机器学习评估方法与评估指标

# 离线评估方法

离线评估的基本原理是在离线环境中，将数据集划分为“训练集”和“测试集”，用“训练集”训练模型，用“测试集”评估模型，根据数据集划分方法的不同，离线评估可以分为以下3种。

## Holdout检验

Holdout检验是基础的离线评估方法，它将原始的样本集合随机的划分为训练集和验证集两部分。通常选择7:3作为划分比例，70%的样本用于模型的训练，30%的样本用于模型的评估。

Holdout的缺点是：在验证集上计算出来的评估指标与训练集和验证集的划分有直接的关系，如果仅进行少量的Holdout检验，则得到的结论存在较大的随机性。为了消除这种随机性，“交叉检验”的思想被提出。

## 交叉检验

k-fold交叉验证：先将全部样本划分成k个大小相等的样本子集；依次遍历这k个子集，每次都先把当前子集作为验证集，其余所有子集作为训练集，进行模型的训练和评估；最后把所有k次的评估指标的平均值作为最终的评估指标。在实际经验中，k通常取10。

留一验证：每次留下一个样本作为验证集，其余所有样本作为测试集。样本总数为n，依次遍历所有n个样本，进行n次验证，再将评估指标求平均得到最终指标。在样本总数较多的情况下，留一验证发的时间开销极大。事实上，留一验证是留p验证的特例。留p验证是指每次留下p个样本作为验证集，而从n个元素中选择p个元素有种可能，因此它的时间开销远远高于留一验证，故很少在实际工程中应用。

## 自助法

不管是holdout检验还是交叉检验，都是基于划分训练集和测试集的方法进行模型评估的，然后当样本规模比较小时，将样本集进行划分会让训练集进一步减小，这可能会影响模型的训练效果。自助法可以在一定程度上解决这个问题。

自助法（Bootstrap）是基于自采样的检验方法：对于总数为n的样本集合，进行n次有放回的随机采样，得到大小为n的训练集。在n次采样过程中，有的样本会被重复采样，有的样本没有被采样过，这些没有被采样的样本作为验证集进行模型验证，这就是自助法的验证过程。

# 离线评估指标

## 准确率

准确率（Accuracy）是指分类正确样本占总样本个数的比例，即：

其中，为被正确分类的样本个数，为总样本个数，TP表示预测为正并且实际也为正的样本个数，TN表示预测为负并且实际也为负的样本个数。

准确率是分类任务中较直观的评价指标，虽然其具有较强的可解释性，但也存在缺陷：当不同类别的样本比例非常不均衡时，占比大的类别往往成为影响准确率的最主要因素。例如：如果负样本占99%，那么分类器把所有样本都预测为负样本可以获得99%的准确率。

## 精确率与召回率

精确率（Precision）是分类正确的正样本个数占分类器判定为正样本的个数的比例。可公式化描述如下：

召回率（Recall）是分类正确的正样本个数占真正的正样本个数的比例。可公式化描述如下：

精确率和召回率是矛盾统一的两个指标：为了提高精确率，分类器需要尽量在“更有把握时”，才把样本预测为正样本，但往往会因为过于保守而漏掉很多“没有把握”的正样本，导致召回率很低。

为了综合反映Precision和Recall的结果，可以使用F1-score（F1值），F1-score是精确率和召回率的调和平均值，其定义如下：

## 均方根误差

均方根误差（Root Mean Square Error，RMSE），经常用于评估回归模型的好坏。其公式化定义如下：

其中，是第i个样本点的真实值，是第i个样本点的预测值，n是样本点的个数。

一般情况下，RMSE能够很好的反映回归模型的预测值与真实值的偏离程度。但如果存在一些偏离程度非常大的离群点，那么即使离群点数量非常少，也会让RMSE指标变得很差。为了解决这个问题，可以使用鲁棒性更强的平均绝对百分比误差（Mean Absolute Percent Error，MAPE）进行评估，MAPE的定义如下：

相比RMSE，MAPE相当于把每个点的误差进行了归一化，降低了离群点带来的绝对误差的影响。

## 对数损失函数

## P-R曲线

## ROC曲线

## 平均精度损失

参考：王喆，深度学习推荐系统。