交叉验证

Gao

数学学院（珠海）LaTeX

版本：1.0

日期：2025 年 4 月 13 日

1 基本概念

在机器学习中，为了评估模型的性能，我们通常将数据集分为三个部分：训练集（Training Set）、验证集（Validation Set）和测试集（Test Set）。

1.1 训练集（Training Set）

训练集用于训练模型，即通过学习训练集中的数据来调整模型参数。

1.2 验证集（Validation Set）

验证集用于在模型训练过程中进行模型选择和超参数优化。通过在验证集上评估模型性能，我们可以调整模型结构或参数，以提高模型的泛化能力。

1.3 测试集（Test Set）

测试集用于最终评估模型的性能。测试集应该完全独立于训练过程，只在模型训练完成后使用一次，以评估模型在未知数据上的表现。

2 交叉验证

交叉验证是一种通过将数据集划分为多个子集，并在这些子集上重复训练和验证模型，以提升模型泛化性能的模型训练技术。

3 交叉验证示意图

交叉验证将数据集分成K个大小相等（或尽可能相等）的子集，每次使用其中的一个子集作为验证集，而剩下的$K-1$个子集作为训练集。这个过程重复K次，每次选择不同的子集作为验证集，最终得到K个模型性能评估结果，并计算这些结果的平均值作为最终的模型评估结果。

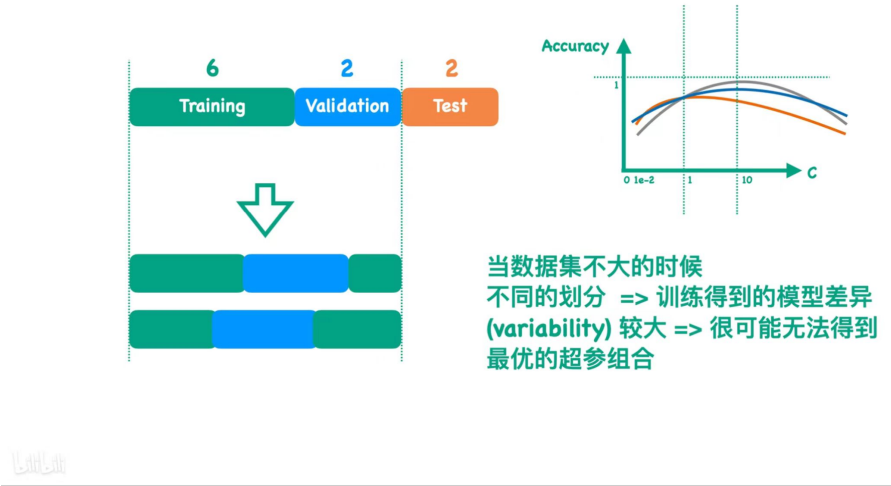

图 1: 数据集划分示意图

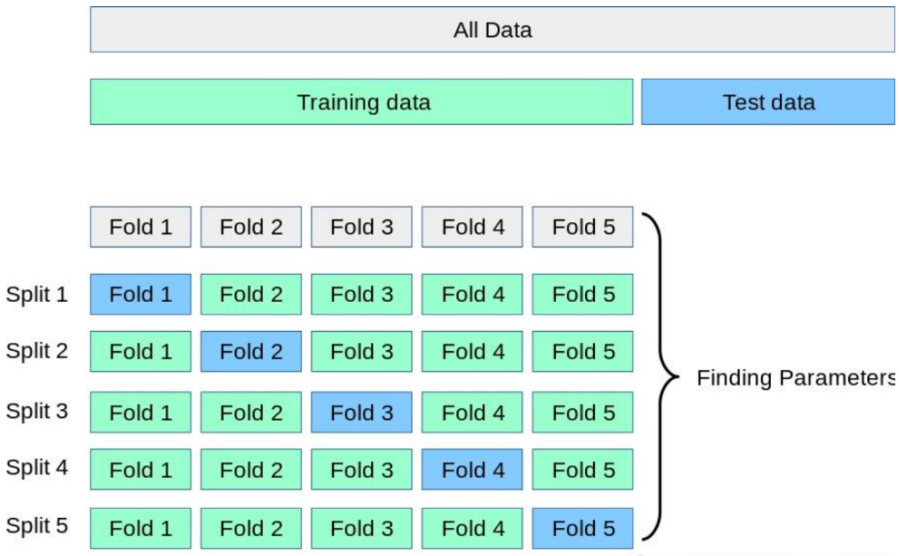

图 2: 交叉验证示意图

4 交叉验证的类型

4.1 K折交叉验证

K折交叉验证是最常见的交叉验证方法。它将数据集分为K个子集，每次选择一个子集作为验证集，其余$K-1$个子集作为训练集。这个过程重复K次，每次选择不同的子集作为验证集。最终的模型性能评估结果是所有K次评估结果的平均值。

