# 特征工程测试题

1. **什么是特征工程？**

答：本质上来说，呈现给算法的数据应该能拥有基本数据的相关结构或属性。当你做特征工程时，其实是将数据属性转换为数据特征的过程，属性代表了数据的所有维度，在数据建模时，如果对原始数据的所有属性进行学习，并不能很好的找到数据的潜在趋势，而通过特征工程对你的数据进行预处理的话，你的算法模型能够减少受到噪声的干扰，这样能够更好的找出趋势。事实上，好的特征甚至能够帮你实现使用简单的模型达到很好的效果。

**2. 特征工程分哪几步？**

答：①数据预处理；②特征选择；③特征提取。

**3. 数据不平衡的问题如何解决？**

答：①采样的方法，采样分为**上采样（Oversampling）和下采样（Undersampling）**，上采样是把小众类复制多份，下采样是从大众类中剔除一些样本，或者说只从大众类中选取部分样本。上采样后的数据集中会重复出现一些样本，训练出来的模型会有一定的过拟合，解决办法，可以在每次生成新数据点时加入轻微的随机扰动；而下采样的缺点也比较明显，那就是最终的训练集丢失了数据，模型只学到了一部分。可能会导致欠拟合，解决办法：多次下采样（放回采样，这样产生的训练集才相互独立）产生多个不同的训练集，进而训练多个不同的分类器，利用模型融合，通过组合多个分类器的结果得到最终的结果。

​②数据合成方法是利用小众样本在特征空间的相似性来生成新样本，这类方法在小数据场景下有很多成功案例，比如医学图像分析等。

​③对于数据极端不平衡时，可以观察观察不同算法在同一份数据下的训练结果的precision和recall，这样做有两个好处，一是可以了解不同算法对于数据的敏感程度，二是可以明确采取哪种评价指标更合适。针对机器学习中的数据不平衡问题，建议更多PR(Precision-Recall曲线)，而非ROC曲线，具体原因画图即可得知，如果采用ROC曲线来作为评价指标，很容易因为AUC值高而忽略实际对少量样本的效果其实并不理想的情况。

​④选择对数据倾斜相对不敏感的算法，如树模型。

**4. 简单介绍特征选择？**

答：当数据预处理完成后，我们需要选择有意义的特征输入机器学习的算法和模型进行训练。通常来说，从两个方面考虑来选择特征：

* 特征是否发散：如果一个特征不发散，例如方差接近于0，也就是说样本在这个特征上基本上没有差异，这个特征对于样本的区分并没有什么用。
* 特征与目标的相关性：这点比较显见，与目标相关性高的特征，应当优选选择。除方差法外，本文介绍的其他方法均从相关性考虑。

根据特征选择的形式又可以将特征选择方法分为3种：

* Filter：过滤法，按照发散性或者相关性对各个特征进行评分，设定阈值或者待选择阈值的个数，选择特征。如方差选择法，先要计算各个特征的方差，然后根据阈值，选择方差大于阈值的特征。
* Wrapper：包装法，根据目标函数（通常是预测效果评分），每次选择若干特征，或者排除若干特征。递归消除特征法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。
* Embedded：嵌入法，先使用某些机器学习的算法和模型进行训练，得到各个特征的权值系数，根据系数从大到小选择特征。类似于Filter方法，但是是通过训练来确定特征的优劣。使用带惩罚项L1的基模型，除了筛选出特征外，同时也进行了降维。L1惩罚项降维的原理在于保留多个对目标值具有同等相关性的特征中的一个，所以没选到的特征不代表不重要。故可结合L2惩罚项来优化。具体操作为：若一个特征在L1中的权值为1，选择在L2中权值差别不大且在L1中权值为0的特征构成同类集合，将这一集合中的特征平分L1中的权值；树模型中GBDT也可用来作为基模型进行特征选择。

**5. 简述SVD(奇异值分解)工作原理?**

答：SVD是一种矩阵分解方法，相当于因式分解，他的目的纯粹就是将一个矩阵拆分成多个矩阵相乘的形式。能用小得多的数据集来表示原始数据集，这样可以去除噪声和冗余信息。适用于数值型数据。

SVD将原始数据集矩阵Data分解成三个矩阵U,E,VT,Data是m行n列，则三个矩阵分别为m×m，m×n，n×n，E为对角阵，这些对角元素称为奇异值。奇异值其实是矩阵Data×Data^T特征值的平方根。

