Анализ набора данных Iris и демонстрация переобучения
鸢尾花数据集分析与过拟合演示

Ёркинжон Валиев - ИУ1-41М

1 Введение

Целью данного проекта является анализ набора данных Iris, выполнение бинарной классификации с использованием логистической регрессии и демонстрация переобучения на примере линейной регрессии. Машинное обучение подвержено переобучению, когда модель запоминает обучающие данные вместо выявления общих закономерностей. В отчёте представлен подробный анализ, реализация моделей и обсуждение результатов.

2 Обзор набора данных

Набор данных Iris содержит 150 образцов цветов, классифицированных на три вида: Setosa (0), Versicolor (1), Virginica (2). Каждый экземпляр описан четырьмя признаками: длина и ширина чашелистика и лепестка. Для классификации выбраны Setosa и Versicolor, класс Virginica исключён.

引言

本项目旨在分析鸢尾花数据集，使用逻辑回归进行二分类，并展示线性回归中出现的过拟合问题。机器学习中，模型可能只记忆训练数据而非发现通用规律。本文展示了数据分析、模型实现及其评估结果。

数据集概述

鸢尾花数据集包括150个样本，分为Setosa（0）、Versicolor（1）和Virginica（2）三类。每个样本由花萼和花瓣的长度与宽度4个特征描述。本实验选择Setosa和Versicolor两类，排除Virginica类。

3 Визуализация данных

Для исследования взаимосвязей признаков был построен pairplot, отображающий распределения и зависимости между признаками.

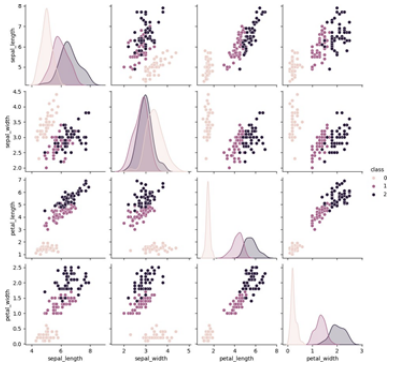

Рис. 1: Pairplot признаков / 特征可视化

Класс Setosa чётко отделяется от других, Versicolor и Virginica — частично пересекаются.

4 Логистическая регрессия

4.1 Математическая модель

Логистическая регрессия использует сигмоидную функцию:

$$P(y = 1|X) = \frac{1}{1 + e^{-(W^T X + b)}}$$

4.2 Реализация

- Удалён класс Virginica.
- Разделение: 80% — обучение, 20% — тест.
- Стандартизация:

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

数据可视化

为研究特征间关系，绘制了pairplot图，展示分布和相关性。

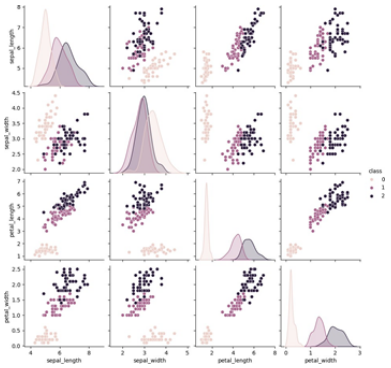

Рис. 1: 特征分布图 / Pairplot

Setosa类别明显分离，而Versicolor与Virginica有重叠。

逻辑回归

数学模型

逻辑回归采用S型函数:

$$P(y = 1|X) = \frac{1}{1 + e^{-(W^T X + b)}}$$

实现步骤

- 移除Virginica类;
- 训练集80%，测试集20%;
- 标准化处理:

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

- 模型训练后准确率达到100%。

- Обучена модель, достигнута точность 100%.

5 Переобучение в линейной регрессии

5.1 Модель

Линейная регрессия:

$$y = W^T X + b + \epsilon$$

- Синие точки — обучающая выборка;
- Красные — тестовая;
- Зелёная линия — обученная модель;
- Фиолетовая — предсказания.

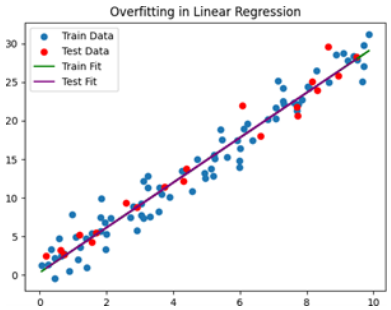

Рис. 2: Переобучение / 过拟合演示

6 Заключение

- Визуализация помогает выявить разделимость;
- Стандартизация — ключ к качеству модели;
- Переобучение можно устранить регуляризацией (L1, L2).

线性回归中的过拟合

模型描述

线性回归:

$$y = W^T X + b + \epsilon$$

- 蓝点为训练数据;
- 红点为测试数据;
- 绿色为拟合曲线;
- 紫色为预测值。

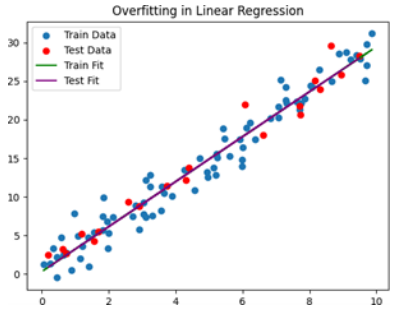

Рис. 2: 过拟合图示 / Overfitting

结论

- 可视化有助于识别类别分离;
- 标准化是提高模型性能的关键;
- 可通过正则化（L1, L2）减轻过拟合。

Список литературы / 参考文献

Список литературы

[1] R.A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 1936.

[2] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.

[3] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.