4. 机器学习基础

WU Xiaokun 吴晓堃

xkun.wu [at] gmail

2021/04/19

本章内容

机器学习的基本分支。 机器学习模型的评估方法。 数据预处理和特征工程。 解决过拟合问题的三种方法。 机器学习的通用工作流程。 实践：数据点拟合，过拟合处理。

重点：机器学习模型的评估方法、数据预处理的基本方法、机器学习的通用工作流程；

难点：超参数的调节流程、解决过拟合问题的三种方法。

学习目标

- 了解机器学习的四大基本分支：监督学习、无监督学习、自监督学习和强化学习；
- 掌握机器学习模型的评估方法和超参数的调节流程；
- 掌握数据预处理的基本方法，并理解特征工程的意义；
- 掌握解决过拟合问题的三种方法：减小网络容量，权重正则化，Dropout正则化；
- 掌握机器学习的通用工作流程。

机器学习的基本分支

监督学习

样本和目标都是给出的。

无监督学习

只有样本，没有目标。

强化学习

强调如何基于环境而行动，以取得最大化的预期利益。

机器学习效能的根本矛盾

优化

优化由理论分析及数值计算来解决；

泛化

- 机器学习的最终目的是得到可以泛化的模型；
- 衡量模型泛化能力的核心是预测其在未知数据上的效能。

评估机器学习模型

数据划分类别

- 训练集：训练模型；
- 验证集：评估模型，调节模型配置（超参数）；
- 测试集：模型的最终评估。

为什么评估需要两个数据集？

评估过程必然造成信息泄露：未知数据被间接地获取。

数据划分方法

- 简单留出；
- K-fold 交叉验证；
- 重复的 K-fold 交叉验证。

评估模型的注意事项

数据代表性

例如，数字图像分类问题，在将数据划分为训练集和验证集之前，通常应该随机打乱数据。

时间箭头

例如，气温预测问题，始终确保验证集中所有数据的时间都晚于训练集数据。

数据冗余

确保训练集和验证集之间没有交集。

数据预处理与特征工程

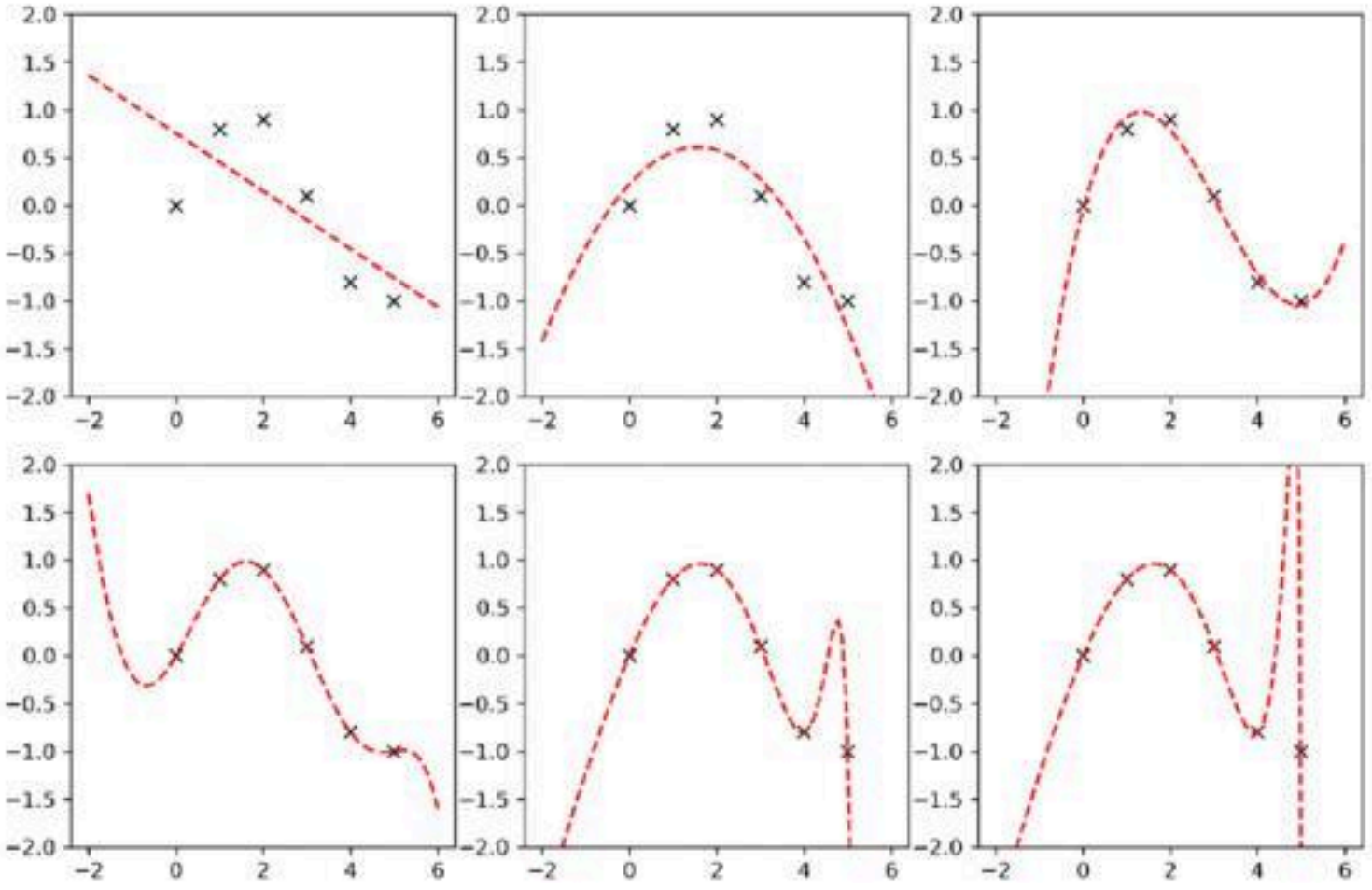
数据预处理

- 向量化；
- 值标准化：标准正态分布；
- 处理缺失值。

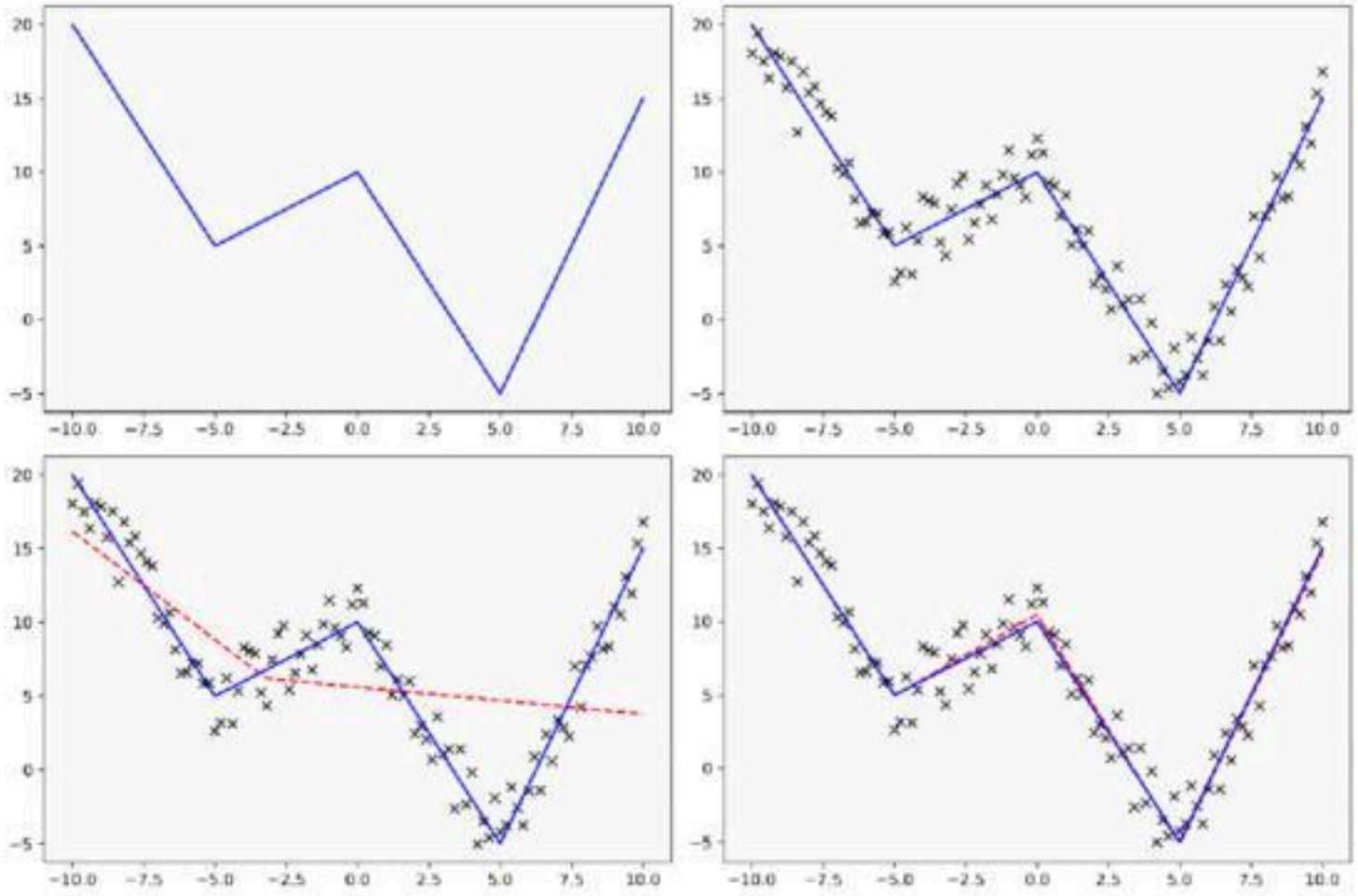
特征工程

