《实用机器学习》课程报告

班级 大数据 22-1　学号 2204050108　姓名　郭子铭

I 题目

本任务功能包含两步:使用修正基础模型预测发电出力值,以及使用发电出力值预测厂用率。以下是任务介绍:

1. 预测数据修正模型

在电力预测数据修正项目中，我们需要开发了一个预测数据修正模型，以解决实际观测数据和预测数据之间的偏差问题。通过对二者关系建模，可以将基础预测数据调整得更接近实际观测值。如图 I-1 所示，具体方法是利用历史的实际观测数据和预测数据来训练模型。使用时，输入包括预测数据、日期、预测日期以及预测类型（如风电、光伏、供电）。模型输出修正后的预测数据。

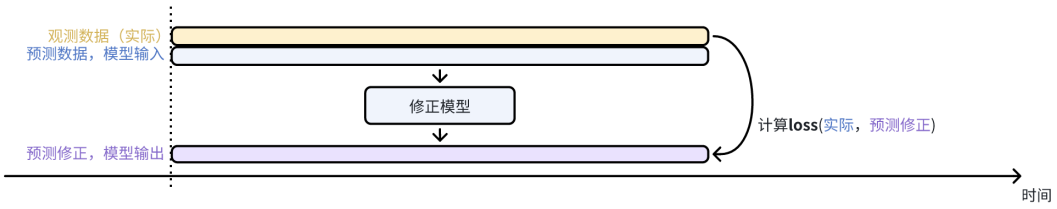


图 I-1 预测数据修正模型工作示意图

例如，输入预测数据为 1200.5，日期为 2024 年 5 月 7 日，预测日期为 2024 年 5 月 9 日，预测类型为光伏，模型将输出修正后的数据例如 1351.3（需要拟合真实数据）。

2. 发电出力值预测厂用率模型

发电厂的厂用电率是指单位时间内厂用变耗电量与发电量的百分比。

$$厂用电率 = \frac{机端发电量 - 上网电量}{机端发电量} \times 100\%$$

在发电出力值预测中，我们针对四种发电类型，分别预测其对应的厂用率。首先，利用已有的发电出力和厂用率数据训练模型。在实际应用中，输入该类型的发电出力值，模型将输出对应的厂用率均值和预测区间。

例如,对于风电,输入发电出力值为 5297.8105 时,模型预测的厂用率区间为-1.5161074%到 3.4591386%。这一方法通过对历史数据进行建模，提供了一个可靠的估计，使得电力资

源的管理更加精准和高效。

II 报告

1、机器学习技术概述

（1）请阐述所设计数据分析项目的背景、目标及数据来源；

项目背景：

在电力系统中，准确的负荷预测和出力值预测对于电网的稳定运行至关重要。由于实际观测数据和预测数据之间常存在偏差，导致电力资源配置效率降低。因此，我们设计了一个数据分析项目，旨在提高预测精度，优化电力资源配置。

项目目标：

- 提高预测精度：通过修正模型减少实际观测数据与预测数据之间的偏差。
- 优化电力资源配置：利用更准确的预测数据，提升电网稳定性和运行效率。
- 开发厂用率预测模型：根据不同发电类型，预测出力值对应的厂用率，提供可靠的决策支持。

数据来源：

本项目的数据来自国网黑龙江省电力有限公司电力科学研究院。数据包括历史的实际观测数据和预测数据，涵盖风电、光伏、供电等多种类型。这些数据为模型训练和验证提供了坚实的基础，确保预测结果的可靠性和准确性。

（2）根据机器学习的任务不同可将机器学习模型分为回归、分类和聚类三个类别，请解释这三类机器学习模型的任务，并结合自拟的数据分析项目，说明使用哪类模型才能达到预期数据分析目标。

分类（Classification）

任务解释：

- 分类用于预测离散的类别标签。
- 适合场景：当目标变量是离散类别时，比如邮件分类（垃圾邮件或正常邮件）、疾病诊断（如“是”或“否”）。

方法案例：

- 支持向量机（SVM）：通过找到最佳分隔超平面将不同类别的样本分开。
- 随机梯度下降（SGD）：用于优化分类模型，适合大规模数据。
- 贝叶斯分类器：基于贝叶斯定理，适合文本分类任务。

