1. 数据集包含1000个样本，其中500个正例，500个反例，将其划分为包含70%样本的训练集和30%样本的测试集用于留出法评估，试估算共有多少种划分方式。

“

解答：

- 保证训练集和测试集合同分布(正反例比例相同)

  问题转换为从500个正例抽取350个样本，从500个反例里面抽取350个样本。总共有多少抽取方式

$$C_{500}^{350} * C_{500}^{350}$$

---

1. 数据集包含100个样本，其中正反例各一半，假定学习算法所产生的模型是将新样本预测为训练样本数较多的类别（训练样本数相同时进行随机猜测），试给出用10折交叉验证法和留一法分别对错误率进行评估所得的结果。

“

解答：

- 10折交叉验证

10折交叉验证：同分布将数据分成10份。取其中9份为训练样本，1份为测试样本。

```
1．正反样例相同并且同分布情况下，本题学习出来的模型就是随机猜测
2．测试样本也是5个正样本，5个负样本
```

因此每次学习后，测试的错误预估样本数为5， 错误率为50%， 10轮的平均错误率50%

- 留1法验证

留1法: 取1个样例做为测试，剩余100-1=99个样例作为训练。依次轮询100次，取平均错误率。

```
1．考虑留下样例为反例，测试集合50个正例，49个反例。模型判断测试样本为正例
2．考虑留下为样例为正例，测试集合为49个正例，50个反例。模型判断测试样例为反例
```

因此每次学习后，错误率100%

---

1. 若学习器A的F1值比学习器B高，试析A的BEP值是否也比B高。

“

解答：

分类结果的混淆矩阵

真实情况	预测结果	预测结果
	正例	反例
正例	TP(真正例)	FN(假反例)
反例	FP(假正例)	TN(真假例)

- 查准率(准确率),预测的正例中，真实的正例的占比。 $P(precision) = \frac{TP}{(TP+FP)}$
- 查全率(召回率),所有正例，其中被预测出来的真正正例的占比 $R(precision) = \frac{TP}{(TP+FN)}$
- 平衡点(Break-Event Point, BEP), 学习器当P = R 时候的取值
- $F_1 = \frac{2*P*R}{P+R} == \frac{2*TP}{Total+TP-TN}$
- 考虑F1为动态曲线时候，既学习器A的F1 在P和R变动时候 恒大于 学习器B的F1

$F_{1A} = \frac{2*P*R}{P+R} > F_{1B} = \frac{2*P*R}{P+R}$ 既然是恒大

---

1. 试述真正例率（TPR）、假正例率（FPR）与查准率（P）、查全率（R）之间的联系。

“

解答:

- TRP真正例率，也就是预测结果中的正例占所有正例的比率。$TRP = \frac{TP}{TP+FN}$
- FPR假正例率，也就是预测结果中假的正例占所有反例的比率。$FPR = \frac{FP}{TN+FP}$

从公式上看，TRP = P(查准率/准确率)

1. 试证明(2.22) $AUC = 1 - l_{rank}$

解答:

- 考虑正反列没有预测分数一样的情况，既$f(x^+) != f(x^-))$

设正例集合

$$D^+ = \{d_i^+ | 0 < i <= m^+\}$$

反例集合为

$$D^- = \{d_i^- | 0 < i <= m^-\}$$

整个ROC的二维面积空间为 1*1的正方形空间。不失一般性，设正例 d_j 在学习器得分排序前有k个反例。

d_j 的ROC坐标为:

