安徽理工大学
ANHUI UNIVERSITY OF SCIENCE & TECHNOLOGY

- 人工智能专业 学科基础教育必修模块

2025

Python与机器学习

Python and Machine Learning

Chapter 7: Model Evaluation and Selection

- Lecturer ： 孟亦凡
- E-mail：myf@aust.edu.cn

Chapter 7: Model Evaluation and Selection

Target

- 理解数据集划分的目的与方法 |Understand the Purpose and Methods of Dataset Division
- 理解欠拟合与过拟合的概念 | Understand the Concepts of Underfitting and Overfitting
- 理解归纳偏好的理论基础 | Understand the Theoretical Basis of Inductive Preference
- 理解性能指标的分类与意义 | Understand the Classification and Significance of Performance Metrics

Chapter 7: Model Evaluation and Selection

Target

- 理解性能指标的分类与意义 | Understand the Classification and Significance of Performance Metrics
- 理解偏差与方差的概念及其对模型性能的影响 | Understand the Concepts of Bias and Variance and Their Impact on Model Performance

Chapter 7: Model Evaluation and Selection

CONTENTS

01 **Dividing the dataset**
数据集划分

02 **Underfitting and Overfitting**
欠拟合与过拟合

03 **Inductive Preference**
归纳偏好

04 **Performance Metrics**
性能指标

05 **Bias and Variance**
偏差与方差

7.1 Dividing the dataset
数据集划分

划分数据集是为了能更好的训练评估模型
The dataset is partitioned to optimize model training and evaluation.

7.1 Dividing the Dataset

➢ Why Divide the Dataset? Evaluate Model Generalization Ability

- Training Set (训练集)
  - Used to train model parameters
  - The model learns patterns and rules from this data
- Validation Set (验证集)
  - Used to evaluate model performance and tune hyperparameters（用于评估模型性能并调整超参数）
  - Can be evaluated multiple times to guide model optimization（可进行多次评估，指导模型优化方向）

7.1 Dividing the Dataset

➢ Why Divide the Dataset? Evaluate Model Generalization Ability

- 测试集 (Test Set)
    - Used for final evaluation of model performance on new data（用于最终评估模型在新数据上的性能）
    - Used only once, simulating real-world performance after deployment
    - （只能使用一次，模拟模型部署后的真实表现）

Key Principles

- **Independence Principle (独立性原则)**
  - **The test set should be independently sampled from the same true distribution as the training set. Ensure training and test sets are as mutually exclusive as possible.**
  - **测试集应从与训练集相同的真实分布中独立采样。尽可能保证训练集与测试集互斥。**

7.1 Dividing the Dataset

➢ Key Principles

- Consistency Principle (一致性原则)
  - The train/test split should maintain consistency in data distribution as much as possible. Avoid data distribution shift caused by splitting.
  - 训练/测试划分应尽可能保持数据分布的一致性。避免因划分导致的数据分布偏移。

Key Principles

- Representativeness Principle (代表性原则)
  - Each subset should sufficiently represent the characteristics of the original dataset. For imbalanced data, stratified sampling should be used.
  - 各子集应充分代表原始数据集的特征。对于类别不平衡数据，应采用分层抽样。

➢ Key Principles

- Representativeness Principle (代表性原则)
  - Each subset should sufficiently represent the characteristics of the original dataset. For imbalanced data, stratified sampling should be used.
  - 各子集应充分代表原始数据集的特征。对于类别不平衡数据，应采用分层抽样。

7.1 Dividing the Dataset

➢ Method 1: Hold-Out Method (留出法)

- **Basic Concept:**
  - **Directly divide dataset D into two mutually exclusive sets. Usually divided by a certain ratio (e.g., 2:1, 4:1, 7:3 or 8:2).**
  - **直接将数据集D划分为两个互斥的集合。通常按一定比例划分（如2:1, 4:1, 7:3或8:2）。**

$$D = D_{\text{train}} \cup D_{\text{test}}, \quad D_{\text{train}} \cap D_{\text{test}} = \emptyset$$

7.1 Dividing the Dataset

➢ Method 1: Hold-Out Method (留出法)

- **Considerations:**
  - **Single division is random → Usually perform multiple random divisions, take average results.**
  - **单次划分具有随机性 → 通常多次随机划分，取平均结果。**

轮次	训练集	测试集
1	S2、S3、S4	S1
2	S1、S3、S4	S2
3	S1、S2、S4	S3
4	S1、S2、S3	S4

➢ Method 1: Hold-Out Method (留出法)

- Maintain data distribution → Stratified sampling can be used to maintain class proportions.
- 保持数据分布 → 可使用分层抽样保持类别比例。
- Common ratio: Training set : Test set ≈ 2:1 to 4:1.
- 常见比例：训练集:测试集 ≈ 2:1 到 4:1。

7.1 Dividing the Dataset

➢ Method 1: Hold-Out Method (留出法)

Dimension	Advantages	Disadvantages
Data Utilization Rate	Training set utilization rate ≈ 100%, retaining all data information	The test set contains only one sample, which is vulnerable to the influence of outliers(异常值).
Evaluating Stability	Unique partitioning, stable results, no randomness	The indicators are biased toward the majority class in imbalanced datasets.
Computer costs	No additional simplification costs	Iterating N times is prohibitively expensive for complex models and large datasets.
Applicable Scenarios	Small sample size, sparse data, and models require precise scenario evaluation.	Large-scale samples, complex models, and imbalanced category scenarios
Information leak	No risk of leakage, objective assessment	No apparent information leakage issues

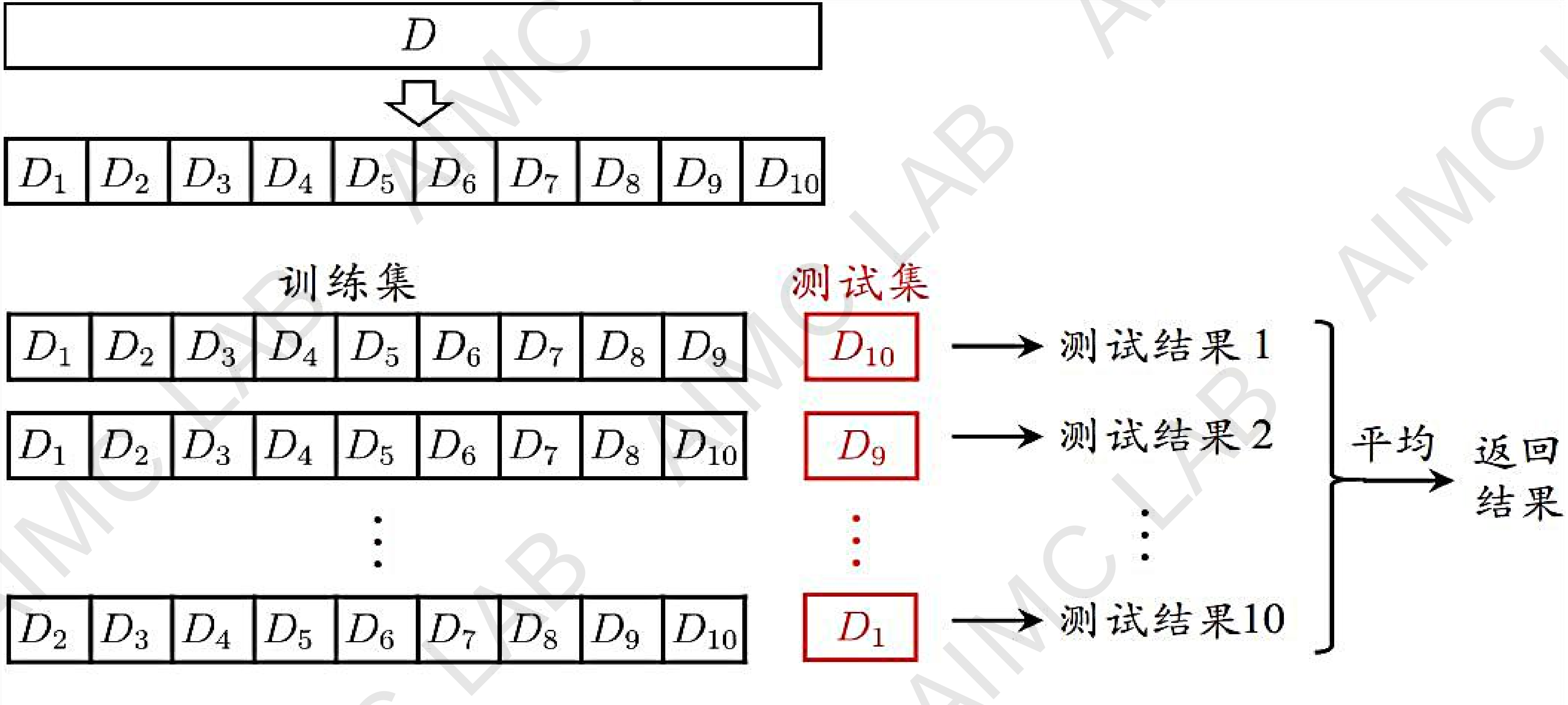
7.1 Dividing the Dataset

➢ Method 2: Cross-validation method (交叉验证法)