本质上是应用先验知识将数据变换成易于训练的表示形式，即人与机器的合作。

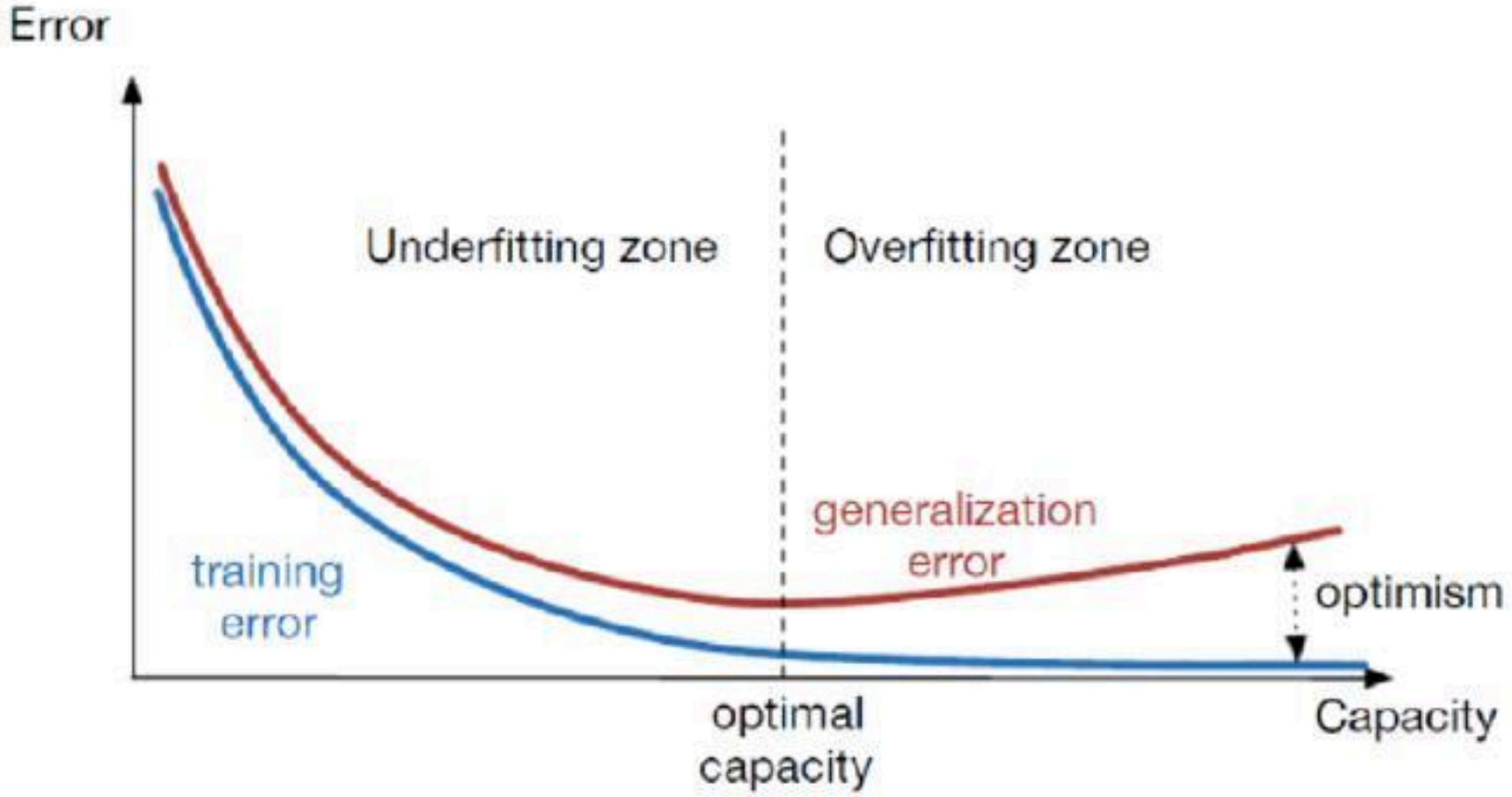
数据拟合：过拟合与欠拟合



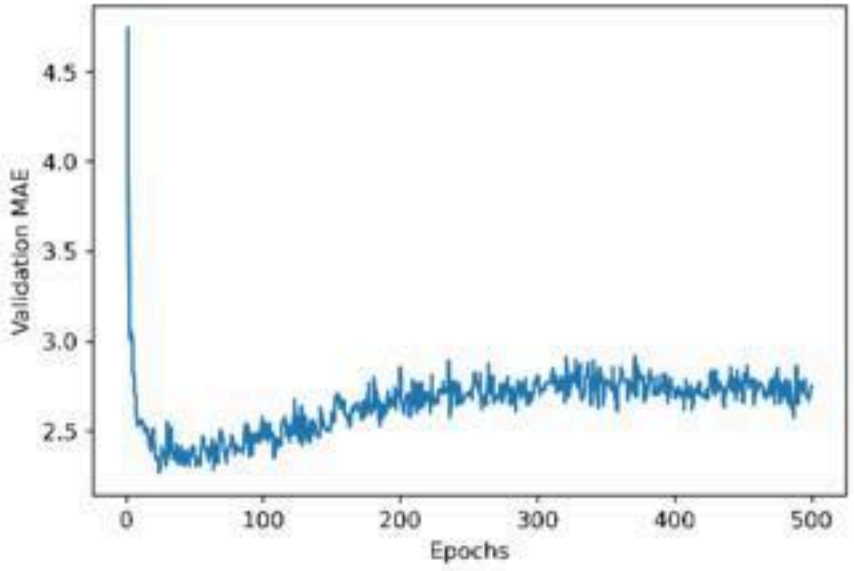
数据拟合：先验知识

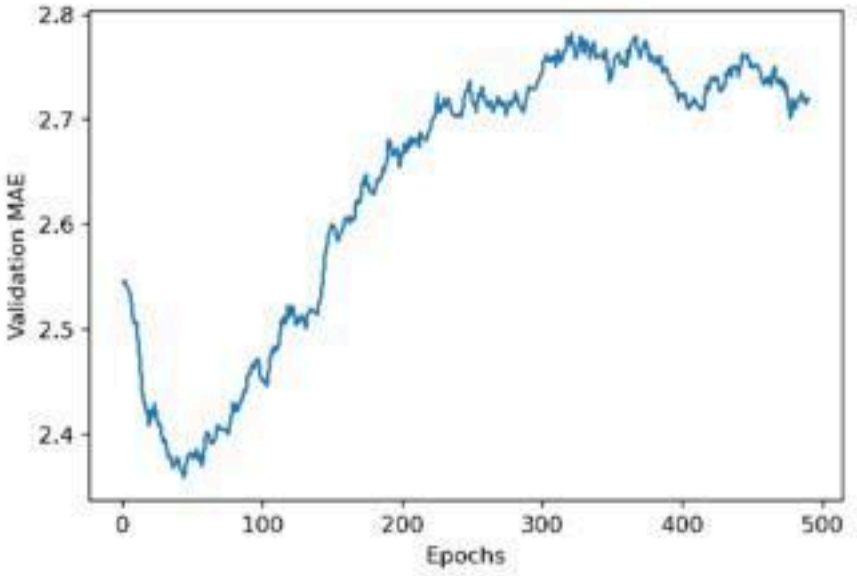


过拟合与欠拟合：理想曲线



过拟合与欠拟合：实际曲线





防止过拟合的基本思路

获取更多的训练数据

最有效、最简单、最耗资源。

正则化

通过降低模型复杂度来防止过拟合的方法。

- 次优解决方法。即调节模型允许存储的信息量，或对模型允许存储的信息加以约束。
- 如果一个网络只能记住非常有限的几个模式，那么优化过程会迫使模型集中学习最重要的模式，这样更可能得到良好的泛化。

防止过拟合的其他常用技巧

减小网络容量

小容量的模型被迫只能记住最关键的几个模式。

添加权重正则化

奥卡姆剃刀原理：最可能正确的解释是最简单、假设最少的那个。

- 简单模型：指参数值分布的熵尽可能小（L2），或参数尽可能少（L1）。

