特征工程

1. 为什么要对特征归一化?

- 提升模型收敛速度：归一化让各特征在数值上处于同一尺度，避免某些特征值过大导致梯度下降缓慢。
- 防止特征主导模型：不同量纲或分布的特征可能使模型偏向于数值大的特征，归一化后权重更均衡。
- 提升模型表现：许多算法（如K近邻、SVM、神经网络等）对特征尺度敏感，归一化有助于提高模型性能。 常见方法有min-max归一化和z-score标准化。

2. 什么是组合特征？如何处理高维组合特征?

- **组合特征**：通过对原始特征进行某种运算（如拼接、乘法、加法等）得到的新特征。例如，如果有性别和年龄两个特征，可以构造“性别+年龄段”这样的组合特征。
- **高维组合特征**（尤其是类别型特征组合后）会导致特征空间爆炸（维度极高，很多组合很少甚至没有样本）。
- 主要的处理方法有：
  - **降维处理**：用特征选择（如卡方、信息增益）筛选有效组合。
  - **嵌入编码**：把高维组合特征用Embedding等方法映射成低维连续向量，常用于深度学习。
  - **限制组合方式**：只做有限几类（如二元交叉）或统计显著性强的组合，减少无意义的高维特征。

3. 比较欧氏距离和曼哈顿距离

- **欧氏距离（L2距离）**
    - 计算方式：两点间的“直线”距离，公式为 $\sqrt{\sum_{i=1}^n (x_i - y_i)^2}$
    - 几何意义：二维下为两点之间的线段长度，高维空间也类似
    - 对特征的影响：受大数值（离群点、单维远离）影响更加敏感
    - 适用场景：空间距离度量、特征方差接近时效果好
- **曼哈顿距离（L1距离）**
    - 计算方式：各坐标轴距离之和，公式为 $\sum_{i=1}^n |x_i - y_i|$
    - 几何意义：类似“城市街区”步行路线（只能水平/竖直走）
    - 对特征的影响：对离群点不那么敏感，更鲁棒
    - 适用场景：高维稀疏数据、特征差异较大时更稳定

4. 为什么一些场景使用余弦相似度而非欧氏距离？

- 余弦相似度常用于衡量“方向”是否一致，而不是“大小”是否接近。
- 使用场景通常是：
    - 关注角度而非幅值：如文本相似度、推荐系统、词向量表示中，向量长度受文本长短或频率影响，但我们更关心内容或语义方向是否一致。
    - 忽略模长差异：余弦相似度对向量的模长（大小）不敏感，只比较方向，能有效过滤掉由数量级造成的误差——如用户兴趣相似但活跃度不同的情况。
    - 高维稀疏数据：欧氏距离在高维空间往往退化为无意义的度量（距离趋于平均），而余弦相似度在高维稀疏向量（如TF-IDF、one-hot文本特征）中仍表现良好。
    - 归一化后差异小：若数据已归一化，欧氏距离和余弦相似度在数值上等价，但在未归一化时，余弦更适用。

5. one-hot作用是什么？为什么不直接使用数字？

- one-hot编码的作用是将类别型特征转化为模型可以处理的数值特征，且避免引入错误的“大小关系”。
- 去除类别间的顺序/距离影响：每个类别用1和0的独立分量表示，保证不同类别之间“等距”且“独立”。

模型评估

1. 模型评估过程中，过拟合和欠拟合分别指什么现象？

- **过拟合（Overfitting）**
    - **定义**：模型在训练集上表现很好，但在验证集/测试集上表现差。
    - **现象**：模型过度学习了训练数据的细节和噪声，无法很好地泛化到新数据。
- **欠拟合（Underfitting）**
    - **定义**：模型在训练集和测试集上都表现不好。
    - **现象**：模型对训练数据规律学习不够，无法捕捉到数据的真实关系。

2. 降低过拟合和欠拟合的方法

- **降低过拟合的方法：**
    - 增加训练数据量，提升泛化能力；
    - 数据增强（如图像平移、翻转等）；

  - 使用正则化（如L1/L2正则、Dropout、早停early stopping）；
  - 降低模型复杂度（减少参数、简化结构）；
- **降低欠拟合的方法：**
  - 增加模型复杂度（引入更多参数、深层网络）；
  - 提高训练轮数或增加学习时间；
  - 减少正则化力度（减小正则化系数）；
  - 特征工程（增加有效特征或非线性特征）。

3. L1和L2正则先验服从什么分布

- L1正则化（Lasso）
  - 对应参数服从拉普拉斯分布（Laplace分布，零均值的双指数分布）
  - L1正则倾向产生稀疏解（多数参数为零）
  - $p(w) \propto \exp(-\lambda|w|)$
- L2正则化（Ridge）
  - 对应参数服从高斯分布（Gaussian分布，零均值的正态分布）
  - 参数趋向零但一般不完全为零。
  - $p(w) \propto \exp(-\lambda w^2)$

4. 对于树形结构为什么不需要归一化?