- Randomly divide dataset D into k similar-sized, mutually exclusive subsets.
- 将数据集D随机划分为k个大小相似且互斥的子集。
- Each time use the union of k-1 subsets as training set, remaining subset as test set.
- 每次使用k-1个子集的并集作为训练集，剩余1个子集作为测试集。
- Repeat k times, obtain k test results, take average as final evaluation.
- 重复k次，得到k个测试结果，取平均值作为最终评估。

7.1 Dividing the Dataset

➢ Method 2: Cross-validation method (交叉验证法)



D

D_1 D_2 D_3 D_4 D_5 D_6 D_7 D_8 D_9 D_{10}

训练集 测试集

D_1 D_2 D_3 D_4 D_5 D_6 D_7 D_8 D_9 | D_{10} → 测试结果 1

D_1 D_2 D_3 D_4 D_5 D_6 D_7 D_8 D_{10} | D_9 → 测试结果 2

⋮ ⋮ ⋮

D_2 D_3 D_4 D_5 D_6 D_7 D_8 D_9 D_{10} | D_1 → 测试结果10

平均 → 返回结果

10 折交叉验证示意图

7.1 Dividing the Dataset

➢ Method 2: Cross-validation method (交叉验证法)

- Similar to the leave-one-out method, there are multiple ways to divide a dataset D into k subsets.
- To minimise variations introduced by different sample divisions, k-fold cross-validation typically employs random divisions repeated p times.
- The final evaluation result is the mean of the p times k-fold cross-validation outcomes. For instance, the common ‘10-fold cross-validation repeated 10 times’.

➢ Method 2: Cross-validation method (交叉验证法)

- Special Case: Leave-One-Out, LOO (特殊形式：留一法)
  - When k equals sample size m: k = m. Each subset contains only one sample. Advantages: Not affected by random division, stable results. Disadvantages: Large computational overhead, suitable for small datasets.
  - 当k等于样本数m时：k = m。每个子集只包含一个样本。优点：不受随机划分影响，结果稳定。缺点：计算开销大，适用于小数据集。

7.1 Dividing the Dataset

Variant Type	Advantages	Disadvantages
Standard k-fold (k=5/10) (标准k折)	Balancing computational cost and evaluation accuracy, with the widest range of applicable scenarios.平衡计算成本与评估准确性，适用场景最广泛。	Normal k-folding is not well-suited for imbalanced data; results exhibit slight fluctuations.不适合不平衡数据；结果存在轻微波动。
Layered k-fold分层k折	Maintain the distribution of each fold category consistent with the original data, adapting to imbalanced data.保持每一折的类别分布与原始数据一致，适应不平衡数据	Applicable only to classification tasks; regression tasks cannot be hierarchical.仅适用于分类任务；回归任务无法使用分层。
Leaving One Method(k=N)	Training set utilization rate: 100%, with stable results and no fluctuations.训练集利用率：100%，结果稳定，无波动。	Extremely costly to compute (unusable when N is large); highly susceptible to outliers.计算成本极高（当N很大时无法使用）；对异常值非常敏感。

7.1 Dividing the Dataset

Variant Type	Advantages	Disadvantages
fold k times(重复k折)	The results demonstrate superior stability compared to conventional k-folds, enabling more reliable evaluation.结果比常规k折更稳定，能够进行更可靠的评估。	The computational cost is the number of repetitions multiplied by the standard k-fold value (e.g., 5 repetitions = 5*k models).计算成本是重复次数乘以标准k折的成本（例如，5次重复 = 5×k 个模型需要训练）
Group k-fold (组k折)	Avoid cross-set association of samples; adapt to data with grouping characteristics (e.g., users, experiment batches).避免样本在不同集合间产生关联；适应具有分组特征的数据（例如，用户、实验批次）。	Clear grouping labels are required; unequal grouping may compromise assessment effectiveness.需要明确的分组标签；分组大小不均可能影响评估效果。
Time Series k-Fold(时间序列k折)	Conforms to temporal causal logic and adapts to time-series forecasting tasks符合时间因果逻辑，适应时间序列预测任务。	The training set contains only historical data, and the training set for some rounds is relatively small.训练集只包含历史数据，某些轮次的训练集相对较小。

Method 3: Bootstrap Method (自助法)

- **Basic Idea:**
- **Randomly sample m times with replacement from dataset D containing m samples. Form training set D' (containing about 63.2% of original samples). Samples not appearing in D' (about 36.8%) serve as test set.**
- **从包含m个样本的数据集D中有放回地随机采样m次。形成训练集D'（包含约63.2%的原始样本）。未出现在D'中的样本（约36.8%）作为测试集。**

➢ Method 3: Bootstrap Method (自助法)

- **Mathematical Principle:**
  - **The probability that a sample is never selected in m samplings:**
  - **一个样本在m次采样中始终不被选中的概率：**

$$\lim_{m\to\infty}\left(1-\frac{1}{m}\right)^m \approx \frac{1}{e} \approx 0.368$$

  - **Therefore, training set D' contains about 63.2% of samples from the original dataset.**
  - **因此，训练集D'大约包含原始数据集中63.2%的样本。**

➢ Comparison and Selection of Three Methods

Method	Advantages (优点)	Disadvantages (缺点)	Suitable Scenarios (适用场景)
Hold-Out (留出法)	- 简单直接 - 计算效率高	- 结果受随机划分影响大 - 数据利用率较低	- 大数据集 - 初步快速评估
k-Fold CV (k折交叉验证)	- 数据利用率高 - 评估结果稳定	- 计算开销较大 - k值选择影响结果	- 中等规模数据集 - 需要稳定评估结果
Bootstrap (自助法)	- 适合小数据集 - 可生成多个训练集	- 改变数据分布，引入估计偏差 - 测试集分布有偏	-数据集很小 - 集成学习

7.1 Dividing the Dataset

➢ Selection Guide:

- When data is abundant (数据充足时): Prefer hold-out or cross-validation (优先使用留出法或交叉验证法)
- When data is limited (数据较少时): Consider k-fold cross-validation (k can be increased appropriately) (考虑使用k折交叉验证（k可适当增大）)
- When data is very limited (数据非常有限时): Try bootstrap, but be aware of its bias (可尝试自助法，但需注意其偏差)
- Ensemble learning scenarios (集成学习场景): Bootstrap can generate multiple diverse training sets, helping improve model diversity (自助法可生成多个有差异的训练集，有助于提升模型多样性)

7.1 Dividing the Dataset

➢ Key Points:

- **Strictly distinguish the purposes of training, validation, and test sets. (严格区分训练集、验证集和测试集的用途)**
- **Test set can only be used once for final evaluation. (测试集只能使用一次，用于最终评估)**
- **Choose appropriate division method based on data size and task requirements. (根据数据规模和任务需求选择合适的划分方法)**
- **For important decisions, use multiple random divisions or cross-validation for more reliable results. (对于重要决策，使用多次随机划分或交叉验证以获得更可靠的结果)**

7.1 Dividing the Dataset

Practical Advice:

- **Use random seeds to ensure experiment reproducibility. (使用随机种子保证实验可复现性)**
- **Use stratified sampling for imbalanced data. (分层抽样处理类别不平衡数据)**
- **Record information of each division for traceability. (记录每次划分的信息便于追溯)**
- **Consider the special of time series data (avoid future information leakage). (考虑时间序列数据的特殊性（避免未来信息泄漏）)**

7.2 Underfitting & Overfitting
欠拟合&过拟合

机器学习的目标不是完美拟合训练数据，而是构建能够泛化到新数据的模型
The goal of machine learning is not to perfectly fit the training data, but to build models that generalize to new data.

7.2 Underfitting & Overfitting

Two Common Issues:

- 欠拟合 (Underfitting)
  - 模型过于简单，无法捕捉数据中的基本模式。
  - The model is too simple to capture the underlying patterns in the data.