优点：简化数据，去除噪声，提高算法的结果；

缺点：数据的转换可能难以理解，分解出的矩阵解释性往往不强。

**6. 简述线性判别分析LDA（Linear Discriminate Analysis）工作原理？**

答：LDA是一种监督学习的降维技术，也就是说它的数据集的每个样本是有类别输出的。这点和PCA不同，PCA是不考虑样本类别输出的无监督降维技术。LDA的思想可以用一句话来概括，就是“投影后类内方差最小，类间方差最大”。我们要将数据在低维度上进行投影，投影后希望每一种类别数据的投影点尽可能地接近，而不同类别的数据的类别中心之间的距离尽可能地大。

**7. 标准化和归一化异同？**  
答：这俩都是线性变换，将原有数据进行了放缩，归一化一般放缩到[0,1]，标准化则转为服从标准正态分布，数据的大小顺序没有发生改变。影响归一化的主要是两个极值，而标准化里每个数据都会影响，因为计算均值和标准差。

**8. 你是如何处理CTR类特征？**  
答：放缩然后离散化。

**9. 序号编码、one-hot编码、二进制编码都是什么？适合怎样的类别型数据？**

答：

* 序号编码通常用于处理类别建具有大小关系的数据
* one-hot编码用于处理类别建不具有大小关系的特征。但是当类别取值较多的时候需要注意 (1) 用稀疏向量来节省空间 (2) 配合特征选择来降低维度。
* 二进制编码，先用序号编码给每个类别赋予一个类别id，然后将类别id转为二进制编码，本质上是利用二进制对id进行了hash映射，得到0-1特征向量，且维度少于one-hot编码。

**10. 时间类型数据你的处理方法是什么？原因？**  
答：一般取到小时或者天，太细维度一来可解释性不强，而来造成特征维度增加，数据过于稀疏，也增加训练资源。

**11. 你怎样理解组合特征？举个例子，并说明它和单特征有啥区别**答：组合特征其实就是特征的交叉。举个例子，比方说浏览新闻，给用户推荐新闻的时候如果按性别或者兴趣这个特征来推，可能效果不是特别好，但是经过观察发现，性别和兴趣有比较强的关系，例如女性可能更喜欢看娱乐类，男性更喜欢体育财经类，把这俩特征做组合加入到模型，效果比单特征要强很多。

**12. 如何理解笛卡尔积、外积、内积？**  
答：笛卡尔积是两两相乘，m条记录和n条记录的结果为mn条记录，内积是两个向量点乘，结果是一个数值，外积是列向量乘以行向量，结果是一个矩阵。

**13. 文本特征表示有哪些模型？他们的优缺点都是什么？**

答：

* 词袋模型：词汇需要仔细的设计，特别是为了管理文档的大小，这会影响文档表示的稀疏性；简单但是失去了词的上下文结构；由于计算的原因（空间和时间复杂性）以及信息的原因，稀疏表示更难模拟，因为模型在如此庞大的代表空间中利用这么少的信息面临着巨大挑战
* N-gram：将连续出现的n个词组成词组作为一个单独的特征放到向量表示中  
  参数空间过大；数据稀疏
* 主题模型：能够计算出每篇文章的主题分布
* 词嵌入模型：将词映射成 K K K维的向量。

**14. 哪些机器学习算法不需要做归一化（Normalization）处理？**

答：概率模型不需要归一化，因为它们不关心变量的值，而是关心变量的分布和变量之间的条件概率。 如决策树、随机rf。而像adaboost、svm、lr、KNN、KMeans之类的最优化问题就需要归一化。

**15. 对于树形结构为什么不需要归一化？**

数值缩放，不影响分裂点位置。因为第一步都是按照特征值进行排序的，排序的顺序不变，那么所属的分支以及分裂点就不会有不同。对于线性模型，比如说LR，我有两个特征，一个是(0,1)的，一个是(0,10000)的，这样运用梯度下降时候，损失等高线是一个椭圆的形状，这样我想迭代到最优点，就需要很多次迭代，但是如果进行了归一化，那么等高线就是圆形的，那么SGD就会往原点迭代，需要的迭代次数较少。

另外，注意树模型是不能进行梯度下降的，因为树模型是阶跃的，阶跃点是不可导的，并且求导没意义，所以树模型（回归树）寻找最优点是通过寻找最优分裂点完成的，从而不需要数据归一化。