人工智能-知识梳理与面试真题

监督学习中的数据、特征处理与模型评价

目录 (Table of Contents)

缺失值处理

缺失情况检查：为什么缺失? 缺失状态？

1. 完全随机缺失 (MCAR)：变量 Y 有缺失数据。如果 Y 缺失数据的概率与 Y 本身的值或在该数据组中任何其他变量的值都无关的话，那么 Y 的数据就是完全随机缺失的。
2. 随机缺失 (MAR)：控制了其他变量后，Y 缺失数据的概率与 Y 值无关，则称 Y 的数据为随机缺失。
3. 非随机缺失，依赖于缺失数据本身：Y 缺失数据的概率与 Y 值相关。
4. 非随机缺失，依赖于未观测到的预测值：Y 缺失数据不是随机的，依赖于某一未记录的信息，且该信息可预测缺失值。

处理缺失数据的方法：

1. 完全删除：删除含有缺失变量的数据。
    - 优点：可用于任何类型的统计分；不需要特别的运算方法；如果任何因变量缺失数据的概率不取决于自变量的值，则使用成列删除的回归估计值将会是无偏误的。
    - 缺点：误差通常较大；如果数据不是 MCAR 而只是 MAR，那么删除可能会产生有偏误的估计值。

2. 简单插补：以某些合理的猜测插补来替代缺失值，然后再接着按没有缺失数据的情况进行分析。插补的值，一般是零值，极大或极小值，平均值，条件均值，中位数。
    - 优点：保证数据的完整性，模型的适应性
    - 缺点：低估标准误、高估检验统计量
3. 多重插补：以两个或多个插补来替代缺失值，分析不同的插补值带来的数据结构的变化及误差，来选择最优的插补值。
    - 优点：当数据为 MAR 时，正确使用多重插补会产生一致的、渐近有效的估计值；可以被任何一种模型使用。
    - 缺点：操作繁琐易出错；每次使用多重插补时，都会产生不同的估计值。
4. 最大似然：利用最大期望算法来估测缺失值。
    - 优点：适用于大样本。
    - 缺点：只适用于线性模型。

不均衡数据处理

不均衡数据：因变量在不同类别中的分布不均衡。

不均衡数据的影响：影响模型的准确性。

- 分类模型的表现会偏向于数据多的类别
- 在整体模型最小化误差的过程中，数据少的类别贡献非常的少
- 一些模型应用的前提条件是数据均衡分布
- 一些模型应用的前提条件是误差在不同类别中的权重是相同的

处理不均衡数据的方法：

1. 欠采样：从数据多的类别中随机抽取一部分
    - 优点：适用于数据量较大的情况；可以提高计算效率，减少计算时间和内存要求。
    - 缺点：数据原有模式被破坏，模型准确率变差
2. 过采样：从数据少的类别中随机选取数据复制
    - 优点：无信息缺失
    - 缺点：容易导致过拟合
3. SMOTE (Synthetic Minority Over-sampling Technique): 模拟数据少的类别中的数据形式，并生成数据填入其中。生成方法是：选取数据少的类别中的一点，找出它的最近点，算出差值，随机生成一个 0 和 1 之间的随机数，差值乘以随机数。
    - 优点：无信息缺失；保持了模型原有模式；不易过拟合

特征工程

数据的预处理方法

1. 数据标准化：

$$x' = \frac{x - \mu}{\sigma^2}$$

2. 数据缩放：

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

3. 数据归一化：

$$x' = \frac{x}{\|x\|_2}$$

4. 数据二分化：

$$x' = 1, \ x > \eta; \quad x' = 0, \ x \leq \eta$$

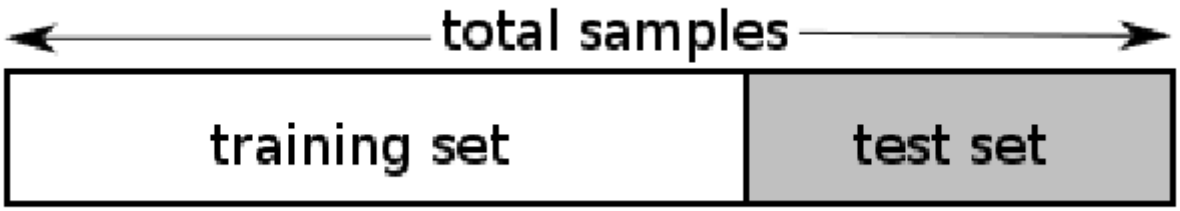
5. 独热编码处理分类型数据

特征选取

1. 方差原则：去掉方差较低的特征。
2. 相关性原则：去掉完全相关的变量以防止共线性；选取得分较高的特征。
3. 递归式特征消除：使用一个基模型来进行多轮训练，每轮训练后，消除若干权值系数的特征，再基于新的特征集进行下一轮训练。

模型验证与评价

交叉验证



- 缺点：
  i. 方差较高
  ii. 在稀疏数据集上表现差
  iii. 测试集数据越多，偏差越高

K-fold 交叉验证

- 缺点：
  i. 在不均衡数据集上表现不好

分层交叉验证

模型评价

误差矩阵	预测正值	预测负值
真实正值	TP	FN
真实负值	FP	TN

- 准确率：

$$\frac{TP+TN}{TP+FN+FP+TN}$$

- 误差率：

$$\frac{FP+FN}{TP+FN+FP+TN}$$

- 精确率:

$$\frac{TP}{TP+FP}$$

- 召回率:

$$\frac{TP}{TP+FN}$$

- ROC 曲线 和 AUC 值：

面试真题

1. 建模的时候，遇到数据缺失怎么办？

2. 如果数据不均衡，可以直接建模吗？ 为什么？

3. 介绍一下 SMOTE 这个方法？他是用来做什么的？

4. 举例三种数据预处理的方法？

5. 为什么要进行数据预处理？

6. 举出三种特征工程常用的方法？

7. 准确率和精确率的区别？

8. 什么是 ROC 曲线？ 和 PR 曲线的区别是什么？

9. 什么是 AUC?