- 过拟合 (Overfitting)
  - 模型过于复杂，不仅学习了规律，还学习了训练数据中的噪声和随机波动。
  - The model is too complex, learning not only the patterns but also the noise and random fluctuations in the training data.

➢ Key Metrics: Error rate & Error（误差率 & 误差）

- Error Rate (误差率/错误率)
  - Proportion of misclassified samples.
  - $E = a/m$ (m is total number of samples; a is misclassified samples)
- Accuracy (精度)
  - $Accuracy = 1\text{-}E$
- *Since we do not know the features of new samples beforehand, we can only strive to minimise empirical error;*
- *Although we can often achieve zero classification error on the training set, in most cases such a learner is not desirable.*

Key Metrics: Error rate & Error（误差率 & 误差）

- **Training Error (训练误差)**
  - **The error rate of the model on the training set. (模型在训练集上的错误率)**

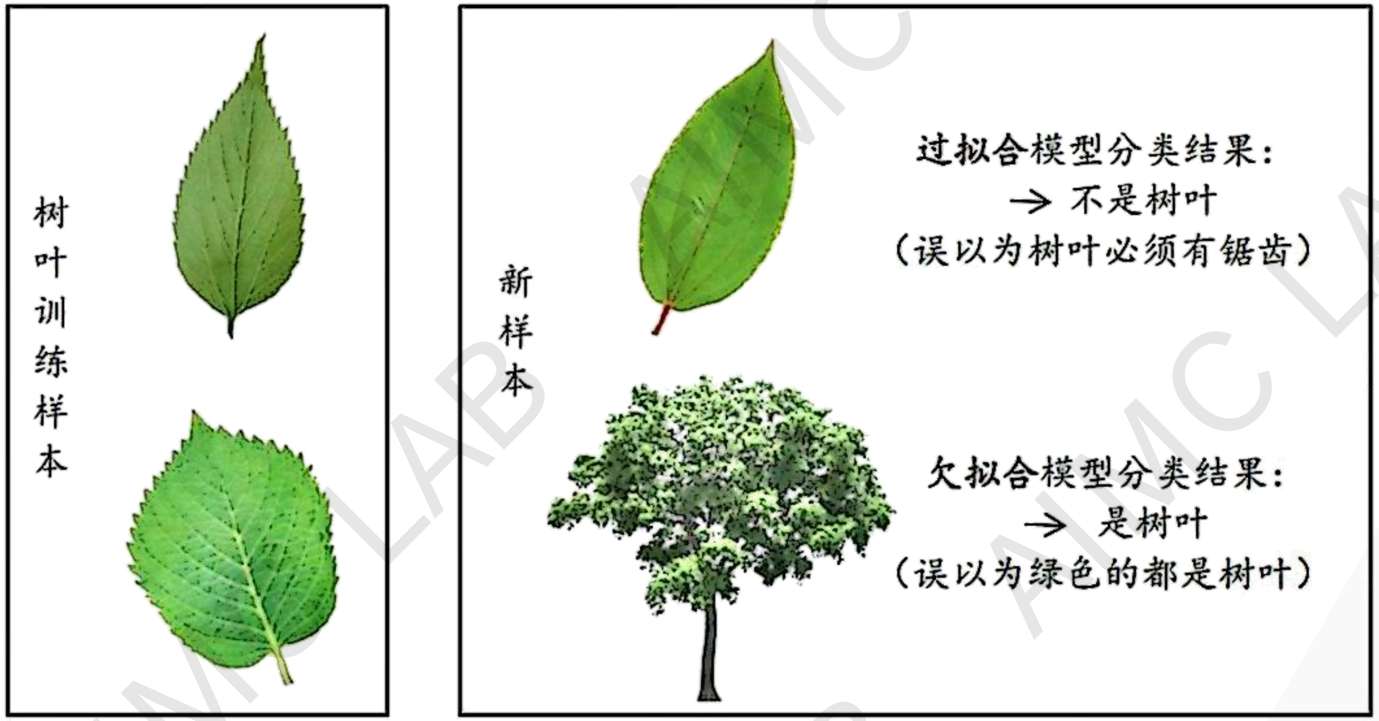
- ***泛化误差 / 测试误差 (Generalization / Test Error)***
  - **The expected error rate of the model on unseen data (test set).(模型在未见过的数据（测试集）上的预期错误率。)**

➢ Understanding Underfitting & Overfitting

Phenomenon	Training Error	Generalization Error	Model State
Underfitting	High	High	Too Simple
Good Fit	Low	Low	Appropriate
Overfitting	Low	High	Too Complex

7.2 Underfitting & Overfitting

➢ Understanding Underfitting & Overfitting

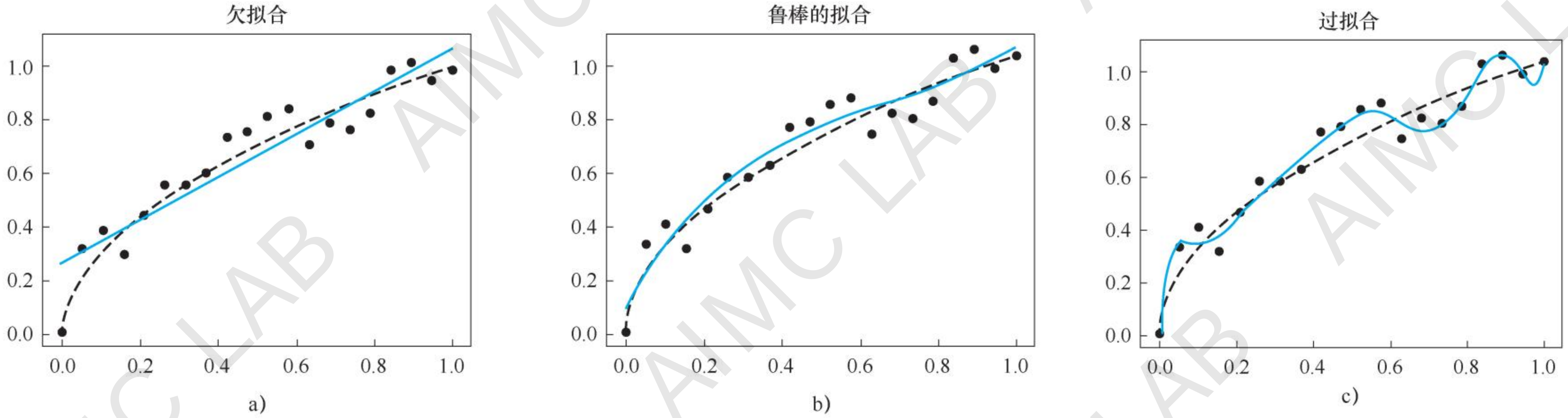


过拟合、欠拟合的直观类比

Overfitting: The learner treats the features of the training samples themselves as general properties that all potential samples will possess.

Underfitting: The general properties of the training samples have not been adequately learned by the learner.

➢ Understanding Underfitting & Overfitting



➢ Strategies to Address Underfitting

- **Increase Model Complexity (增加模型复杂度): Use more powerful models .**
- **Add More Features (添加更多特征): Introduce more informative features or feature combinations.**
- **Reduce Regularization Strength (减少正则化强度):Reduce constraints on model weights.**
- **Increase Training Epochs/Time (增加训练轮次/时间): Give the model more opportunities to learn patterns.**

➢ Strategies to Address Overfitting

- **Get More Training Data (获取更多训练数据): One of the most effective methods.**
- **Reduce Model Complexity (降低模型复杂度): Choose a simpler model architecture.**
- **Apply Regularization(应用正则化): L1/L2 regularization to increase model smoothness.**
- **Use Dropout - for Neural Networks (使用Dropout（神经网络）Randomly “turn off” a portion of neurons to prevent co-adaptation.**
- **Early Stopping (早停法): Stop training when validation set performance starts to degrade.**
- **Feature Selection / Dimensionality Reduction (特征选择/降维):Remove redundant or irrelevant features.**

7.2 Underfitting & Overfitting

➢ Key Takeaways

- 目标是低泛化误差，而非低训练误差。
- The goal is low generalization error, not low training error.
- 欠拟合和过拟合是模型能力与数据复杂度不匹配的表现。
- Underfitting and Overfitting indicate a mismatch between model capacity and data complexity.
- 需要通过验证集监控模型泛化性能来诊断问题。
- Diagnose issues by monitoring model generalization performance on a validation set.
- 没有“一招鲜”，需根据具体问题和数据选择应对策略。
- There is no “one-size-fits-all” solution; choose strategies based on the specific problem and data.

