# KS值

KS用于模型风险区分能力进行评估，KS指标衡量的是好坏样本累计分布之间的差值。好坏样本累计差异越大，KS指标越大，那么模型的风险区分能力越强。

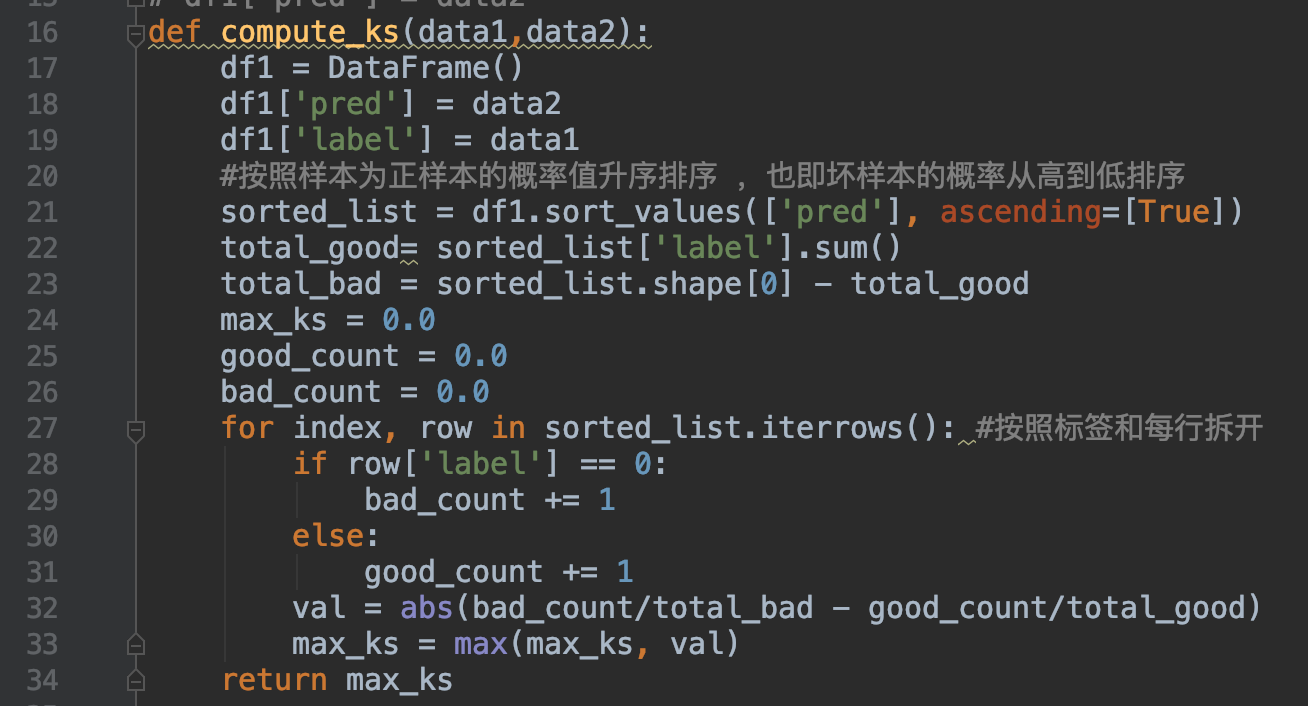
KS值取值范围为[0,1]。通常来说，值越大，表明正负样本区分的程度越好，但并不是所有情况KS都越高越好。如果KS值越大，一般超过0.9，就可以认为正负样本分得过开了，不太可能是正态分布的，反而是比较极端化的分布状态。KS值所代表的仅仅是模型的分割能力，并不代表分割的样本是准确的，即正负样本完全分错，但KS值仍然可以很高。

