Assignment 5

Name: Han Siyue

ID: 17307110012

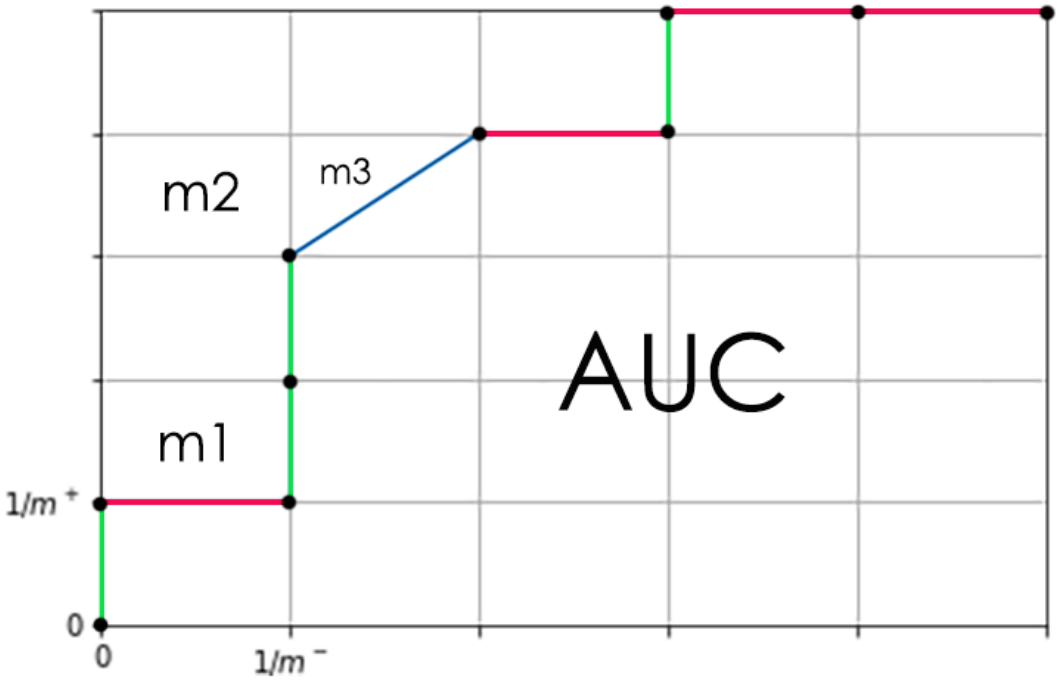
October 26, 2019

假设当真正例与假正例预测值相等时，ROC曲线会呈斜线上升，证明西瓜书p35式(2.22)

已知公式 (2.21)

$$l_{rank}=\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}\left(f(x^+)<f(x^-)\right)+\frac{1}{2}\mathbb{I}\left(f(x^+)=f(x^-)\right)\right)$$

此公式正如书上所说，l_{rank}为ROC曲线之上的面积，要证 $\mathrm{AUC}=1-l_{rank}$，且我们已知 AUC 为ROC曲线之下的面积，所以只需证明公式 (2.21) 等号右边表示的是ROC曲线之上的面积即可，假设某ROC曲线如下图所示：



观察ROC曲线易知：

- 每增加一条绿色线段对应着有1个正样例（x_i^+）被模型正确判别为正例，且该线段在Y轴的投影长度恒为$\frac{1}{m^+}$；
- 每增加一条红色线段对应着有1个反样例（x_i^-）被模型错误判别为正例，且该线段在X轴的投影长度恒为$\frac{1}{m^-}$；

- 每增加一条蓝色线段对应着有a个正样例和b个反样例同时被判别为正例，且该线段在X轴上的投影长度=$b*\frac{1}{m^-}$，在Y轴上的投影长度=$a*\frac{1}{m^+}$；
- 任何一条线段所对应的样例的预测值一定小于其左边和下边的线段所对应的样例的预测值，其中蓝色线段所对应的a+b个样例的预测值相等。