7.3 Inductive Preference
归纳偏好

Inductive preference is the "philosophy" behind your machine learning practice. Being aware of it helps you choose models more wisely and understand their limitations.
归纳偏好是你机器学习实践背后的“哲学”。意识到它可以帮助你更明智地选择模型并理解其局限性。

➢ The Core Question:

当多个模型在验证集上表现相似时，我们应该选择哪一个?
When multiple models show similar performance on the validation set, which one should we choose?

- **Answer:**
- **We need guiding principles—this is where Inductive Preference comes in.**
- **我们需要指导原则——这就是归纳偏好的作用。**

7.3 Inductive Preference

➢ What is Inductive Preference?

- **Inductive preferences are the heuristics or "values" that guide a learning algorithm to select one hypothesis over another in a vast hypothesis space. In other words, it is the algorithm's built-in bias toward certain types of solutions.**
- 归纳偏好是指导学习算法在庞大的假设空间中选择某个假设而非其他假设的启发式规则或“价值观”。换句话说，它是算法内置的对特定类型解决方案的偏好。
- **Key Insight: The alignment between an algorithm's inductive preference and the true nature of the problem often determines its success.**
- 关键洞察：算法的归纳偏好与问题真实性质之间的匹配度，常常决定其成败。

➢ What is Inductive Preference?

- **A carpenter (算法 / algorithm) has many tools (模型 / models).**
- **木匠（算法）拥有许多工具（模型）。**
- **Their preference for a hammer over a saw for driving nails is an inductive preference.**
- **他们倾向于用锤子而非锯子来钉钉子，这就是一种归纳偏好。**
- **This preference is generally good but may fail for special nails.**
- **这种偏好通常是好的，但对于特殊的钉子可能会失败。**

Principle 1: Occam's Razor (奥卡姆剃刀)

- "Entities should not be multiplied beyond necessity." Or more simply: Among competing hypotheses that explain the data equally well, choose the simplest one.
- “如无必要，勿增实体。” 或者更简单地说：在能同样好地解释数据的竞争假设中，选择最简单的那个。

Principle 1: Occam's Razor (奥卡姆剃刀)

- **Application in Machine Learning:**
  - **Favors models with fewer parameters, smoother functions, or shorter decision trees when performance is comparable.在性能相近时，偏好参数更少、函数更平滑或决策树更短的模型。**
  - **Acts as a guard against overfitting 作为防止过拟合的一种手段。**
  - **Simplicity here is not about computational ease, but about explanatory complexity.**
  - **这里的简单性不是指计算容易，而是指解释的复杂性。**

➢ Principle 2: No Free Lunch Theorem (NFL) (没有免费午餐定理)

- **"If algorithm A outperforms algorithm B on some set of problems, then there will necessarily exist other sets of problems where B outperforms A."**
- **“如果算法A在某些问题上比算法B表现更好，那么必然存在另一些问题集，在那些问题上B的表现优于A。”**

- **No universally best algorithm: There is no single learning algorithm that is inherently superior for all possible problems.没有普遍最优的算法：没有哪一个学习算法天生对所有可能的问题都更优。**

➢ Principle 2: No Free Lunch Theorem (NFL) (没有免费午餐定理)

- **Context is king: The discussion of which algorithm is "better" is meaningless without specifying the problem domain and data distribution.场景至上：如果不指定问题领域和数据分布，讨论哪个算法“更好”是没有意义的。**
- **Justifies specialization: It's okay and necessary to design algorithms tailored for specific types of problems (e.g., CNNs for images, RNNs for sequences).**
- **证明了专业化的合理性：为特定类型问题（例如，CNN用于图像，RNN用于序列）设计算法是合理且必要的。**

➢ Balancing the Two Principles

- **Occam's Razor gives a general guideline for selection within a given context.**
- **奥卡姆剃刀在给定场景内为选择提供了通用指南。**

- **No Free Lunch reminds us that there is no universal guideline valid across all contexts.**
- **没有免费午餐定理提醒我们，不存在跨所有场景都有效的通用指南。**

➢ Balancing the Two Principles

- **Define the problem and data context clearly. (明确界定问题和数据场景。)**
- **Among top performers, apply Occam's Razor to prefer simpler models, all else being equal. (在表现最佳的模型中，在其他条件相同时，应用奥卡姆剃刀偏好更简单的模型。)**
- **Accept that a model chosen this way may not perform well on a radically different problem. (接受这样选择的模型可能在完全不同的问题上表现不佳。)**

7.4 Performance Metrics
性能指标

Performance metrics are evaluation criteria for the generalization capability of models, which reflect task requirements. Using different metrics often leads to different conclusions.

性能指标是模型泛化能力的评估标准，反映了任务需求。使用不同的指标常常会导致不同的结论。

7.4 Performance Metrics

➢ Why Performance Metrics Matter? （为什么性能指标很重要？）

- We divide data to evaluate models (我们划分数据以评估模型)
- We monitor training to avoid under/overfitting (我们监控训练以避免欠/过拟合)
- Now we need quantitative measures to judge how good a model actually is (现在我们需要量化指标来判断模型到底有多好)

Performance metrics are evaluation criteria for the generalization capability of models, which reflect task requirements. Using different metrics often leads to different conclusions.

性能指标是模型泛化能力的评估标准，反映了任务需求。使用不同的指标常常会导致不同的结论。

Different Tasks, Different Metrics:

- **Regression (回归): Mean Squared Error, Mean Absolute Error, R-squared**

- **Classification (分类): Accuracy, Precision, Recall, F1-Score, ROC-AUC**

7.4 Performance Metrics

安徽理工大學
ANHUI UNIVERSITY OF SCIENCE & TECHNOLOGY

➢ Metrics for Regression Tasks (回归任务的性能指标)

- **Mean Squared Error (MSE) (均方误差)**

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

- **其中y_i是真实值，$\widehat{y_i}$是预测值。**
- ***Characteristics (特点):***
- ***Sensitive to large errors (对大的误差很敏感)***
- ***Commonly used as loss function (常用作损失函数)***

➢ Metrics for Regression Tasks (回归任务的性能指标)

- **Mean Absolute Error (MAE) (平均绝对误差)**

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|$$

- 其中y_i是真实值，$\widehat{y_i}$是预测值。
- *Characteristics (特点):*
- *More robust to outliers (对异常值更鲁棒)*
- *Interpretable in original units (可按原单位解释)*

➢ Metrics for Regression Tasks (回归任务的性能指标)

- **R-squared (Coefficient of Determination) (决定系数)**

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

- 其中y_i是真实值，$\widehat{y_i}$是预测值，$\overline{y_i}$是平均值。
- *Interpretation (解释): Proportion of variance explained by the model (模型解释的方差比例)*

➢ The Confusion Matrix(混淆矩阵) - Foundation for Classification

- Binary Classification Scenario (二分类场景):

	Predicted: Positive	Predicted: Negative
Actual: Positive	True Positive (TP)	False Negative (FN)
Actual: Negative	False Positive (FP)	True Negative (TN)

- TP (True Positive): Correctly predicted positive (正确预测的正例)
- FP (False Positive): Incorrectly predicted positive (Type I Error)
- FN (False Negative): Incorrectly predicted negative (Type II Error)
- TN (True Negative): Correctly predicted negative (正确预测的负例)
- All classification metrics are derived from these four basic counts.
- 所有分类指标都源自这四项基本计数。

Basic Classification Metrics

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

- Accuracy (准确率)
- Meaning : The proportion of ALL samples that are correctly classified.

  （被正确分类的样本占总样本的比例。）

- Error Rate (错误率)

$$\text{Error Rate} = 1 - \text{Accuracy} = \frac{FP + FN}{TP + TN + FP + FN}$$

- Best For: Perfectly balanced datasets where the cost of FP and FN is roughly equal.
- 最适用于：类别完全平衡的数据集，且假正例(FP)和假负例(FN)的代价相近。

➢ Basic Classification Metrics

- Precision (精确率/查准率)

$$\text{Precision} = \frac{TP}{TP + FP}$$

- "Of all samples predicted as positive, how many are actually positive?""在所有被预测为正例的样本中，有多少是真正的正例？ "
- The trustworthiness of a positive prediction. “When the model says ‘yes’， how often is it right?”模型做出正例预测的可信度。“当模型说‘是’的时候，它有多大概率是对的？ ”

➢ Basic Classification Metrics

- Web Search (示例 - 网络搜索): When you search for “Python”, you want the top results to be highly relevant (high precision). It‘s okay if some relevant pages are not on the first page (lower recall).
- 当你搜索“Python”时，你希望最前面的结果高度相关（高精确率）。可以接受一些相关页面没出现在第一页（较低召回率）。

Can we afford to miss a positive case?