添加 dropout

随机将 `dropout` 层的一些输出特征舍弃（设置为 0）。

- 核心思想是在层的输出值中引入噪声，从而避免模型学到偶然模式。

机器学习的通用工作流程

- 定义问题，收集数据集；
- 选择评价模型效能的终极指标：一般与领域相关；
- 确定调节模型超参数的验证方法：注意信息泄露问题；
- 准备数据：向量化、值标准化、处理缺失值；
- 开发比基于常识的基准方法更好的模型：确保问题可以解决；
- 扩大模型规模：开发略微过拟合的模型；
- 模型正则化与调节超参数：从效能的两级趋向最优。

分类和回归术语表 I

样本（sample）或输入（input）

进入模型的数据点。

目标（target）

真实值。

预测（prediction）或输出（output）

从模型出来的结果。

预测误差（prediction error）或损失值（loss value）

模型预测与目标之间的距离。

分类和回归术语表 II

类别（class）

进入模型的数据点。

标签（label）

分类问题中供选择的一组标签。

真值（ground-truth）或标注（annotation）

数据集的所有目标，通常由人工收集。

分类和回归术语表 III

二分类（binary classification）

一种分类任务，每个输入样本都应被划分到两个互斥的类别中。

多分类（multiclass classification）

一种分类任务，每个输入样本都应被划分到两个以上的类别中。

多标签分类（multilabel classification）

一种分类任务，每个输入样本都可以分配多个标签。

分类和回归术语表 IV

标量回归（scalar regression）

目标是连续标量值的任务。

向量回归（vector regression）

目标是一组连续值（比如一个连续向量）的任务。如果对多个值（比如图像边界框的坐标）进行回归，那就是向量回归。

小批量（mini-batch）或批量（batch）

模型同时处理的一小部分样本（样本数通常为 8~128）。样本数通常取 2 的幂，这样便于 GPU 上的内存分配。训练时，小批量用来为模型权重计算一次梯度下降更新。