- 集成方法（如随机森林）：通过多个决策树的组合提高预测准确性。
- K 近邻（KNN）：通过测量与训练样本的距离进行分类。

示例：

- 输入患者数据（如年龄、症状），预测是否患有某种疾病。

回归（Regression）

任务解释：

- 回归用于预测连续数值输出。
- 适合场景：当目标变量是连续的实数，比如房价预测、负荷预测。

方法案例：

- 线性回归：通过拟合直线来预测目标变量。
- 支持向量回归（SVR）：使用支持向量机的概念应用于回归问题。
- 随机梯度下降（SGD）：用于优化回归模型，适合大规模数据。
- 集成方法（如梯度提升树）：通过多个回归树的组合提高预测精度。

示例：

- 预测电力负荷或房价，根据历史数据和相关特征。

聚类（Clustering）

任务解释：

- 聚类是一种无监督学习方法，用于将数据分成多个组或簇。
- 适合场景：没有标签数据时，用于发现数据的内在结构，比如市场细分。

方法案例：

- K-均值（K-means）：通过迭代优化簇中心，最小化簇内的平方误差。
- 高斯混合模型（GMM）：假设数据是由多个高斯分布组成，使用期望最大化算法进行优化。
- 层次聚类：通过构建层次树状结构对数据进行聚类。

示例：

- 对客户数据进行聚类，识别不同的客户群体以进行目标营销。

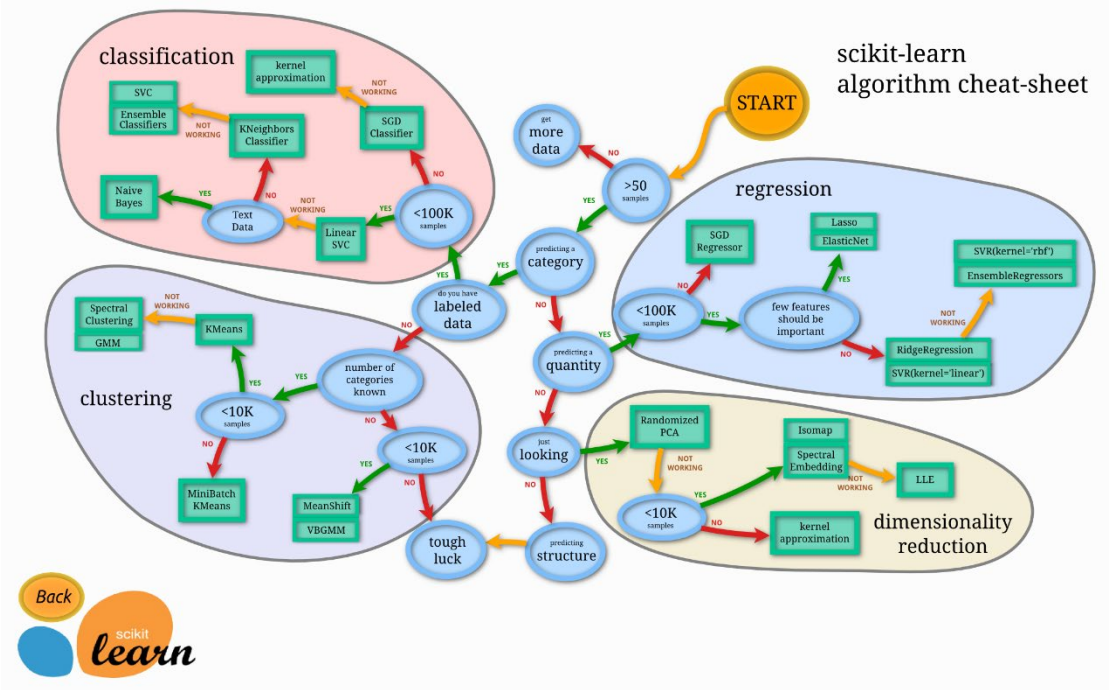


图 II -1 Scikit-learn 方法分类

在电力负荷预测与优化项目中，我们选择**回归模型**。原因如下：

- **任务需求**：我们需要预测和修正电力负荷和厂用率，这些都是连续的数值。
- **模型适配**：回归模型能够处理连续数据的预测任务，帮助我们提高预测精度，优化电力资源配置。
- **数据特性**：项目中涉及的预测数据（如电力负荷）是连续变量，适合使用回归模型进行建模。