➢ Basic Classification Metrics

$$\text{Recall} = \frac{TP}{TP + FN}$$

- Recall (召回率/查全率)
- The model‘s ability to find all relevant positive cases. “Of all the actual ‘yes’ cases, how many did we find?”
- 含义: 模型找出所有相关正例的能力。“在所有真实的‘是’中，我们找出了多少？”
- Example - Medical Screening (疾病筛查): In a cancer screening test, it is critical to identify every potential patient (high recall). It is acceptable to have some healthy people undergo further testing (lower precision, i.e., some FP).在癌症筛查中，识别出每一个潜在患者至关重要（高召回率）。可以接受一些健康的人接受进一步检查（较低精确率，即一些FP）。

7.4 Performance Metrics

➢ Basic Classification Metrics

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

- F1-Score: Harmonic Mean of Precision and Recall (精确率和召回率的调和平均数)The harmonic mean of Precision and Recall. It punishes extreme values. （精确率和召回率的调和平均数。它惩罚极端值。）
- Purpose: Provides a single, balanced score when you need to consider both Precision and Recall equally.当你需要同等考虑精确率和召回率时，提供一个单一的、平衡的分数。
- Best For: Situations where there is no clear preference for Precision over Recall or vice versa, especially with imbalanced datasets. It is more informative than accuracy here.

➢ Basic Classification Metrics

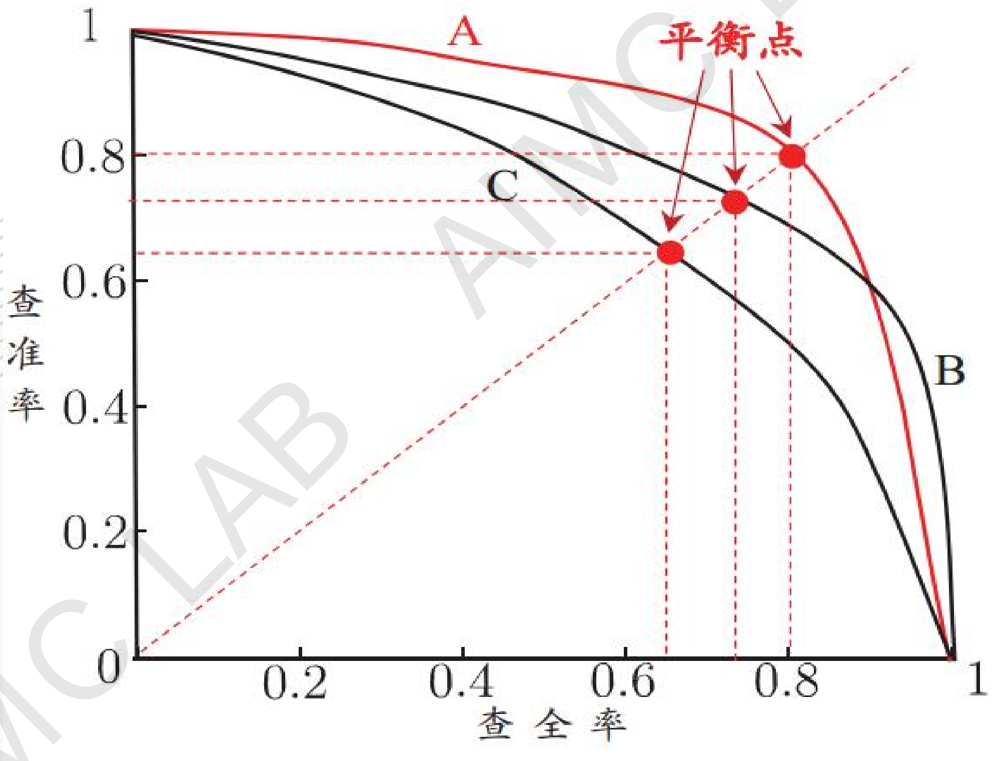
- Fβ分数

$$F_\beta = (1 + \beta^2) \times \frac{\text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}}$$

- β > 1 favors Recall, β < 1 favors Precision, β = 1 is standard F1 (β > 1 更看重召回率，β < 1 更看重精确率)

Basic Classification Metrics

- Precision-Recall (P-R) Curve (P-R曲线)
- Plot Precision vs. Recall at different classification thresholds (在不同分类阈值下绘制精确率 vs. 召回率)
- Better than ROC curve for imbalanced datasets (对于不平衡数据集，比ROC曲线更好)
- Break-Even Point (平衡点): Where Precision = Recall (精确率 = 召回率的点)



P-R曲线与平衡点示意图

7.4 Performance Metrics

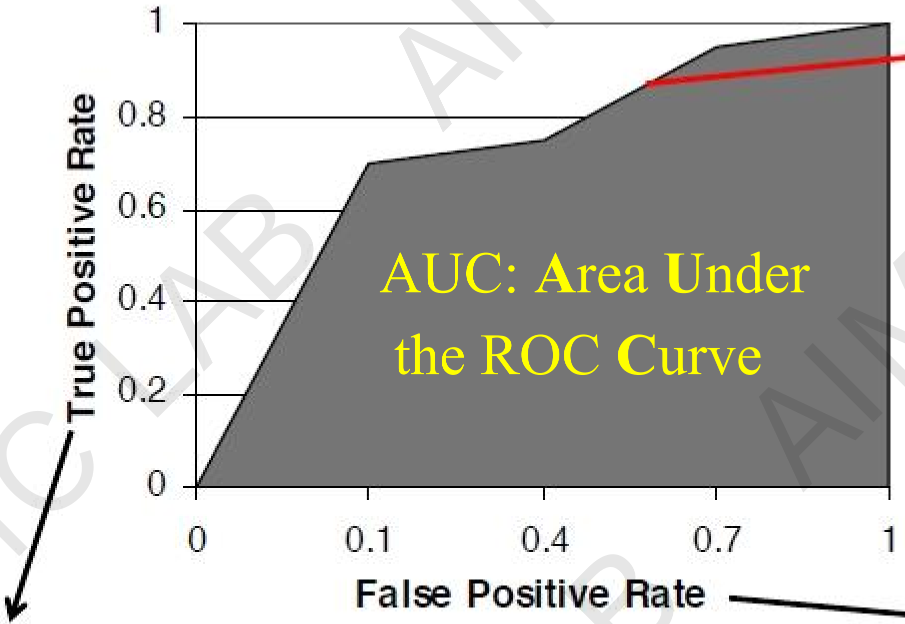
➢ Basic Classification Metrics

- ROC Curve (Receiver Operating Characteristic Curve) (受试者工作特征曲线)
- Plots True Positive Rate (TPR) vs. False Positive Rate (FPR) at various thresholds
- (在不同阈值下绘制真正率(TPR) vs. 假正率(FPR))
- TPR = Recall = Sensitivity (真正率 = 召回率 = 灵敏度)
- $\boldsymbol{FPR = \frac{FP}{PF+TN}}$ (假正率）它是被错误标记为正例的负例样本的比例。可以把它理解为“误报率”。

7.4 Performance Metrics

➢ Basic Classification Metrics

- ROC Curve



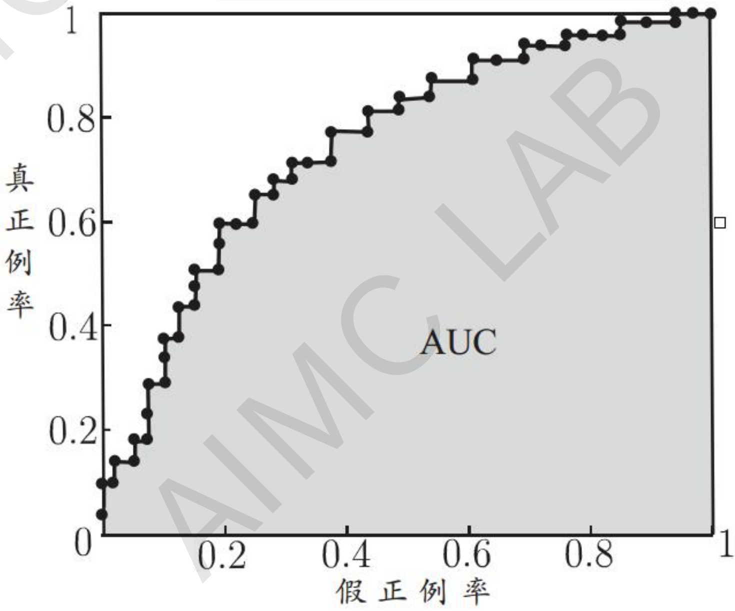
ROC (Receiver Operating Characteristic) Curve