KS曲线的横轴：阈值or数据集百分比；纵轴：TPR（真正率）与FPR（假正率）

其中，TPR的意义是所有真实类别为1的样本中，预测类别为1的比例；FP的意义是所有真实类别为0的样本中，预测类别为1的比例。

计算：

每个样本的预测结果化为一个概率，对概率升序排列，计算正负样本的累计分布。KS值为两个分布中，最大差值的绝对值。



# AUC

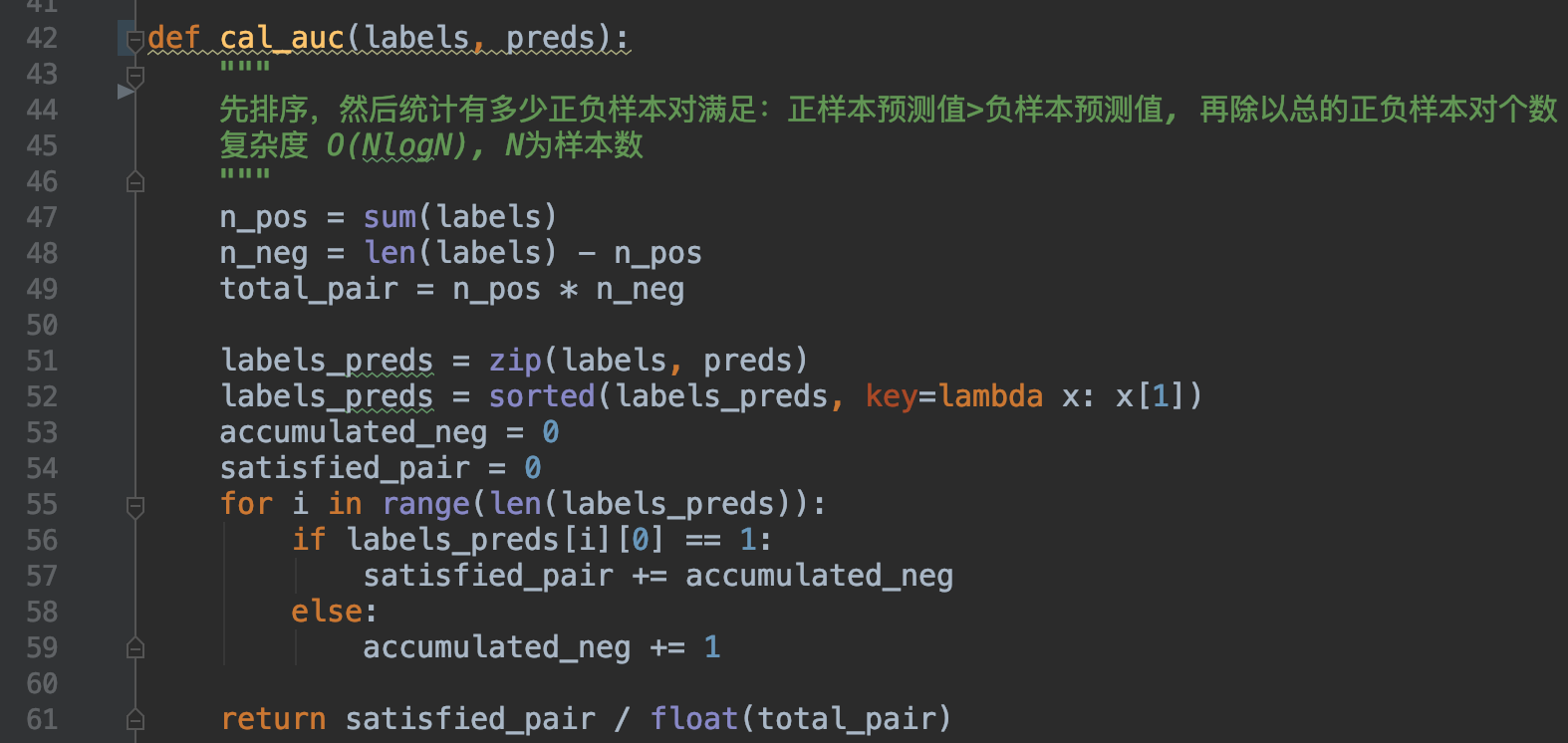
ROC曲线用来评价一个二值学习器泛化性能的好坏。ROC曲线的纵轴是“真正率”（TPR），横轴是“假正率”（FPR）。ROC曲线有个很好的特性：**当测试集中的正负样本的分布变化的时候，ROC曲线能够保持不变。在实际的数据集中经常会出现类不平衡现象，即负样本比正样本多很多（或者相反），而且测试数据中的正负样本的分布也可能随着时间变化。**ROC曲线越接近左上角，该分类器的性能越好。但是，若两个学习器的ROC曲线发生交叉，则需要判别ROC曲线下的面积，即AUC，来决定哪个model更好。