（3）请概述样本集的划分方法，并说明哪种方法更适合当前的机器学习项目；

在发电出力预测修正项目中，如果发现厂用率和发电出力数据的时间间隔对预测结果影响不大，这表明数据的时间依赖性不强。因此，可以将数据视为非时间序列数据进行处理。使用随机划分方法将数据集分为训练集、验证集和测试集是合理的选择，通常的比例是 70%用于训练，15%用于验证，15%用于测试。随机划分能保留数据的整体分布特性，避免时间顺序对模型训练的影响，从而有效评估模型的性能和泛化能力。这种方法适合快速验证模型的效果，并在非时间序列特性明显的数据中表现良好。 对于厂用率预测也是如此。

两种情况下，都需要对每种发电类型分别进行预测，这是因为不同类型的发电方式（如火电、水电、风电等）可能具有不同的特性和影响因素。分别预测可以更准确地捕捉每种类型的特定模式和变化规律，从而提高预测的精度和可靠性，数据示例见图表。（如图

II-2 的厂用率监测数据和图 II-3 的发电出力值监测数据）。

	日期	时间	发电类型	发电出力值	负荷率	受阻率	厂用率
1	20231028	15:30	光伏	1063.2375	0.0	0.0	0.8523
2	20231028	15:30	火电燃煤	11417.4727	0.0	0.0	10.0141
3	20231028	15:30	水电有功	460.5616	0.0	0.0	1.0456
4	20231028	15:30	发电有功	15180.0322	0.0	0.0	0.0
5	20231028	15:30	风电	878.1815	0.0	0.0	6.0089
6	20231028	15:30	火电生物质	1195.7607	0.0	0.0	0.0
7	20231028	15:45	风电	908.1126	0.0	0.0	5.1185
8	20231028	15:45	发电有功	15237.2871	0.0	0.0	0.0
9	20231028	15:45	火电生物质	1207.2881	0.0	0.0	0.0
10	20231028	15:45	光伏	696.3683	0.0	0.0	1.2949

图 II -2 厂用率监测数据

日期	时间	预测时的日期	间隔天数	风电预测	风力发电	光伏预测	光伏发电	供电预测	供电负荷值
2023-10-13	10:00	2023-10-10	3	2070.38	2701.5127	2723.5	3405.2114		10305.153
2023-10-13	10:00	2023-10-10	3	2432.75	2701.5127	2704.07	3405.2114		10305.153
2023-10-13	10:00	2023-10-11	2	2070.38	2701.5127	2723.5	3405.2114		10305.153
2023-10-13	10:00	2023-10-12	1	2045.67	2701.5127	2711.25	3405.2114		10305.153
2023-10-13	10:15	2023-10-10	3	2428.14	2750.6099	2783.25	3529.8398		10462.762
2023-10-13	10:15	2023-10-10	3	2126.91	2750.6099	2802.32	3529.8398		10462.762
2023-10-13	10:15	2023-10-11	2	2126.91	2750.6099	2802.32	3529.8398		10462.762
2023-10-13	10:15	2023-10-12	1	2099.47	2750.6099	2789.76	3529.8398		10462.762
2023-10-13	10:30	2023-10-10	3	2437.77	2877.836	2811.84	3608.3574		10548.4795
2023-10-13	10:30	2023-10-10	3	2182.81	2877.836	2830.73	3608.3574		10548.4795
2023-10-13	10:30	2023-10-11	2	2182.81	2877.836	2830.73	3608.3574		10548.4795
2023-10-13	10:30	2023-10-12	1	2151.93	2877.836	2817.77	3608.3574		10548.4795
2023-10-13	10:45	2023-10-10	3	2240.1	3000.0344	2833.18	3675.2761		10722.039
2023-10-13	10:45	2023-10-10	3	2464.22	3000.0344	2813.85	3675.2761		10722.039

图 II -3 发电出力值监测数据

（4）请说明数据集中的所有特征能否全部参与机器学习过程，如果不可以，请阐述特征提取的方法；

在机器学习中，并非所有特征都适合参与模型训练。以下是一些原因：

- 冗余特征：一些特征可能提供重复的信息，导致模型过拟合。
- 噪声特征：某些特征可能包含噪声，影响模型的性能。
- 高维特征：特征过多可能导致“维度诅咒”，使得模型训练变得困难。
- 相关性：某些特征可能与目标变量的相关性较低，参与训练可能不会带来实际收益。

通过统计测试（Pearson 相关系数）评估每个特征与目标变量的相关性，我们发现厂用率时间特征对于结果相关性极低因此不考虑时间作为预测厂用率特征。

数据集中的特征无法全部参与机器学习过程，因此我们在特征操作中根据实际日期和预测日期计算了间隔天数（图 II -3）计算时间间隔可以作为一个特征，帮助模型理解时间上的变化，并排除无关特征。针对由于数据采集错误（如传感器故障或通信问题）导致的缺失值问题，我们可以检测明显错误（例如数据中的异常值如 0）并使用线性插值方法进行填补。这样能够有效提高模型的预测性能和稳定性。