公式里的$\sum_{x^+\in D^+}$可以看成一个遍历x_i^+的循环：

for x_i^+ in D^+:

$$\frac{1}{m^+}\cdot\frac{1}{m^-}\cdot\sum_{x^-\in D^-}\left(\mathbb{I}\left(f(x^+)<f(x^-)\right)+\frac{1}{2}\mathbb{I}\left(f(x^+)=f(x^-)\right)\right)\qquad(\mathrm{S})$$

由于每个x_i^+都对应着一条绿色或蓝色线段，所以遍历x_i^+可以看成是在遍历每条绿色和蓝色线段，并用式S来求出每条绿色线段与Y轴构成的面积（例如上图中的m1)或者蓝色线段与Y轴构成的面积（例如上图中的m2+m3）。

对于每条绿色线段： 将其式S展开可得：
$\frac{1}{m^+}\cdot\frac{1}{m^-}\cdot\sum_{x^-\in D^-}\mathbb{I}(f(x_i^+)<f(x^-))+\frac{1}{m^+}\cdot\frac{1}{m^-}\cdot\sum_{x^-\in D^-}\frac{1}{2}\mathbb{I}(f(x_i^+)=f(x^-))$其中$x_i^+$此时恒为该线段所对应的正样例，是一个定值。$\sum_{x^-\in D^-}\frac{1}{2}\mathbb{I}(f(x_i^+)=f(x^-))$是在通过遍历所有反样例来统计和$x_i^+$的预测值相等的反样例个数，由于没有反样例的预测值和x_i^+的预测值相等，所以$\sum_{x^-\in D^-}\frac{1}{2}\mathbb{I}(f(x_i^+)=f(x^-))$此时恒为0，于是其式S可以化简为：
$\frac{1}{m^+}\cdot\frac{1}{m^-}\cdot\sum_{x^-\in D^-}\mathbb{I}(f(x_i^+)<f(x^-))$其中$\frac{1}{m^+}$为该线段在Y轴上的投影长度，
$\sum_{x^-\in D^-}\mathbb{I}(f(x_i^+)<f(x^-))$同理是在通过遍历所有反样例来统计预测值大于$x_i^+$的预测值的反样例个数，也即该线段左边和下边的红色线段个数+蓝色线段对应的反样例个数，所以
$\frac{1}{m^-}\cdot\sum_{x^-\in D^-}(\mathbb{I}(f(x^+)<f(x^-)))$便是该线段左边和下边的红色线段在X轴的投影长度+蓝色线段在X轴的投影长度，也就是该绿色线段在X轴的投影长度，观察ROC图像易知绿色线段与Y轴围成的面积=该线段在Y轴的投影长度 * 该线段在X轴的投影长度。

对于每条蓝色线段： 将其式S展开可得：
$\frac{1}{m^+}\cdot\frac{1}{m^-}\cdot\sum_{x^-\in D^-}\mathbb{I}(f(x_i^+)<f(x^-))+\frac{1}{m^+}\cdot\frac{1}{m^-}\cdot\sum_{x^-\in D^-}\frac{1}{2}\mathbb{I}(f(x_i^+)=f(x^-))$ 其中前半部分表示的是蓝色线段和Y轴围成的图形里面矩形部分的面积，后半部分表示的便是剩下的三角形的面积，矩形部分的面积公式同绿色线段的面积公式一样很好理解，而三角形部分的面积公式里面的$\frac{1}{m^+}$为底边长，$\frac{1}{m^-}\cdot\sum_{x^-\in D^-}\mathbb{I}(f(x_i^+)=f(x^-))$为高。

综上分析可知，式S既可以用来求绿色线段与Y轴构成的面积也能求蓝色线段与Y轴构成的面积，所以遍历完所有绿色和蓝色线段并将其与Y轴构成的面积累加起来即ROC曲线之上的面积得式 (2.21)，从而可得式 (2.22)。

证毕

Reference

Datawhale, *pumpkin book* , from https://datawhalechina.github.io/pumpkin-book/#/