ROC曲线的绘制：

给定m+个正例和m-个反例，根据学习器预测结果对样例进行排序，然后把分类阈值设为最大，即把所有样例均预测为反例，此时真正例率和假正例率均为0，在坐标(0,0)处标记一个点。然后，将分类阈值依次设为每个样例的预测值，即依次将每个样例划分为正例。设前一个标记点坐标为(x,y)，当前若为真正例，则对应标记点的坐标为(x,y+1/m+)；（就是纵轴真正例率up）当前若为假正例，则对应标记点的坐标(x+1/m-,y)，（就是横轴假正例率up），然后用线段连接相邻点即得。

AUC被定义为ROC曲线下的面积，显然这个面积的数值不会大于1。又由于ROC曲线一般都处于y=x这条直线的上方，所以AUC的取值范围在0.5和1之间。使用AUC值作为评价标准是因为很多时候ROC曲线并不能清晰的说明哪个分类器的效果更好，而作为一个数值，对应AUC更大的分类器效果更好。

AUC值是一个概率值，表示随机挑选一个正样本以及一个负样本，当前的分类算法根据计算得到的Score值将这个正样本排在负样本前面的概率就是AUC值。即AUC值越大，当前的分类算法越有可能将正样本排在负样本前面，即能够更好的分类。



# PSI

## 特征稳定性

所谓特征稳定性，就是关注该特征的取值随着时间的推移会不会发生大的波动。对特征稳定性的关注，要在建模之前完成，从一开始就避免将那些本身不太稳定的特征选入模型。

通常采用*PSI*（*Population Stability Index*，群体稳定性指数）指标评估特征稳定性。计算公式如下：

*PSI*是对两个日期的特征数据进行计算，可以任选其一作为base集，另一则是test集（通常日期比较新的作为base）。一般认为PSI小于0.1时候模型稳定性很高，0.1-0.25一般，大于0.25模型稳定性差，建议重做。

计算方法如下：

1. 特征取值等频分段：对这个特征在base集的取值进行等频划分（通常等频分10份即可），用字母表示第个分段区间。
2. 计算：统计落在每个区间内的目标数量（如果是用户特征就是用户数），得到数量占比，表示该特征在base集中第个取值分段中的数量占比。
3. 计算：按照base集的分段标准，采用相同的方法计算。
4. 得到该特征基于这两个日期的*PSI*。

通常，如果一个特征跨度6个月的*PSI*取值小于0.1，那么这个特征被认为是稳定的（也可以根据具体情况适当放宽0.1的标准）。

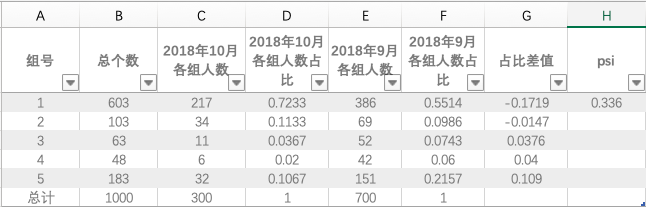
## 模型稳定性

二分类模型的输出一般都会有一个取值为0~1之间的概率值（记作：prediction\_prob），模型*PSI*监控的就是这个值的稳定性。

将模型产出的prediction\_prob理解为一个特征，就可以像计算特征*PSI*一样计算得到模型*PSI*了，不同的地方在于，特征*PSI*一般是对很多特征一起做计算（假如准备了200个特征进行建模，那就是对200个特征计算*PSI*），而模型*PSI*通常只是对prediction\_prob这一个字段做计算。

## PSI计算

下表是某个变量，以2018年10月为基准，每个月（以2018年9月为例）都和2018年10月去做一个下面表格的运算



2018年10月各组人数占比：217/300

占比差值：386/700-217/300

*PSI*：（386/700-217/300）\* ln（(386/700)/(217/300)）

对所有的组的*PSI*进行累加，只要有一个月对比得到的*PSI*值大于0.1，就要把这个变量剔除。