（5）机器学习模型训练数据时可能会出现过拟合或欠拟合现象，请解释过拟合、欠拟合现象产生的原因，并说明解决方法；

过拟合

定义

过拟合是指模型在训练数据上表现良好，但在测试数据上表现不佳的现象。模型学习到了训练数据中的噪声和细节，而不是捕捉到数据的真实模式。

原因

1. 模型复杂度过高：使用了复杂的模型（如深度神经网络）来拟合简单的数据。
2. 训练数据量不足：训练数据量较小，模型容易记住训练样本。
3. 噪声数据：训练数据中存在较多噪声，使模型学习到错误的信息。
4. 特征过多：特征数量过多，模型可能会在特征空间中找到不必要的复杂关系。

解决方法

1. 正则化：使用 L1（Lasso）或 L2（Ridge）正则化来限制模型的复杂度。
2. 交叉验证：使用 k 折交叉验证评估模型性能，确保模型在不同数据集上的表现一致。
3. 简化模型：选择更简单的模型或减少模型的参数。
4. 增加训练数据：收集更多的训练数据，以帮助模型学习更普遍的模式。
5. 提前停止：在训练过程中监控验证集的性能，当验证集性能开始下降时停止训练。

欠拟合

定义

欠拟合是指模型在训练数据和测试数据上均表现不佳，未能捕捉到数据的基本趋势和模式。

原因

1. 模型复杂度过低：使用了过于简单的模型（如线性回归）来拟合复杂的数据。
2. 特征不足：未使用足够的特征来描述数据，导致模型无法学习到足够的信息。
3. 数据预处理不足：未对数据进行适当的预处理和特征工程，影响模型的学习能力。

解决方法

1. 增加模型复杂度：选择更复杂的模型（如多项式回归、决策树等）。
2. 增加特征：通过特征工程创建新的特征，或使用特征选择方法增加重要特征。
3. 调整模型参数：优化模型的超参数，以提高模型的拟合能力。
4. 数据预处理：对数据进行适当的预处理，如归一化、标准化等，以提升模型学习的效果。

2、针对你所构建的数据分析项目解决方案，对方案中的机器学习模型、模型改进、评估等关键问题进行分析和描述

（1）请使用两种不同的机器学习模型完成项目的最终目标，请给出相应模型的算法描述。

算法 1：梯度提升（Gradient Boosting）

算法描述：

梯度提升是一种集成学习方法，通过逐步构建多个弱学习器（通常是决策树），每个新模型针对前一个模型的残差进行训练。主要步骤包括：

- 初始化模型，通常使用均值。
- 计算当前模型的残差（预测与实际值之间的差异）。
- 训练新的弱学习器来拟合这些残差。
- 更新模型，将新模型的预测结果加入当前模型。
- 重复上述步骤，直到达到预设的迭代次数或性能不再提升。

应用于项目：

在发电出力值和厂用率预测中，梯度提升能有效捕捉复杂的非线性关系，提高预测精度。

算法 2：线性回归（Linear Regression）

算法描述：

线性回归通过线性方程拟合自变量与因变量之间的关系，其形式为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

