2.20

$AUC = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)\cdot(y_i+y_{i+1})$

[解析]：由于图2.4(b)中给出的ROC曲线为横平竖直的标准折线，所以乍一看这个式子的时候很不理解其中的$\frac{1}{2}$和(y_i+y_{i+1})代表着什么，因为对于横平竖直的标准折线用$AUC = \sum_{i=1}^{m-1}(x_{i+1}-x_i)\cdot y_i$就可以求出AUC了，但是图2.4(b)中的ROC曲线只是个特例罢了，因为此图是所有样例的预测值均不相同时的情形，也就是说每次分类阈值变化的时候只会划分新增**1个**样例为正例，所以下一个点的坐标为$(x+\frac{1}{m^-},y)$或$(x,y+\frac{1}{m^+})$，然而当模型对某个正样例和某个反样例给出的预测值相同时，便会划分新增**两个**样例为正例，于是其中一个分类正确一个分类错误，那么下一个点的坐标为$(x+\frac{1}{m^-},y+\frac{1}{m^+})$（当没有预测值相同的样例时，若采取按固定梯度改变分类阈值，也会出现一下划分新增两个甚至多个正例的情形，但是此种阈值选取方案画出的ROC曲线AUC值更小，不建议使用），此时ROC曲线中便会出现斜线，而不再是只有横平竖直的折线，所以用**梯形面积公式**就能完美兼容这两种分类阈值选取方案，也即 **(上底+下底)*高***$\frac{1}{2}$

2.21

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

[解析]：

首先，重新对AUC图进行一下解释，AUC的X轴是TPR，Y轴是FPR，这里对TPR和FPR重新进行一下说明，以加深理解，同时点名AUC的变化趋势。

TPR：真正例率，True Positive Rate

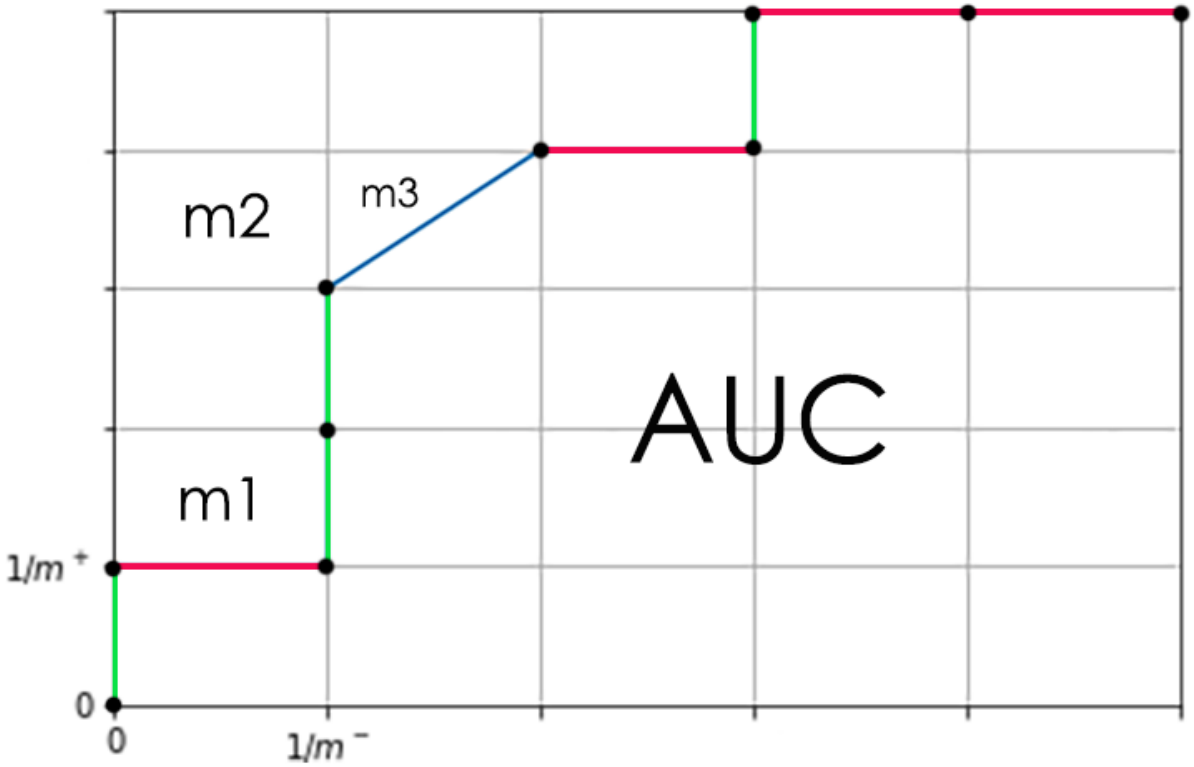
$$TPR = \frac{TP}{TP+FN} = \frac{\text{正}\rightarrow\text{正}}{\text{正}\rightarrow\text{正}+\text{正}\rightarrow\text{反}} = \frac{\text{正}\rightarrow\text{正}}{\text{所有正例的数量}} = \frac{\text{正}\rightarrow\text{正}}{m^+}$$

