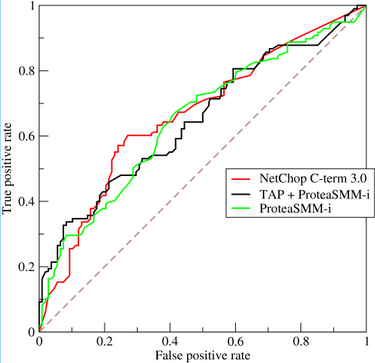
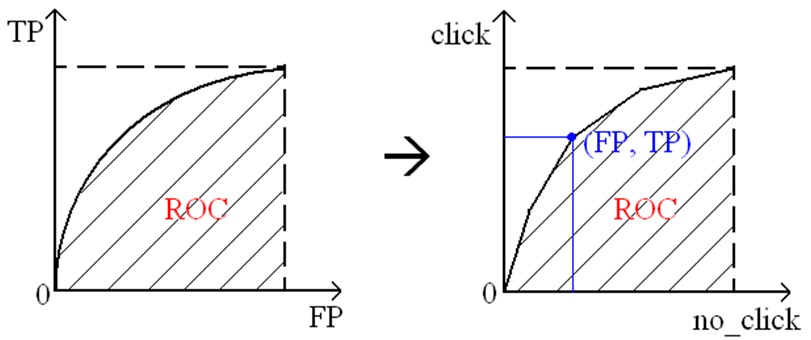
[AUC](http://wiki.babel.baidu.com/twiki/bin/view/Com/Ecom/AUC)是The Area Under an ROC Curve的缩写。Wikipedia上对于ROC curve的定义和画法描述如下:

In signal detection theory, a receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings.



很多同学都有疑问，我们工作中所用的AUC，看似与此不同。以ctr为例，曲线横轴是no\_click，纵轴是click，如下图右侧所示。



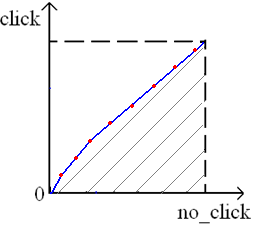
实际上，两者是一致的。

按照预估ctr排序，依次绘制线段的过程就等价于变动阈值绘制ROC曲线的过程。横轴的no\_click就是FP（false positive，表示prediction为positive，但是predict错了），对应过来就是，预估会发生点击（预估ctr大于当前阈值），但是预估错了，实际没有发生点击（no\_click）的样本量。纵轴的click就是TP（true positive，表示prediction为positive，并且predict对了），就是预估会发生点击，实际确实发生了点击的数量。算法最后会除以矩形面积进行归一化，因此可以认为横、纵坐标就是false positive rate和true positive rate，与标准定义是一致的。

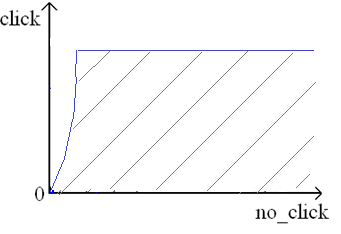
实现时，需要注意一个细节：预估score相同的样本，要提前合并，否则计算结果会不稳定（不同排序算法得到的结果可能不同）。除非限定show=1，click=0或1，并且排序时把整行数据作为key排序（不只把pctr作为key）。

[AUC](http://wiki.babel.baidu.com/twiki/bin/view/Com/Ecom/AUC)的大小反映了预估score与真实结果序关系的一致性。从图上很容易理解，绘制ROC曲线时，每条线段的斜率为click/no\_click，与真实ctr的序是一致的。斜率越大（真实ctr越大）的线段排在前面，曲线的面积越大，AUC就越大，反则反之。当预估结果为与真实结果的序关系完全一致时，AUC达到上限。

通常我们只关注AUC的相对变化，绝对值意义不是很大。有人会存在这样的误区：AUC接近0.5表明预估结果差。这是不准确的，需要参考数据分布。例如大量数据自身label为0.5，此时AUC上限也不会太高（如下图所示）。



类似地，AUC很高（例如0.9x）也不能表明预估结果很好。此时需要参考ROC曲线形状，避免出现尾部平坦，头部预估较差的情况（如下图所示）。



更多详细内容，请参考网上相关资料。

一句话求AUC

单个pctr文件可以用的方法：

sort PCTR\_FILE | awk '($3==0){++x;a+=y;}($3==1){++y;}END{print 1.0-a/(x\*y);}'

注：上面这个命令是其他同学给出的（在此感谢一下）。注意该方法默认show=1，click=1或0，否则不适用。

一句话求auc中a/x\*y实际上是曲线上方的面积。 求下方面积的求法是： sort -k1nr PCTR\_FILE |awk '($3==0) ++x; a+=y}($3==1){++y;}END(print a\*1.0/(x\*y)}' 只是sort方向变了，然后不需要用1去减。