应用于项目：

线性回归适合用于预测发电出力与厂用率之间的线性关系。如图 II -4，使用线性回归模型拟合真实值曲线。

图 II -4 线性回归模型拟合真实值曲线，在原预测值基础上的修正

（2）请评价所使用的机器学习模型，并给出该模型的改进方法。

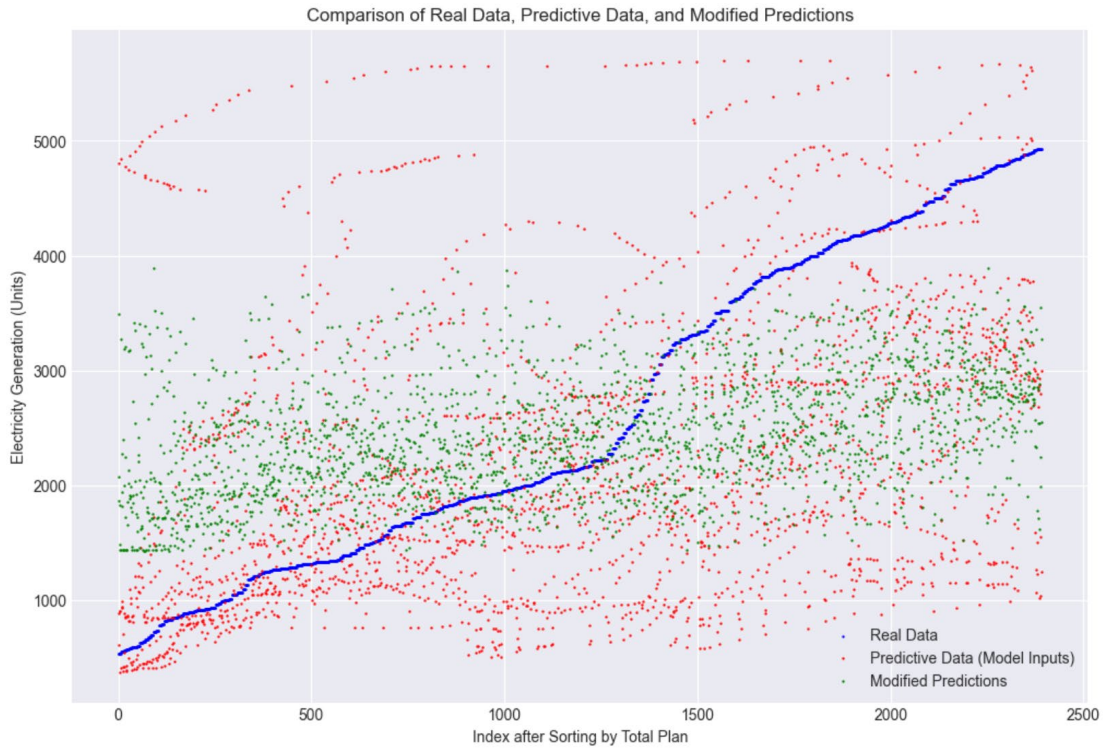
算法 1：梯度提升（Gradient Boosting）

1）优点：

- 能处理复杂的非线性关系。
- 高预测精度。

2）缺点：

- 训练时间较长。
- 容易过拟合。
- 实测效果提升有限。




```
修正前: 0.11336202941690372, RMSE: 0.33669278194951513, MAE: 0.25397675040441214
修正后: 0.09378999253547249, RMSE: 0.30625151842149695, MAE: 0.24222200221854268
```

图 II -5 梯度提升模型拟合真实值曲线

3）改进方法：

- 使用交叉验证选择最佳参数。
- 采用早停法监控验证集性能。

算法 2：线性回归（Linear Regression）

1）优点：

- 模型简单，易于理解。
- 训练速度快。

2）缺点：

- 只能捕捉线性关系。
- 对异常值敏感。

3）改进方法：

- 进行特征工程，添加多项式特征。
- 使用正则化防止过拟合。

➢ 采用异方差回归，同时预测厂用率的均值和方差，以更好地捕捉数据的分布特征。

如图 II -6，厂用率数据比较符合正态分布，因此采用异方差回归的方法是非常合适的。在这种方法中，您不仅预测目标变量（厂用率）的均值，还同时预测其方差。这种做法能够更准确地反映数据的不确定性，尤其是在目标变量的方差随输入特征变化时。

