一些想法

2019 年 11 月 11 日

目录

1 数据和模型

{aaaa}

机器学习中，数据可以被分为四大类：

- 图像：Image
- 序列：Sequence
- 图：Graph
- 表格：Tabular

其中，前三类有着比较明显的模式。比如**图像和图的空间局部性**，**序列的上下文关系和时序依赖性**。而表格数据常见于**各种工业界的任务**，比如广告点击率预测，推荐系统等。在表格数据中，每一个特征表示一个属性，如性别、价格，特征之间一般没有明显且通用的模式。

神经网络适合前三类数据，也就是有明显模式的数据。**针对不同的数据模式，设计对应的网络结构**，从而实现高效地自动抽取"高级"的特征表达。如 CNN（图像）、RNN（序列）。**而表格数据，没有明显的模式**。神经网络无法针对设计。因此对于表格数据，除了专门对特定的任务设计的网络结构如 DeepFM 等，更多时候还是用传统的机器学习模型。尤其是 LGBT（梯度提升术）。因其自动的特征选择能力及动态的模型复杂度，算得上是一个万金油模型，在各种类型的表格数据上都表现很好。

对于表格数据而言，特征工程更加关键，在给定数据的情况下，模型决定了下限，特征决定上限。**特征工程类似于神经网络的结构设计，目的是把先验知识融入数据**。

No free lunch，没有万能的模型。用神经网络，需要结构设计；使用传统模型，需要特征工程。

2 基本的机器学习方法

机器学习的基础方法大概有六种：K 近邻算法、主成分分析法、逻辑回归算法、朴素贝叶斯分类器、决策树算法、随机森林、支持向量机算法、K-Means 聚类、人工神经网络 ANN。**每种方法都有其适应的场景、对象，以及其内涵**

3 机器学习的分类

机器学习分为有监督和无监督两种。

对于有监督学习，是给定数据集 $\{x^i, y^i\}$，学习出 $\hat{y} = f(x)$。**对于分类问题**，y 是**类别**。**对于回归问题**，y 是**连续数**；**对于排序问题**（尤其是信息检索和网页排序等应用上），y 是**序值**。一般情况下，有监督学习就分为这几种。每一个有监督学习都可以归结到这几类问题中，并对照数据和模型1选择合适的方法。

无监督学习是给定数据集 $\{x^i\}$，学习出 $\hat{y} = f(x)$。**对于密度估计**，y 是**密度**。**对于聚类**，y 是**类簇**。**对于数据规约、数据可视化**，y 是**数据 x 的低纬表示**（例如 Autoencoder）。无监督想是发展的方向，当前无监督学习经常被作为监督学习的预处理步骤（这是当前监督学习的有种流行的范式）。