$$tpr = \frac{TP}{TP+FN} = \frac{TP}{m_+}$$

$$FPR = \frac{FP}{PF + TN}$$

Basic Classification Metrics

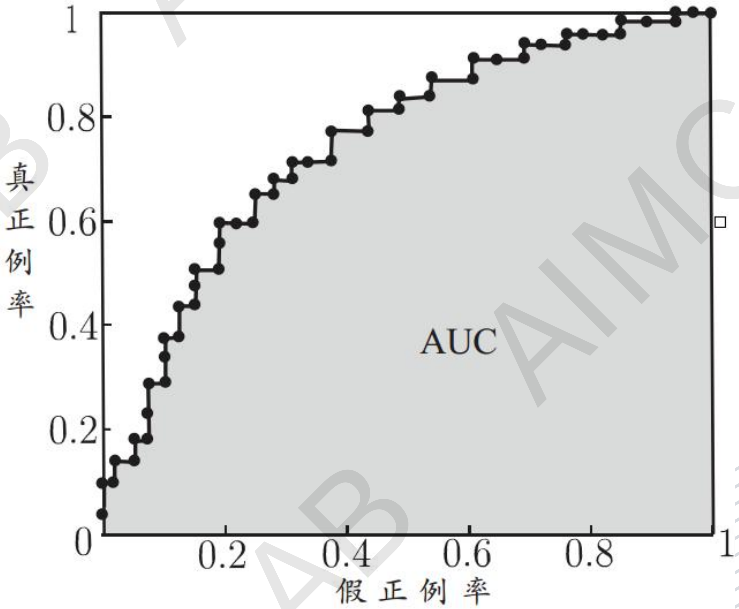
- ROC Curve：类比解释
- 你是一名侦探，手头有100个人的档案。
- 事实是： 其中只有 10个人 是真正的罪犯（正例）。
- 剩下的 90个人 都是无辜群众（负例）。



基于有限样例绘制的 ROC 曲线

- 你的任务不是立刻抓人，而是根据手中的情报（模型预测的“犯罪概率”分数），把这100个人从“最可疑”到“最不可能”排个序。

7.4 Performance Metrics



基于有限样例绘制的 ROC 曲线

➢ Basic Classification Metrics

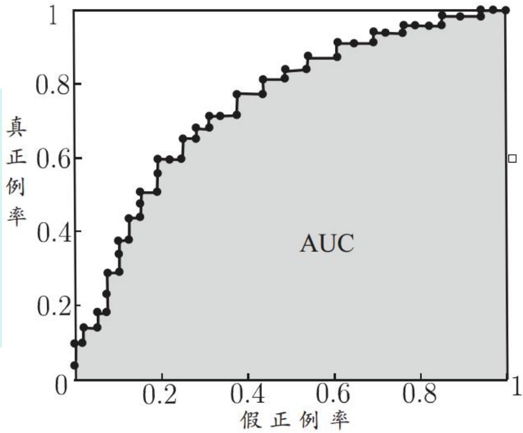
- ROC Curve：类比解释
- X轴（假警报率 - FPR）：
  - 这衡量的是 你有多“扰民”。
  - 公式是：被你错当成坏人的好人数量 / 所有好人的总数 (90人)。
  - 越低越好。0 表示没冤枉一个好人；1 表示你把所有好人都当成坏人了（全抓错了）。
- Y轴（抓捕率 - TPR/Recall）：
  - 这衡量的是 你有多“能干”。
  - 公式是：被你成功找出来的真正坏人数量 / 所有坏人的总数 (10人)。
  - 越高越好。1 表示你把所有坏人都揪出来了（完美破案）。

7.4 Performance Metrics

➢ Basic Classification Metrics

- 标准非常严苛（阈值很高）： 比如，只有证据确凿、100%肯定是坏人的才抓。
  - 结果： 你几乎不会抓错好人（FPR很低），但很多狡猾的坏人可能因为证据不足而漏网（TPR也很低）。
  - 在ROC图上，这个点靠近左下角 (0, 0)。
- 标准非常宽松（阈值很低）： 比如，“宁可错抓一千，不可放过一个”，看起来有点可疑的都先抓起来。
  - 结果： 你确实抓到了大部分坏人（TPR很高），但也会冤枉一大堆好人（FPR也很高）。
  - 在ROC图上，这个点靠近右上角 (1, 1)。

ROC曲线，就是你从“最严苛”到“最宽松”不断调整标准时，所得到的一系列（FPR, TPR）点连成的线。



Basic Classification Metrics

- Plotting the ROC Curve
- Data Preparation Phase: Let the dataset contain m^+ positive examples and m^- negative examples.
- Sort all samples in descending order by the learner's predicted scores.
- Dynamic Threshold Update: Sequentially take each sample's predicted value as the classification threshold, initializing the coordinate origin at (0,0).

➢ Basic Classification Metrics

- Plotting the ROC Curve
- Coordinate Recursion Rule:
- For positive examples: New coordinates (x, y + 1/(m+))
- For negative examples: New coordinates (x + 1/(m−), y)
- Curve Generation Mechanism: Connect each coordinate point sequentially with straight lines to form a stepped ROC curve.

Basic Classification Metrics

Plotting the ROC Curve

待测样本	x1	x2	x3	x4	x5	x6	x7	x8	x9
样本标记	−	+	−	−	+	−	+	+	−
P(+|x_i)	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1

┃ − + − − + − + + −

− ┃ + − − + − + + −

− + ┃ − − + − + + −

− + − ┃ − + − + + −

− + − − + − + + ┃ −

− + − − + − + + − ┃

TPR=0/4;FPR=0/5

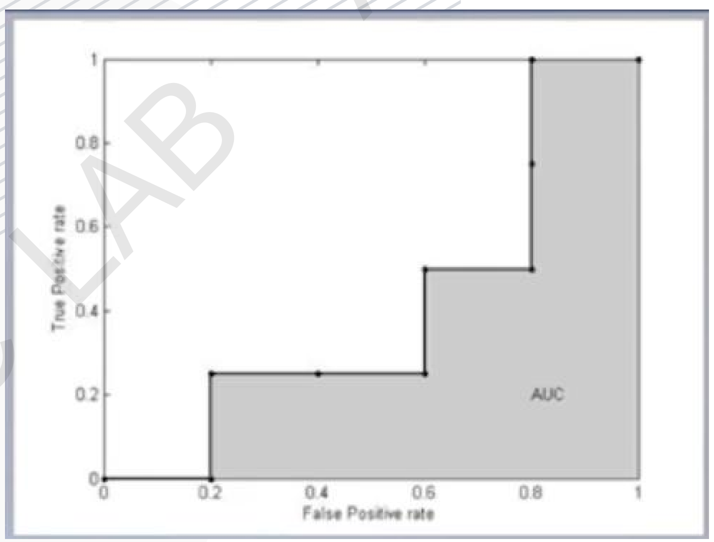
TPR=0/4;FPR=1/5

TPR=1/4;FPR=1/5

TPR=1/4;FPR=2/5

TPR=4/4;FPR=4/5

TPR=4/4;FPR=5/5



Basic Classification Metrics

- ROC Curve (Receiver Operating Characteristic Curve) (受试者工作特征曲线)
- Diagonal Line (y = x): Represents the performance of a random classifier (AUC = 0.5).代表随机分类器的性能（AUC = 0.5）。
- Top-Left Corner (0, 1): Represents a perfect classifier (FPR=0, TPR=1).代表完美分类器（FPR=0, TPR=1）。
- A Good Model: Its ROC curve bulges towards the top-left corner.
- 其ROC曲线向左上角凸起。

Basic Classification Metrics

- AUC Purpose & Role (AUC的目的与作用):
- Primary Purpose (主要目的): To provide a single scalar value that summarizes the entire ROC curve. It represents the probability that a randomly chosen positive instance is ranked higher (has a higher predicted probability/score) than a randomly chosen negative instance.
- 提供一个单一的标量值来概括整个ROC曲线。它表示随机选择一个正例样本，其排名（预测概率/得分）高于随机选择一个负例样本的概率。
- AUC衡量了样本预测的排序质量。