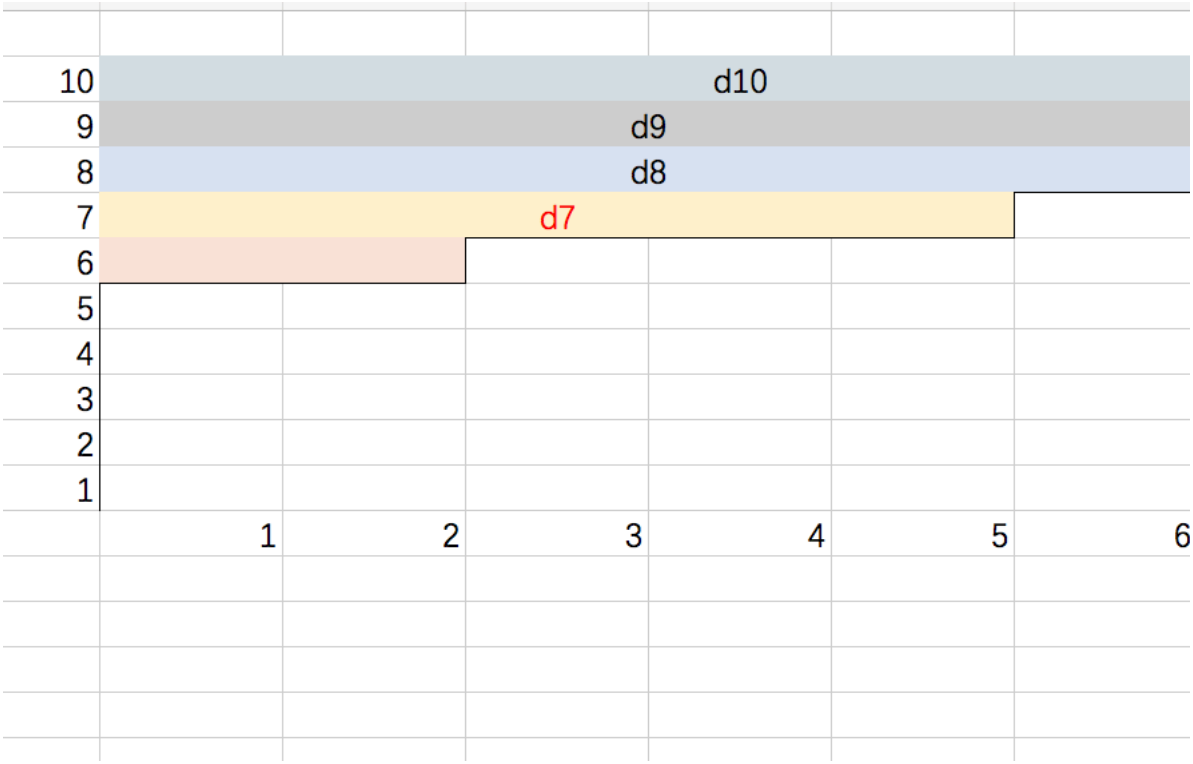
$$\left(x_j, y_j\right), x_j = \frac{k}{m^-}$$

跟进ROC曲线图规则可知，在依次画图中，每一个正例点相对前一个样例点(x,y)其实垂直上升 $\frac{1}{m^+}$ 。每个反例点相对前一个样例(x,y)其实是水平横移动了 $\frac{1}{m^-}$

定义单个正例点产出的ROC曲线外覆盖面积如下

$$V(j,k) = \frac{k}{m^-} * \frac{1}{m^+} = \frac{1}{m^+ \times m^-} \times k$$

ROC图上的V(j,k)面积如下图所示:



图为选10个正样例，10个负样例的ROC图， 其中黑色的折线图为ROC曲线，白色空间为AUC。 d8表示第8个正例，容易知d8 前面有6个反例，d8的 $V(8,6) = \frac{6}{m^-} * \frac{1}{m^+=10} | m^-, m^+ = 10$

令

$$OUC = \sum_{d_i^+ \in D^+} V(i,k) = \frac{1}{m^+ \times m^-} \sum_{i \in D^+} k$$

其中k 为学习器中得分大于d_i的反例个数。 可知

$$AUC = 1 - OUC$$

因此

$$k = \sum_{d_j^- \in D^-} Count(f(d_i^+) < f(d_j^-))$$

代入OUC中得

$$OUC = \frac{1}{m^+ \times m^-} \sum_{i \in D^+} \sum_{d_j^- \in D^-} Count(f(d_i^+) < f(d_j^-)) = l_{rank}$$

- 考虑存在正反列预测分数一样的情况

不失一般性，假设在依次变量样本点 $d \in D$ ，上存在i个正例和k个反例得分都为$f(d \in D)$。这时候我们考虑相对前一个点(x, y)，当前点的增幅情况.

```
1. 由于负例增加了k个，相对横坐标则右平移了 k个单位
2. 由于正例增加了i个，相对纵坐标则上移动了 i个单位
```

其坐标为$(x + \frac{k}{m^-}, y + \frac{i}{m^+})$.
计算相等的正负例，其覆盖的面积为一个矩形对角划分的三角形，其面积为

$$\frac{1}{2} \times \frac{k}{m^-} \times \frac{i}{m^+}$$

代入到OAU中得证明。

---

1. 试述错误率与ROC曲线之间的关系
2. TRP真正例率，也就是预测结果中的正例占所有正例的比率。$TRP = \frac{TP}{TP+FN}$
3. FPR假正例率，也就是预测结果中假的正例占所有反例的比率。$FPR = \frac{FP}{TN+FP}$
4. $ErrorRate = \frac{FP+FN}{m^{+}+m^{-}}$

ROC曲线由以上两个值组成的点构成(FPR, TRP)

$$ErrorRate = \frac{FP + FN}{m^+ + m^-} = \frac{m^- \times FPR + m^+ - m^+ \times TRP}{m^- + m^+}$$

因此可以让学习器选择一个阈值，使得错误率最低。