COMP 551 Lecture 1 - Introduction of machine learning (M1.1)

Junji Duan

2024/1/8

Supervised learning

监督学习是机器学习中的一种方法，它使用包含输入特征和相应正确输出（标签）的标记数据集来训练模型。目标是学习一个函数 f，这个函数可以将输入（特征）映射到输出（标签），使模型能够预测新的、未见过的输入数据的输出。The task is to learn a mapping function f from inputs $\mathbf{x} \in \mathcal{X}$ to outputs $y \in \mathcal{Y}$ (i.e., $f(\mathbf{x}) \to y$).

Inputs

- 输入 (inputs) x 是用来预测输出的数据点。在机器学习中, 这些通常被称为特征 (features)、协变量 (covariates)、预测因子 (predictors) 或独立变量 (independent variables)。
- 输入通常表示为固定维度空间 $\mathcal{X} = \mathbb{R}^D$ 中的向量, 其中 D 代表输入向量的维度, 即每个输入数据点拥有的特征或属性的数量。

Outputs

- 输出 y 是训练集中每个输入向量对应的期望结果。在监督学习的背景下, 这些输出在训练数据中提供, 并用于训练模型。
- 输出 y 也被称为 label, target, or response, 它们的类型可以不同（例如, 分类任务中的类别型输出, 回归任务中的连续型输出）。

Training Set

- 训练集 ($\mathcal{D}$) 包含 N 个输入-输出对 $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$, 其中 N 是数据集中的样本数量 (Sample Size)。
- 这个集合用来训练模型, 即模型根据这些示例学习输入与输出之间的映射关系。

Evaluation Metrics

- 评估指标用来衡量学习函数 $f(\mathbf{x})$ 在预测输出方面的表现如何。指标的选择取决于输出的类型:
  - 对于分类任务 [Classification] (输出为类别型), 常见的指标包括准确率、精确率、召回率和 F1 分数。
  - 对于回归任务 [Regression] (输出为连续型)，常用的指标有均方误差 (MSE) 、均方根误差 (RMSE) 和平均绝对误差 (MAE)。

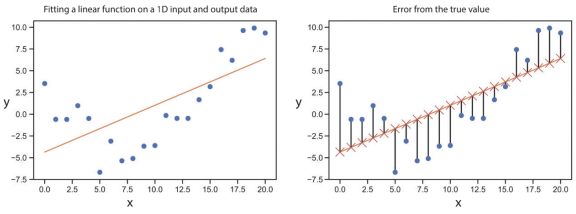
通过在一个已知正确输出的数据集上训练, 模型可以学习输入与输出之间的关系, 这使得它能够对新数据进行预测。这些预测的有效性通过适当的评估指标来衡量。

Classification

在分类任务中，输出空间 Y 由一组无序且互斥的标签组成，这些标签被称为类别。标签通常表示为 $Y = \{1, 2, \ldots, C\}$ 。分类的目标是基于学习到的映射函数 f, 预测输入 $\mathbf{x}$ 属于哪个类别 y 。常见的例子包括电子邮件垃圾邮件检测（垃圾邮件或非垃圾邮件）、图像识别（识别图像中的对象）和医学诊断（确定疾病的存在与否）。

Regression

回归与分类不同, 它涉及预测一个连续的输出而不是类别标签。在回归中, 输出 y 是一个实数值量, 目标是找到一个从输入 x 到实数值输出 y 的映射函数。这通常被视为拟合一条线或曲线，最好地描述数据中输入和输出之间的关系。一个例子是根据各种特征（如房屋的大小、位置和年龄）预测房屋价格。



Overfitting and Generalization

过拟合 [Overfitting] 是分类和回归中常见的问题, 其中模型学习了训练数据中的细节和噪声, 以至于它对新数据的表现产生了负面影响（即模型过度适应了训练数据, 而未能很好地泛化到未见过的数据）。这通常发生在模型过于复杂的情况下，如回归中的高阶多项式函数。

过拟合的例子:

- 多项式回归: 如果我们拟合一个 K-阶多项式函数 $f(x) = 1 + b_1 x + \ldots + b_K x^K$ ，我们可能

会发现随着 K 的增加，多项式开始越来越好地拟合训练数据。然而, 高阶多项式可能开始捕捉噪声而不是底层关系, 从而导致对新的未见数据表现不佳。

- 面板示例：在一个示例中，低阶多项式可能欠拟合（未捕获数据中的所有趋势），而非常高阶的多项式可能过拟合（拟合数据中的噪声而不是实际信号）。

决策树与过拟合:

- 决策树是一种容易过拟合的典型模型，特别是如果允许它们无限制地生长变深。一个深度决策树可能通过创建高度特定的规则来完美分类所有训练数据点, 但这些规则一般不适用于新数据, 因此在新数据上失败。

泛化 [Generalization] 指的是模型在新的、未见过的数据上的表现，不仅仅是在其训练数据上的表现。机器学习的终极目标是开发出能够很好泛化的模型。使用交叉验证、正则化和选择更简单的模型等技术可以帮助防止过拟合并提高泛化能力。

Unsupervised learning

- 在非监督学习中，数据集 $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N$ 仅包含输入数据 x_n 而没有对应的输出 y_n。这意味着模型需要从这些数据中自行发现结构或模式。
- 探索潜在模式：非监督学习的目的是探索输入数据中的潜在模式和结构，而不是预测特定的输出。这包括识别数据中的群组、常见模式、数据分布的特征等。
- 处理高维输入：非监督学习的一个挑战是处理高维数据，寻找方式来 “解释” 或 “理解” 这些数据的内在结构，而不仅仅是输出简单的结果
- 应用场景：非监督学习特别适用于那些标签信息稀缺或完全缺失的情况。由于大多数实际应用中的数据未被标记，非监督学习在处理真实世界数据中扮演了重要角色。
- 非监督学习的方法
  - 聚类 [Clustering:]：这是一种常见的非监督学习技术，目的是将数据集中的样本分组，使得同一组内的样本相似度较高，而不同组间的样本相似度较低。
  - 降维 [Dimensionality reduction]：降维技术如主成分分析（PCA）和 t-SNE 帮助简化数据，去除噪声和冗余信息，同时保留最重要的结构特征，以便更容易地进行分析和可视化。
  - 异常检测：在数据中自动识别异常或不寻常的模式，这在欺诈检测和网络安全等领域非常重要。

总的来说，非监督学习为处理和分析未标记的大数据集提供了强大的工具，使我们能够洞察数据的内在结构和动态。