7.4 Performance Metrics

➢ Basic Classification Metrics

- AUC Purpose & Role (AUC的目的与作用):
- Interpretation (解读):
- AUC = 1.0: Perfect ranking. (完美排名。)
- AUC = 0.5: No discriminative power (like random guessing). (没有区分能力（如同随机猜测）。)
- AUC between 0.5 and 1.0: The higher, the better the model's ranking ability. (数值越高，模型的排序能力越好。)
- 它是与阈值无关的。你不需要确定一个具体的决策阈值就能评估模型区分类别的根本能力。这使其非常适合初步模型筛选和比较。

7.4 Performance Metrics

➢ Basic Classification Metrics

Feature (特点)	**Precision-Recall (P-R) Curve**	**ROC Curve**
Focus (关注点)	Performance on the positive (minority) class.正例（少数）类的性能。	Overall ranking ability between both classes.两类之间的整体排序能力。
Best for	Highly imbalanced datasets where you care most about the rare class.高度不平衡的数据集，且你最关心稀有类。	Moderately balanced or imbalanced datasets for overall model comparison.中等平衡或不平衡的数据集，用于整体模型比较。
X-axis (X轴)	Recall (TPR)	False Positive Rate (FPR)
Y-axis (Y轴)	Precision	True Positive Rate (TPR/Recall)
Baseline (基线)	A horizontal line at the ratio of positives in the dataset. 数据集中正例比例处的一条水平线。	The diagonal line (y=x). 对角线 (y=x)。

➢ Cost-Sensitive Error Rate (代价敏感错误率)

- **In standard classification, we treat all errors as equally bad. But in reality:**
- **在标准分类中，我们将所有错误视为同等严重。但在现实中：**

Error Type (错误类型)	Example Scenario (示例场景)	Real-World Consequence (实际后果)
False Positive (FP) (假正例)	Medical test says you have cancer (but you don't) 医疗检测说你患癌（但你并没有）	Unnecessary anxiety, invasive tests, wasted resources 不必要的焦虑、侵入性检查、资源浪费
False Negative (FN) (假负例)	Medical test says you're healthy (but you have cancer) 医疗检测说你健康（但你患癌）	Delayed treatment, disease progression, potential fatality 延误治疗、疾病恶化、潜在死亡风险

Cost-Sensitive Error Rate (代价敏感错误率)

- The consequences of different types of errors in real-world tasks are likely to vary. To weigh the differing losses caused by various error types, an 'unequal cost' may be assigned to errors.
- For example, in binary classification, a "cost matrix" may be defined based on domain knowledge, as shown in the table below. Here, $cost_{ij}$ denotes the cost of predicting class *i* as class *j*. The greater the loss severity, the larger the difference between $cost_{01}$ and $cost_{10}$.

➢ Cost-Sensitive Error Rate (代价敏感错误率)

- **Under non-uniform cost, the objective shifts from minimising the number of errors to minimising the 'overall cost'. The corresponding cost-sensitive error rate is then:**

$$E(f;D;cost)=\frac{1}{m}\left(\sum_{\boldsymbol{x}_i\in D^+}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)\neq y_i\right)\times cost_{01}+\sum_{\boldsymbol{x}_i\in D^-}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)\neq y_i\right)\times cost_{10}\right)$$

➢ Cost curve

- ROC curves show performance across thresholds but don't directly account for different costs of errors. They assume equal cost for FP and FN.
- ROC曲线展示了不同阈值下的性能，但没有直接考虑不同错误的代价。它假设FP和FN的代价相等。
- The horizontal axis of the cost curve represents the positive example probability cost with values in the range [0,1].

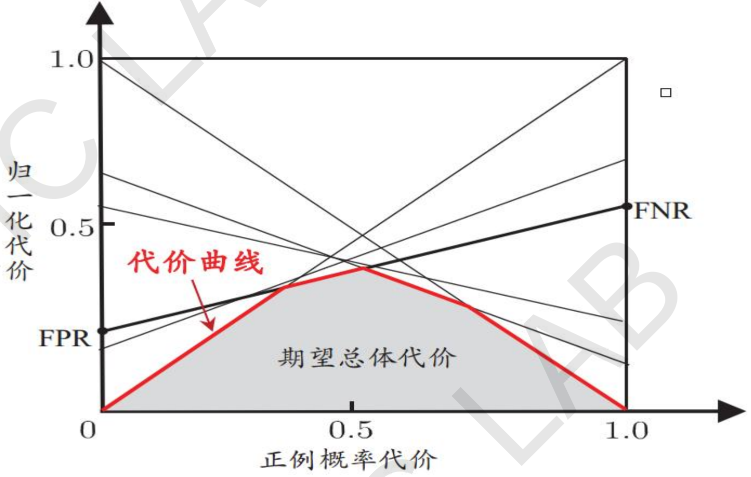
$$P(+)cost = \frac{p \times cost_{01}}{p \times cost_{01} + (1-p) \times cost_{10}}$$

- The vertical axis represents the normalized cost with values ranging from [0,1].

$$cost_{norm} = \frac{\mathrm{FNR} \times p \times cost_{01} + \mathrm{FPR} \times (1-p) \times cost_{10}}{p \times cost_{01} + (1-p) \times cost_{10}}$$

7.4 Performance Metrics

➢ Cost curve



- **Cost Curve Drawing Method:**
- **1. Single-Point Transformation: For any point (FPR, TPR) on the ROC curve, calculate its false negative rate (FNR = 1 - TPR), then connect the endpoints (0, FPR) and (1, FNR) on the cost plane to form a line segment;**
- **2. Area Overlay Principle: The region under this line segment represents the expected total cost at the current classification threshold;**
- **3. Global Lower Bound Fitting: Generate line segments for all points on the ROC curve. The lower bound envelope formed by these segments represents the minimum expected overall cost of the learner across all scenarios.**

7.5 Bias and Variance
偏差与方差

Understanding Model Error

理解模型误差

7.5 Bias and Variance

➢ Why Do Models Make Mistakes?

- **When a model fails to make accurate predictions, the errors come from three fundamental sources:（当模型无法做出准确预测时，误差来自三个基本来源）**

Bias (偏差)	Systematic error in aiming (瞄准的系统性误差) Like always shooting too high 就像总是射得太高	Using linear regression to fit nonlinear data 用线性回归拟合非线性数据
Variance (方差)	Inconsistency in shooting (射击的不一致性)Like shots scattered widely around target就像子弹在目标周围分散很广	Complex model overfitting to training noise 复杂模型对训练数据噪声过拟合
Noise (噪声)	Inherent randomness in the target (目标的固有随机性)Like wind affecting arrow flight 就像风影响箭的飞行	Measurement errors, random fluctuations 测量误差、随机波动

7.5 Bias and Variance

➢ Why Do Models Make Mistakes?

- **Symbol Definitions and Problem Modeling:**
  - **Let the test sample be denoted as x, which has data annotations y_D (observations) and true labels y (true values). The prediction function generated by the learner on the training set D is represented as f (x;D). For regression tasks, the expected prediction error of the learning algorithm can be formally defined as:**

$$\bar{f}(\boldsymbol{x}) = \mathbb{E}_D[f(\boldsymbol{x}; D)]$$

  - **The variance produced by different training sets with the same sample number is** $var(\boldsymbol{x}) = \mathbb{E}_D\left[\left(f(\boldsymbol{x}; D) - \bar{f}(\boldsymbol{x})\right)^2\right]$

➢ Why Do Models Make Mistakes?

- The difference between the expected output and the actual label is called the bias, i.e. $bias^2(\boldsymbol{x}) = \left(\bar{f}(\boldsymbol{x}) - y\right)^2$

- For the convenience of discussion, assume the noise expectation is zero, i.e., $\mathbb{E}_D[y_D - y] = 0$, the generalization error decomposition.

$$
\begin{aligned}
E(f;D) &= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x}) + \bar{f}(\boldsymbol{x}) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y_D\right)^2\right] \\
&\quad + \mathbb{E}_D\left[2\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)\left(\bar{f}(\boldsymbol{x}) - y_D\right)\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y_D\right)^2\right]
\end{aligned}
$$

$\bar{f}(\boldsymbol{x})$是所有数据集上的期望

交叉项的期望为 0

7.5 Bias and Variance

➢ Why Do Models Make Mistakes?

$$= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y+y-y_D\right)^2\right]$$

$$= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y\right)^2\right]+\mathbb{E}_D\left[\left(y-y_D\right)^2\right]$$

$$+\,2\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y\right)\left(y-y_D\right)\right]$$

Furthermore, assuming the noise expectation is zero, we obtain

$$E(f;D)=\mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\left(\bar{f}(\boldsymbol{x})-y\right)^2+\mathbb{E}_D\left[\left(y_D-y\right)^2\right]$$