$$\text{CV}_{K\text{-fold}} = \frac{1}{K}\sum_{i=1}^{K}\text{Model}_i \tag{1}$$

4.2 留一交叉验证

留一交叉验证（Leave-One-Out Cross-Validation, LOOCV）是K折交叉验证的一个特例，其中K等于数据集中的样本数量。每次选择一个样本作为验证集，其余所有样本作为训练集。这个过程重复N次（N为样本数量），每次选择不同的样本作为验证集。最终的模型性能评估结果是所有N次评估结果的平均值。

$$\text{LOOCV} = \frac{1}{N}\sum_{i=1}^{N}\text{Model}_i \tag{2}$$

4.3 广义交叉验证

广义交叉验证（Generalized Cross-Validation, GCV）是一种统计方法，用于评估模型的预测性能，特别是在回归分析中。它提供了一种更集成的方法，直接从数据中估计预测误差，而无需显式数据拆分。这种方法在处理过拟合风险较高的小型数据集时特别有用。广义交叉验证的数学公式植根于最小化预测误差的概念。具体来说，GCV 根据残差平方和计算得分，并通过考虑模型复杂性的惩罚项进行调整。这个惩罚至关重要，因为它可以防止模型变得过于复杂，而这会导致对未知数据的泛化能力较差。广义交叉验证是一种改进的交叉验证方法，它不仅考虑了训练误差，还考虑了模型的复杂度。广义交叉验证的公式如下：

$$\text{GCV} = \frac{\text{RSS}}{(N-2p)^2}\cdot\frac{N}{N-p-1} \tag{3}$$

其中，RSS表示残差平方和，p表示模型参数的数量，N表示样本数量。

5 GCV对数据集的处理方式

GCV本质上是一种留一交叉验证（LOOCV）的形式。它将每个数据点依次作为测试集，然后在整个数据集上对加权预测误差求平均。具体步骤如下：

1. **模型拟合**：使用训练数据对模型进行拟合，并针对不同的正则化参数值计算帽子矩阵。
2. **计算GCV得分**：对于每个正则化参数，利用公式计算GCV得分，该公式涉及到残差平方和、帽子矩阵的迹以及观测值的数量。
3. **选择最优参数**：选择使GCV得分最小的正则化参数，该参数代表了模型拟合与复杂度之间的最佳平衡。

5.1 GCV得分的计算

GCV得分的计算公式如下：

$$\text{GCV}(\lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\left(\frac{n}{n-\text{tr}(H(\lambda))}\right)^2} \tag{4}$$

其中：

- y_i 是第 i 个观测值的实际值。

- $\hat{y}_i$ 是第 i 个观测值的预测值。
- n 是观测值的总数。
- $\text{tr}(H(\lambda))$ 是帽子矩阵 $H(\lambda)$ 的迹，它依赖于正则化参数 λ。

6 GCV与普通交叉验证方法的区别

6.1 计算效率

普通交叉验证，如k折交叉验证，需要将数据集分成k个子集，然后进行k次模型训练和测试，计算量较大。而GCV通过巧妙的数学推导，无需显式地进行多次拟合，就能估计出预测误差，在处理大型数据集或复杂模型时，计算效率更高。

6.2 适用场景

普通交叉验证适用于各种模型和数据集，常用于模型选择和超参数调整。GCV则更侧重于线性模型和正则化问题，特别是在选择岭回归的正则化参数或样条平滑的平滑参数时表现出色。

6.3 对数据划分的依赖程度

普通交叉验证的结果可能会受到数据划分方式的影响，不同的划分可能导致不同的评估结果。GCV是基于整个数据集进行计算的，对数据划分的依赖较小，结果相对更稳定。

7 GCV的实际应用优势

1. **模型选择**：在正则化模型中，GCV可以帮助选择最优的正则化参数，从而在模型复杂度和拟合度之间找到平衡。
2. **预测性能评估**：GCV提供了一种无需额外数据集分割的方法来评估模型的预测性能，这对于数据量有限的情况特别有用。
3. **计算效率**：在处理大型数据集或复杂模型时，GCV的计算效率优势明显，可以节省大量的计算资源。

8 总结

交叉验证是一种有效的模型评估方法，可以帮助我们评估模型的泛化能力。K折交叉验证、留一交叉验证和广义交叉验证各有优缺点，适用于不同的场景。选择合适的交叉验证方法可以显著提高模型的性能和可靠性。