1.评估指标的局限性

准确率的局限性

准确率（Accuracy） = n_correct / n_total

存在问题：当负样本占99%的时候，分类区把所有样本都预测为负样本，也可以获得99%的准确率。当不同类别的样本比例不均衡时，占比大的类别往往称为影响准确率的最主要因素。

可以选择平均准确（每个类别下的样本准确率的算术平均）作为模型评估的指标。

精确率与召回率的权衡

精确率（Precision）= TP/ FP+FP 指分类正确的样本占分类器判定为正样本的样本个数的比例，针对预测结果而言

召回率（Recall） = TP/ TP+FN 指分类正确的样本占真正的正样本个数的比例，针对原本数据集而言

P-R曲线的绘制：通过将阈值（判定样本为正样本的阈值）从高到低移动生成，参考书P24

F1 score = 2*Precision*Recall / precision + recall

平方根误差RMSE

一般用来衡量回归模型

假如模型在95%的时间区段内的RMSE都很小，但在5%的时间区段内的RMSE很大，为什么?

很可能是由于在5%时间区段内存在非常严重的离群点。

解决方法：

（1）在数据预处理的时候把离群点当作噪声点过滤掉

（2）如果离群点不是噪声点，将利群店产生的机制建模，提高模型的预测能力

（3）找更合适的评估指标，比如MAPE（平均绝对百分比误差）

2.ROC曲线

ROC曲线横坐标为假阳性率（FP/N），纵坐标为真阳性率(TP/P)

P为真实的正样本的数量，N为真实的负样本的数量

如何绘制ROC曲线?

通过不断移动分类器的截断点（阈值）来生成曲线上的一组关键点，参考书P28

如何计算AUC?

AUC指的是ROC曲线下的面积，计算AUC只需要沿着ROC横轴做几份。AUC越大，说明分类器性能越好。

ROC相比P-R曲线有什么特点?

当正负样本的分布发生变化时，ROC曲线的形状能够基本保持不变，而P-R曲线的形状一般会发生较剧烈的变化。

3.余弦距离的应用

为什么在一些场景中要使用余弦相似度而不是欧式距离？、

余弦距离 = 1-cos(A,B)， 余弦相似度 = cos(A, B)

余弦相似度只关注向量之间的夹角，并不关心它们的绝对值大小。
余弦相似度在高维度情况下依然保持“相同时为1，正交时为0，相反时为-1”的性质，而欧式距离的数值则受维度的影响，范围不固定。
总体来说，欧式距离提现数值上的绝对差异，而余弦距离体现方向上的相对差异。

余弦距离是否是一个严格定义的距离?

严格定义的距离满足的三条公理：
（1）正定型
（2）对称性
（3）三角不等式（两边之和大于第三边）

4.A/B测试的陷阱

在进行过充分的离线评估之后，为什么还要进行在线A/B测试?

（1）离线评估无法完全消除模型过拟合的影响。
（2）离线评估无法完全还原先上的工程环境。如先上环境的延迟，数据丢失，标签数据缺失等情况。
（3）线上系统的某些商业指标在离线评估中无法计算。以推荐系统为例，离线评估往往只关心ROC曲线、P-R曲线等的改进，而线上评估可以全面了解该算法带来的用户对岸纪律、留存市场、PV访问量等。

如何进行先上A/B测试?

主要手段是进行用户分桶，即将用户分为实验组和对照组。对实验组的用户施以新模型，对对照组用户施以旧模型。确保每个用户每次只能分到一个同中，所选取的用户id是一个随机数，保证样本无偏和独立。

5.模型评估的方法

有哪些主要的验证方法，它们的优缺点是什么?

Holdout验证（70%样本用于训练，30%样本用于模型验证），缺点是计算出来的最后评估指标与原始分组有很大的关系。

交叉验证解决了Holdout的缺点，缺点是时间开销较高。

这两种方法共同缺点是当数据量小的时候，对样本集的划分会让训练集进一步减少，可能会影响模型训练效果。自助法可以解决这一问题。

自助法是基于自主采样法的检验方法。对于总数为n的样本集合，进行n次有放回的随机抽样，得到大小为n的训练集。n次采样过程中有的样本会被重复采样，有的样本没被抽出过。这些没有被抽出过的样本作为验证集，进行模型验证。

6.超参数调优

网格搜索:

采用较大的搜索范围及较小的步长有很大的概率找到全局最优值，但这样花费时间较长。在实际应用中，一般先使用较广的搜索范围和较大的补步长来寻找全局最优值可能的位置，然后会逐渐缩小搜索范围和步长，来寻找更精确的最优值。但由于目标函数一般是非凸的，所以很可能会错过全局最优值。

随机搜索:

在搜索范围中随机选取样本点

贝叶斯优化算法:

网格搜索和随机搜索在测试一个新点时都会忽略前一个点的信息，而贝叶斯优化算法则充分利用了之前的信息，通过对目标函数形状进行学习，找到使目标函数向全局最优值提升的参数。

7.过拟合与欠拟合

降低过拟合风险的方法:

（1）从数据入手，获取更多的训练数据。在图像分类问题上，还可以使用GAN来合成大量的训练数据。

（2）降低模型复杂度。

（3）正则化方法。

（4）集成学习方法。

降低欠拟合风险的方法：

（1）添加新特征。

（2）增加模型复杂度。

（3）减小正则化系数。