# Tutorial: Cross-Validation

*Cheat Sheet and Practice Problems*

ES335 - Machine Learning
IIT Gandhinagar

July 23, 2025

## 1 Summary from Slides

### 1.1 Introduction and Motivation

**Limitations of Single Train/Test Split**:

- Does not utilize the full dataset for training
- Cannot optimize hyperparameters systematically
- Results depend on the particular split chosen
- May not get reliable performance estimates

**Goal**: Use the full dataset for both training and testing to get robust performance estimates

### 1.2 K-Fold Cross-Validation

**Basic Procedure**:

1. Divide dataset into $k$ equal parts (folds)
2. For each fold $i = 1, 2, ..., k$:
   - Use fold $i$ as test set
   - Use remaining $k - 1$ folds as training set
   - Train model and evaluate performance
3. Average the $k$ performance scores

**Key Properties**:

- Each data point is used for testing exactly once
- Each data point is used for training $(k - 1)/k$ of the time
- Provides more robust performance estimates than single split

### 1.3 Hyperparameter Optimization

**Train/Validation/Test Split**:

- Training set: Train model with different hyperparameters
- Validation set: Select best hyperparameters
- Test set: Final evaluation (remains untouched until end)

**Purpose**: Prevents overfitting to the test set during hyperparameter tuning

## 1.4 Nested Cross-Validation

**Structure**: Cross-validation within cross-validation

- Outer loop: Performance estimation
- Inner loop: Hyperparameter optimization
- Provides unbiased performance estimates with hyperparameter tuning

## 1.5 Cross-Validation Variants

**Leave-One-Out CV (LOOCV)**:

- Special case: $k = n$ (number of samples)
- Maximum use of training data
- Computationally expensive for large datasets
- High variance in performance estimates

**Stratified K-Fold**:

- Maintains class distribution in each fold
- Important for imbalanced datasets
- Ensures each fold is representative of the whole dataset

## 1.6 Time Series Cross-Validation

**Forward Chaining**:

- Respects temporal order of data
- Training set always precedes test set in time
- Cannot use future data to predict past (no data leakage)

**Expanding Window**: Training set grows with each fold **Rolling Window**: Fixed-size training window

## 1.7 Best Practices and Pitfalls

**Common Pitfalls**:

- Data leakage between folds
- Inappropriate splitting for time series data
- Not maintaining class balance in stratified scenarios
- Performing feature selection on entire dataset before CV

**Best Practices**:

- Choose appropriate $k$ (typically 5 or 10)
- Use stratified CV for classification
- Perform all preprocessing within CV folds
- Use time-aware splits for temporal data

## 2 Practice Problems

**Problem : Basic K-Fold Calculation**

You have a dataset with 120 samples and want to use 6-fold cross-validation.
a) How many samples will be in each fold? b) How many samples will be used for training in each iteration? c) How many samples will be used for testing in each iteration?

**Problem : LOOCV Analysis**

For a dataset with 50 samples:
a) How many models will be trained in Leave-One-Out Cross-Validation? b) What is the size of each training set? c) What are the advantages and disadvantages compared to 5-fold CV?

**Problem : Stratified CV Design**

You have a binary classification dataset with 1000 samples: 800 class A and 200 class B. Design a 5-fold stratified cross-validation setup.
a) How many samples of each class should be in each fold? b) Why is stratification important for this dataset? c) What could go wrong with regular k-fold CV?

**Problem : Hyperparameter Tuning Setup**

You want to tune the hyperparameter $C$ for SVM using values $[0.1, 1, 10, 100]$ with 5-fold cross-validation on 200 training samples.
a) How many models will be trained in total? b) Design the complete evaluation procedure c) How do you select the final hyperparameter value?

**Problem : Nested CV Structure**

Explain nested cross-validation with outer 5-fold and inner 3-fold CV:
a) Draw the structure showing outer and inner loops b) If you have 150 samples, how many models are trained total? c) What is the purpose of each level of cross-validation?

**Problem : Time Series CV Design**

Design a time series cross-validation for stock price prediction with 365 daily observations:
a) Why can't you use standard k-fold CV? b) Design a forward-chaining approach with 5 splits c) What is the training and test size for each split?

**Problem : Data Leakage Identification**

Identify the data leakage in these scenarios:
a) Normalizing the entire dataset before splitting into folds b) Using future stock prices to predict past prices c) Feature selection on the entire dataset before CV d) Using validation set multiple times during hyperparameter tuning

**Problem : CV Variance Analysis**

You get these 5-fold CV scores: [0.85, 0.92, 0.78, 0.90, 0.88]
a) Calculate the mean and standard deviation b) What does high variance in CV scores indicate? c) How would you interpret these results? d) Compare with LOOCV variance characteristics

**Problem : Computational Complexity**

Compare the computational cost of different CV approaches for a dataset with $n$ samples:
a) Single train/test split (70/30) b) 5-fold cross-validation c) 10-fold cross-validation d) Leave-One-Out cross-validation
Express in terms of number of model training operations.

**Problem : Feature Selection with CV**

You want to perform feature selection with cross-validation. Design the correct procedure:
a) Where should feature selection be performed in the CV loop? b) What happens if you do feature selection before CV? c) How does this affect computational cost? d) Design nested CV with feature selection and hyperparameter tuning

**Problem : Model Comparison**

Compare three models (Linear Regression, SVM, Random Forest) using 10-fold CV:
Results: - Linear Regression: Mean=0.82, Std=0.05 - SVM: Mean=0.85, Std=0.12 - Random Forest: Mean=0.84, Std=0.03
a) Which model would you choose and why? b) How would you test for statistical significance? c) What if you only had these single scores: LR=0.82, SVM=0.89, RF=0.81?

**Problem : Choosing K Value**

For the following scenarios, recommend an appropriate value of $k$ and justify:
a) Small dataset with 50 samples b) Large dataset with 1 million samples c) Imbalanced dataset with 90% majority class d) High-dimensional dataset with many features e) Time-constrained scenario needing quick results

**Problem : Cross-Validation Bias**

Analyze potential biases in cross-validation:
a) What is the bias of k-fold CV performance estimate? b) How does the choice of $k$ affect bias and variance? c) Why might CV overestimate performance for certain algorithms? d) How does sample size affect CV reliability?

**Problem : Practical Implementation**

Design a complete cross-validation pipeline for image classification:
a) How would you handle data augmentation? b) Where should normalization be applied? c) How to handle class imbalance? d) What metrics would you use for evaluation? e) How to ensure reproducible results across folds?

Problem : Advanced Challenge

You're working with grouped data (e.g., multiple images per person):
a) Why is standard k-fold CV problematic? b) Design a group-aware cross-validation strategy c) How does this affect your performance estimates? d) What are the implications for real-world deployment? e) How would you handle time series data with groups?