1、The difference between expected output and actual output

2、The performance variation caused by changes in training sets of the same size

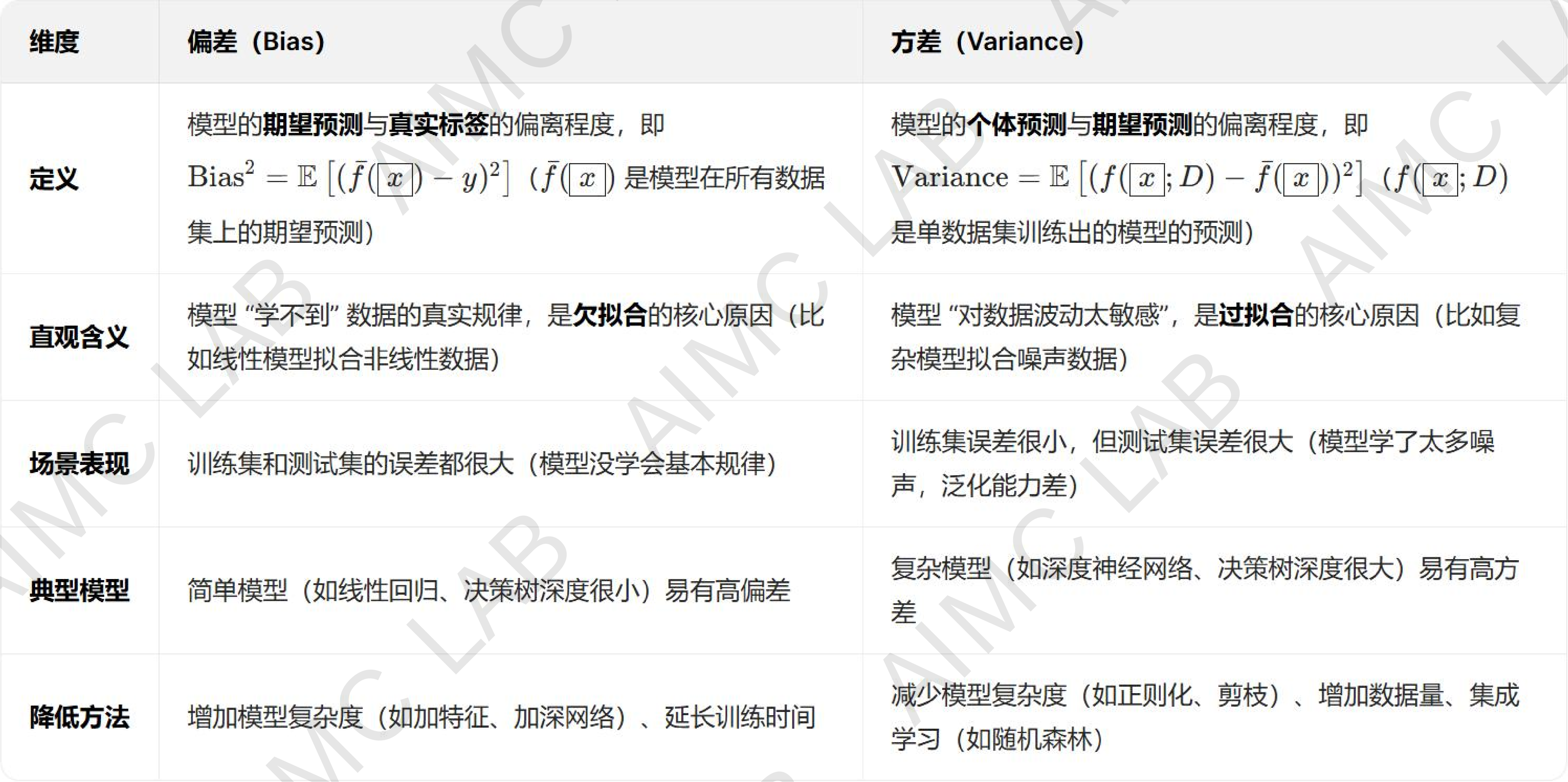
3、The difference between training sample labels and actual labels

So then: $E(f;D)=bias^2(\boldsymbol{x})+var(\boldsymbol{x})+\varepsilon^2$

泛化误差可分解 偏差 方差与噪声 .

7.5 Bias and Variance

➢ Why Do Models Make Mistakes?



维度	偏差（Bias）	方差（Variance）
定义	模型的**期望预测**与**真实标签**的偏离程度，即 $\text{Bias}^2 = \mathbb{E}\left[(\bar{f}(\boxed{x}) - y)^2\right]$（$\bar{f}(\boxed{x})$ 是模型在所有数据集上的期望预测）	模型的**个体预测**与**期望预测**的偏离程度，即 $\text{Variance} = \mathbb{E}\left[(f(\boxed{x}; D) - \bar{f}(\boxed{x}))^2\right]$（$f(\boxed{x}; D)$ 是单数据集训练出的模型的预测）
直观含义	模型 “学不到” 数据的真实规律，是**欠拟合**的核心原因（比如线性模型拟合非线性数据）	模型 “对数据波动太敏感”，是**过拟合**的核心原因（比如复杂模型拟合噪声数据）
场景表现	训练集和测试集的误差都很大（模型没学会基本规律）	训练集误差很小，但测试集误差很大（模型学了太多噪声，泛化能力差）
典型模型	简单模型（如线性回归、决策树深度很小）易有高偏差	复杂模型（如深度神经网络、决策树深度很大）易有高方差
降低方法	增加模型复杂度（如加特征、加深网络）、延长训练时间	减少模型复杂度（如正则化、剪枝）、增加数据量、集成学习（如随机森林）

➢ Why Do Models Make Mistakes?

- **Fitting highly nonlinear housing price data with linear regression**
- **High bias (underfitting, failed to capture price patterns);**

- **Fitting a simple task with only 10 samples using a deep neural network**
- **High variance (overfitting, learned sample noise, leading to inaccurate predictions with different samples).**

7.5 Bias and Variance

Three-stage evolution of bias-variance tradeoff:

1. Underfitting stage: When model complexity is insufficient, bias dominates (high bias, low variance)

2. Optimization transition stage: As training progresses, the model gradually converges toward the variance-dominant interval

3. Overfitting risk zone: After sufficient training, variance becomes the primary conflict (low bias, high variance)

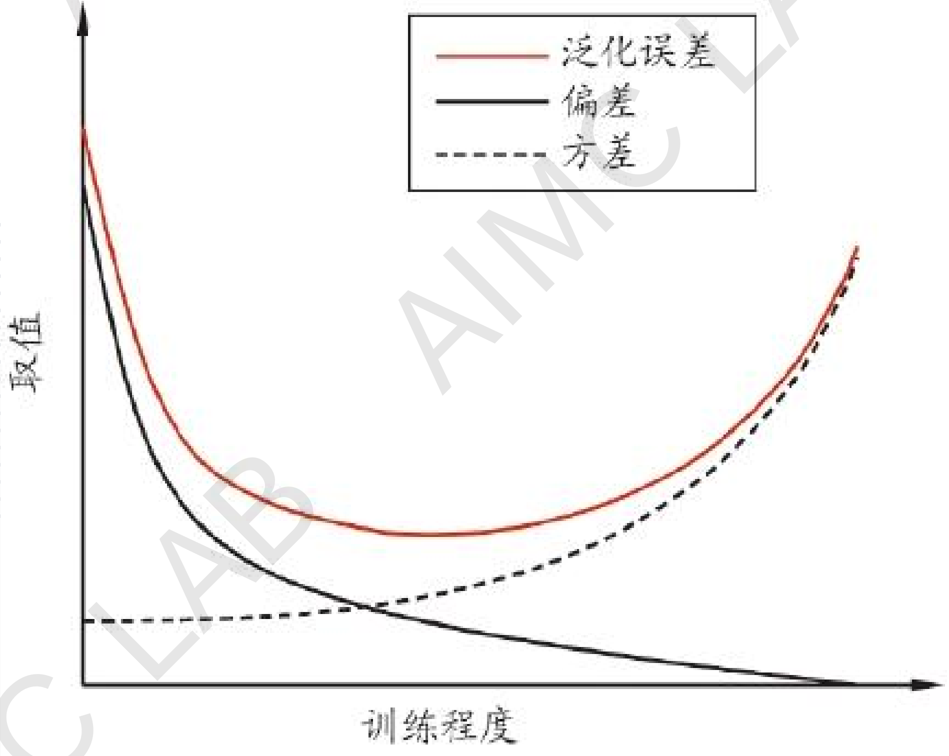


图 2.9　泛化误差与偏差、方差的关系示意图