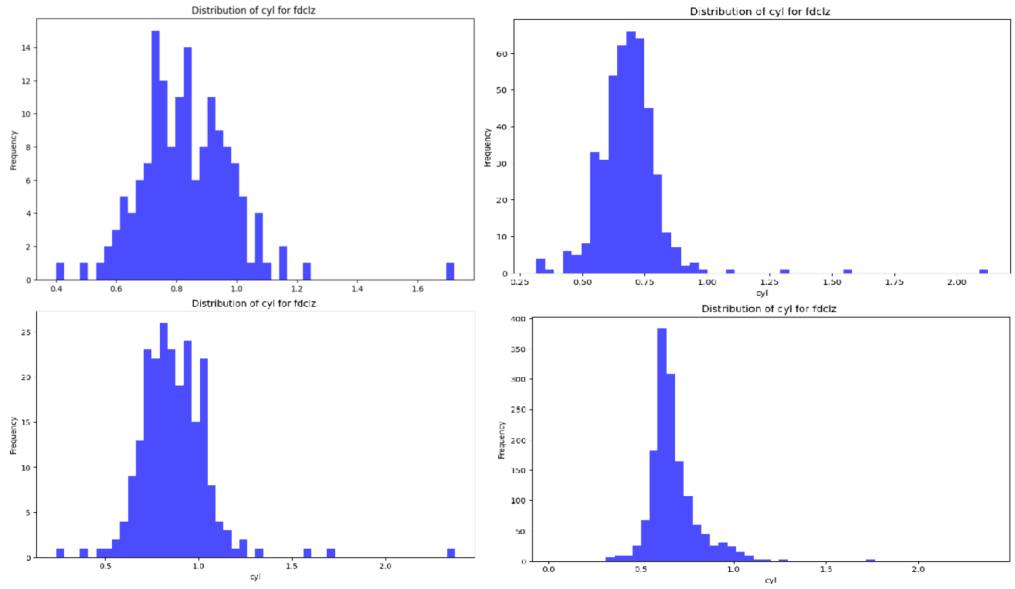
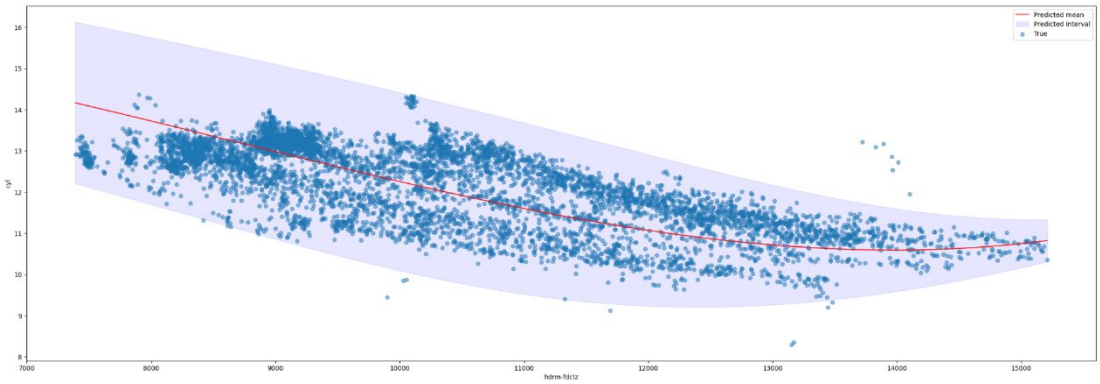


图 II -6 不同出力值区间厂用率分布




```
厂用率预测覆盖 hdrm: 99.75%

输入x= [[13109.719]]
预测均值mean= [[10.690131]]
预测方差std: [[9.468965]]

预测区间: [[9.283293]] ~ [[12.09697]]

预测均值的多项式:  0.9674301147460938 * x^0 + -0.9090405702590942 * x^1 + -0.33470386266708374 * x^2 + 0.6924334764480591 * x^3
预测方差的多项式:  exp(-1.3111510276794434 * x^0 + 0.3933233618736267 * x^1 + 0.4173588752746582 * x^2 + -2.156473159790039 * x^3)
```

图 II -7 采用采用异方差回归训练常用率预测模型

（3）请说明如何对使用到的机器学习模型的执行效果进行评估。

1. **训练误差**

定义：训练误差是模型在训练数据集上的表现，通常通过损失函数（如均方误差、交叉熵等）来衡量。

计算方法：在训练过程中，模型通过最小化损失函数来优化参数，从而计算训练误差。这可以通过以下公式表示：

$$\text{训练误差} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y_i})$$

其中，$\boldsymbol{L}$ 损失函数，$\boldsymbol{yi}$ 是实际值，$\hat{y_i}$ 是预测值，n 是样本数量。

2. 测试误差

定义：测试误差是模型在独立测试数据集上的表现，反映了模型的预测能力。

计算方法：与训练误差相似，测试误差通过在测试数据上计算损失函数得出：

$$\text{测试误差} = \frac{1}{m}\sum_{j=1}^{m} L(y_j, \hat{y_j})$$

其中，m 是测试样本数量。

表 II -1 出力值预测模型测试误差示例，最优值**加粗**

测试误差	MSE	RMSE	MAE
预测数据	0.306	0.553	0.547
预测修正	**0.007 (-0.299)**	**0.082 (-0.471)**	**0.063 (-0.483)**

3. 泛化误差

定义：泛化误差是指模型在所有可能的未见数据上的平均预测误差，通常难以直接计算。

计算方法：泛化误差可以通过交叉验证等方法间接评估。在交叉验证中，数据集被划分为多个子集，模型在不同的训练和测试组合上进行训练和评估，最终得出一个平均误差。