FPR：假正例率，False Positive Rate

$$FPR = \frac{FP}{TN+FP} = \frac{\text{反}\rightarrow\text{正}}{\text{反}\rightarrow\text{反}+\text{反}\rightarrow\text{正}} = \frac{\text{反}\rightarrow\text{正}}{\text{所有反例的数量}} = \frac{\text{反}\rightarrow\text{正}}{m^-}$$

按照书上所示，在刚开始的时候，分类阈值设置到最大，这个时候，反例会被为反例，FPR=0，正例会被判为反例，TPR=0，之后降低分类阈值，正例被判为正例的数量越来越多，反例被判为正例的也越来越多，直至最后，分类阈值被降为最低的时候，正例和反例都会被判为正例。TPR和FPR都为1。此外，m^+代表样例中正例的数量，是定值。m^-代表样例中负例的数量，是定值。

此公式正如书上所说，l_{rank}为ROC曲线**之上**的面积，假设某ROC曲线如下图所示：



上图的要点如下：

- Y轴向上变化，就是正例预测为正例，
- X轴向右变化，就是反例预测为正例。
- 计算的是曲线与Y轴的面积。

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，这个地方可以根据分类阈值的变化规律去理解，左下的部分（即在开始的时候），分类阈值较大，现在（任何一条线段对应的时刻）分类阈值较小，而TPR和FPR是和分类阈值息息相关的，举个生活中的例子，A在及格线是90 的时候就及格了，而B在及格线未60的时候才及格，则B的成绩肯定小于A的。在这里，就是线段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，这个地方是因为从蓝线的左端点到右端点，分类阈值只变化了一次，而根据AUC的绘制过程，将分类阈值一次设定为每个样例的预测值，所以，蓝线对应的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^+:

$$\frac{1}{m^+}\cdot\frac{1}{m^-}\cdot\sum_{x^-\in D^-}(\mathbb{I}(f(x_i^+)<f(x^-))+\frac{1}{2}\mathbb{I}(f(x_i^+)=f(x^-))) \text{ #记为式S}$$

这里x_i^+的含义是正例中的一个样例，其预测结果只可能有两种正 $\rightarrow$ 正和正 $\rightarrow$ 负，在计算的时候只关心后者，即正 $\rightarrow$ 负，所以每个x_i^+都对应着一条绿色或蓝色线段，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，即$x^-\in D^-$是遍历所有的负例，$\mathbb{I}(f(x_i^+)=f(x^-)$的含义是正例的预测值和反例的预测值是相等的。由于没有反样例的预测值和x_i^+的预测值相等（在蓝线会存在这种情况，在绿线不存在）。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，这个时候，反例会被预测为正例，原因是反例的预测值太高。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（只能数左下的，左下的阈值较大），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（绿线在X的投影长度为0），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得l_{rank}。

脚注：ROC曲线

roc曲线：接收者操作特征（receiveroperating characteristic）,roc曲线上每个点反映着对同一信号刺激的感受性。**横轴：负正类率(false postive rate FPR)特异度**，划分实例中所有负例占所有负例的比例；(1-Specificity)，**纵轴：真正类率(true postive rate TPR)灵敏度**，Sensitivity(正类覆盖率)。 参考： ROC和AUC介绍以及如何计算AUC ROC曲线-阈值评价标准