- **不依赖距离或梯度计算**：不像线性模型或神经网络等，树模型不基于距离度量或梯度更新，特征尺度不会导致训练过程失衡。
- **模型的分割过程对不同量纲天然鲁棒**：分类或回归树分裂时查找最优节点，只关心某特征上的分割阈值，无需考虑特征间比例关系。

5. 什么是数据不平衡？如何解决?

- 数据不平衡指的是分类任务中样本分布极为不均，某（些）类别样本远多于其他类别。例如：正样本10%、负样本90%。
- 容易导致模型“偏见”，优先预测数量多的类别，导致少数类别的召回率或准确率较低。
- 常见解决方法
  - 过采样 / 欠采样
  - 用算法合成少数类样本
  - 在损失函数中增加少数类权重，如设置class_weight。
  - 使用集成学习方法，对少数类进行重点学习。
  - 用AUC、F1-score、召回率等对模型效果进行评价，而不是仅看准确率。

线性回归与逻辑回归

1. logistic回归公式是什么?

Logistic回归的基本公式如下：

- 输出概率计算公式（Sigmoid函数）： $$ P(y=1|x) = \sigma(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}} $$
- 对数几率（Logit）公式： $$ \log \left(\frac{P(y=1|x)}{1-P(y=1|x)} \right) = w^T x + b $$

- 损失函数（对数损失）： $$ L = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1-y_i) \log(1-p_i)] $$
- 梯度：

$$ \frac{\partial L}{\partial w} = \frac{\partial L}{\partial p} \cdot \frac{\partial p}{\partial z} \cdot \frac{\partial z}{\partial w} $$

$$ \frac{\partial L}{\partial p} = -\frac{y}{p} + \frac{1-y}{1-p} $$ $$\frac{\partial p}{\partial z} = p(1-p) $$ $$ \frac{\partial z}{\partial w} = x $$

$$ \frac{\partial L}{\partial w} = [-\frac{y}{p} + \frac{1-y}{1-p}] \cdot p(1-p) \cdot x $$

$$ \frac{\partial L}{\partial w} = (p - y) x $$

$$ \frac{\partial L}{\partial w} = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)x_i $$

$$ \frac{\partial L}{\partial b} = p - y $$

2. 逻辑回归相较于线性回归，有何异同?

属性	线性回归（Linear Regression）	逻辑回归（Logistic Regression）
任务类型	回归（预测连续数值）	分类（二分类/多分类）
输出范围	实数（无界）	概率（0到1之间）
激活函数	无	Sigmoid函数
损失函数	均方误差（MSE）	交叉熵损失（对数似然损失）
输出解释	直接给出预测值	输出为类别概率，再通过阈值生成分类结果
先验分布	要求残差服从正态分布	对数几率服从线性模型

3. 逻辑回归处理多标签分类问题怎么做?

逻辑回归处理多标签时，实际上就是为每个标签分别拟合一个二分类器，结果是每个标签得到一个概率，样本可以有多个标签为正。

逻辑回归用于**多类别分类（multiclass classification）**时，主要有两种常见方法：

4. 逻辑回归处理多类别分类问题怎么做?

- **一对多（One-vs-Rest, OvR）策略**
    - 原理：针对每个类别，训练一个二分类器，将“该类别”与“其它所有类别”区分开来。
    - 如果有 (K) 个类别，要训练 (K) 个二分类逻辑回归模型。
    - 预测时，对每个类别得到一个分数/概率，选最大者作为最终类别。
    - 很多库（如 scikit-learn）默认实现多类别逻辑回归采用 OvR。
- **Softmax（多项逻辑回归，Multinomial Logistic Regression）**
    - 原理：对所有类别同时建模，输出每类属于该样本的概率，满足所有概率和为1。
    - 使用 Softmax 函数将线性组合结果映射为概率分布：

$P(y=k|x) = \frac{\exp(w_k^T x + b_k)}{\sum_{j=1}^K \exp(w_j^T x + b_j)}$

- 损失函数为多类别交叉熵（categorical cross-entropy）：$ L = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \log P_{ik}$
- 此方式是“多项回归”或“Softmax回归”，是最直接的多类别方法。
- scikit-learn 支持 `multi_class='multinomial'` 实现。

