fold 4.**

5. Fifth Iteration:

- **Training Set: Folds 1, 2, 3, 4 (Samples 1, 2, 3, 4, 5, 6, 7, 8)**
- **Test Set: Fold 5 (Samples 9, 10)**
- **Train the model on Folds 1, 2, 3, and 4.**
- **Test the model on Fold 5.**

Final Calculation After All 5 Iterations

After performing 5 iterations, you will have 5 evaluation scores (e.g., accuracy, mean squared error, etc.). The final performance of the model is obtained by averaging these scores across all folds.

Example:

- **Accuracy for each fold: [0.90, 0.85, 0.88, 0.89, 0.87]**
- **Mean Accuracy = (0.90 + 0.85 + 0.88 + 0.89 + 0.87) / 5 = 0.878**

Example: Predicting House Prices

- • Dataset: Predict house prices based on features like size, location, and number of rooms.
- • Steps:
- 1. Collect and preprocess data (handle missing values, normalize features).
- 2. Divide data into training, testing, and validation sets.
- 3. Train the model (e.g., Linear Regression) on the training set.
- 4. Evaluate model performance on the test set.
- 5. Use cross-validation for more reliable results.

Important Points to Consider

- • Ensure data quality: Clean and preprocess the dataset properly.
- • Avoid data leakage: Test data should never be used to train the model.
- • Handle imbalanced datasets: Use techniques like SMOTE or stratified sampling for balanced datasets.
- • Use cross-validation for robust model evaluation.