Model Validation (Selection) in ML (DL)

by Dr. Rishikesh Yadav

October 22, 2025

Assistant Professor, School of Mathematical and Statistical Sciences, IIT Mandi, India

Table of Contents

Model Validation (Choosing the Best Model)

Model Validation: Why Do We Need It?

How do we know if our model is actually any good?

- Is a **linear regression model** good enough for a given data, OR
- Do we need a **polynomial model**? If so, what **degree** is enough?
- Are any models other than **polynomial** needed to be looked at?
- A model that performs well on **training data** may fail on **new data**.
- We need to assess the model's **generalization performance**.
- Model validation helps avoid **overfitting** and supports better **model selection**.

Agenda

Training, Validation, and Test Splits

- **Training Set** ($\approx$**60–70%** of data): Used to **fit** the model - the model **"learns"** from this data.
- **Validation Set** ($\approx$**15–20%**):
  - Used to **tune hyperparameters** and monitor performance.
  - Allows techniques like **early stopping** or **model selection**.
- **Test Set** ($\approx$**15–20%**):
  - Held out until **final evaluation**.
  - Measures true **generalization** to **unseen data**.

Important Note: In most cases, we only divide the whole data into training and testing datasets.

Most Common Data Splitting Strategies

- **Random Split**:
  - Common and simple.
  - Risky if the dataset is small or imbalanced.
- **Stratified Split**:
  - Preserves class proportions in classification.
- **Time-Based Split**:
  - Used for time-series data where order matters.

Cross-Validation (CV)

- Splits data into multiple folds and rotates the validation set.
- Common types:
  - **k-Fold CV**: Divides data into k equal parts
  - **Leave-One-Out CV**: Each data point is its own fold
- Helps in more reliable model evaluation when data is limited.

Agenda

Model Evaluation Metrics: Regression Metrics

- **Mean Squared Error (MSE)**:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, and n is the total number of observations.

- **Root Mean Squared Error (RMSE)**:

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Gives error in the same unit as the response variable.

- **Coefficient of Determination R^2**:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the observed values. It represents the proportion of variance in y explained by the model.

Model Evaluation Metrics: Classification Metrics

- **Accuracy**:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

  where:

  - TP = True Positives
  - TN = True Negatives
  - FP = False Positives
  - FN = False Negatives

- **Precision and Recall**:

$$\textbf{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \textbf{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

  Precision measures correctness of positive predictions; recall measures ability to find all actual positives.

- **F1 Score** (harmonic mean of precision and recall):

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

  Useful when classes are imbalanced or when both precision and recall are important.