- **如何选择**
  - 类别数较多且彼此独立时，建议用 Softmax；
  - 类别严重不平衡时，可以考虑 One-vs-Rest；
  - 多标签分类不是用 Softmax，而用 One-vs-Rest（参见上一问）。

朴素贝叶斯模型

1. 写出全概率公式和贝叶斯公式

- **全概率公式**
  假设事件A与一组互斥且完备的事件${B_i}$有关，则 $ P(A) = \sum_{i} P(A \mid B_i) \cdot P(B_i)$

- **贝叶斯公式**
  在上面的条件下，贝叶斯公式为 $ P(B_i \mid A) = \frac{P(A \mid B_i) \cdot P(B_i)}{P(A)}$
  - $P(A \mid B_i)$：在 B_i 发生的前提下，A 发生的概率（似然）。【观测到数据，在某条件下数据出现的概率。】
  - $P(B_i)$：先验概率。【不看新数据，原本的主观相信。】
  - $P(B_i \mid A)$：【结合原有认知和新数据后修正的概率。】

2. 朴素贝叶斯为什么朴素？naive?

- 朴素贝叶斯之所以叫“朴素”（Naive），是因为它对特征之间的关系做了一个非常朴素、简单的假设：**在类别已知的前提下，所有特征彼此条件独立。** $ P(x_1, x_2, \ldots, x_n | y) = \prod_{i=1}^n P(x_i | y) $

3. 朴素贝叶斯有无可调的超参数?

- 朴素贝叶斯核心可调超参数是拉普拉斯平滑系数（alpha），其余参数主要是模型类型选择和相关实现细节。超参数较少，也是它简单高效的一大特色。

4. 朴素贝叶斯的工作流程?

朴素贝叶斯的工作流程可以分为训练和预测两个阶段，简明梳理如下：

- 训练阶段
  - **计算先验概率**：统计每个类别的样本出现频率，得到 $P(y)$。
  - **计算条件概率**:对每个类别，统计每个特征在该类别下每种取值的频率，得到$P(x_i | y)$。连续型特征常假设高斯分布，估计均值和方差。

- 预测阶段

- **组合概率**:对新样本，计算它属于每个类别的**后验概率**,可以用对数概率防止下溢。 $ P(y|x_1,x_2,...,x_n) \propto P(y) \prod_{i=1}^n P(x_i|y)$
- **类别判定**:选取后验概率最大（MAP准则）的类别作为预测结果。

- 补充连续数值问题
    - 对于连续型特征，一般**假设其在每个类别下都服从正态分布（高斯分布）**： $ P(x_i \mid y = c) = \frac{1}{\sqrt{2\pi\sigma_{c,i}^2}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right) $
    - 用训练数据，分别统计每个类别下每个连续特征的**均值和方差**。
    - 测试样本时，就用统计得到的高斯分布公式，计算当前特征值在对应类别下的概率密度。

- 补充离散数值问题
    - **零概率问题**: 如果训练数据里在某一类别下，特征x_i从未出现过某个取值v，那么分子为0，算出来的概率就是0。在后验概率连乘（即所有特征概率相乘）时，只要有一个乘数为0，则整条概率链也为0，导致该类别永远不可能被选中。
    - 拉普拉斯平滑（Laplace Smoothing）:给所有取值的计数都加上一个常数 α（通常为1），即“假装见过一次”。
    - 公式调整为： $ P_{\text{Laplace}}(x_i = v \mid y = c) = \frac{\text{计数}(x_i = v, y = c) + \alpha}{\text{计数}(y = c) + \alpha \cdot N} $

5. 朴素贝叶斯对异常值的敏感程度?

朴素贝叶斯对异常值的敏感程度取决于**特征类型**和**分布假设**：

- **高斯朴素贝叶斯（处理连续型特征）**
    - 通常假设每个特征在每一类别下服从正态分布，参数用**均值和方差**估计。
    - **对异常值很敏感**：极端值会明显影响均值和方差，使概率估计偏向异常值，从而影响分类。
    - 连续特征一旦有离群点，模型计算的概率密度会被拉低或拉高，分类效果下降。

- **多项式/伯努利朴素贝叶斯（处理离散/二值特征）**
    - 条件概率用计数/频率估算，异常值的影响相对较小。
    - 只要特征空间足够大、样本数量足够，极少数“特殊取值”不会主导整体概率。