2.1

数据集共 1000 个样本，所以训练集包含 700 个样本，测试集包含 300 个样本。为了保持数据分布一致，采用分层采样，则训练集包含 350 个正例，350 个反例。测试集包含 150 个正例，150 个反例。因此划分流程为从 500 个正例中取出 350 个样本，500 个反例中取出 350 个样本，作为训练集，剩下的样本作为测试集。则划分方式有$\binom{500}{350} \times \binom{500}{350}$种。

2.2

解题思路：首先划分训练集测试集，然后根据模型的训练结果对测试集进行测试。

10 折交叉验证：就是通过分层采样将样本划分为 10 个互斥子集，每个子集共 10 个样本，正反例各一半。每次训练取其中 9 个子集作为训练集，1 个子集作为测试集。因为正反例数相同，因此模型将测试集种的样本预测为正例或反例的概率均为 50%，因此预测错误率为 50%。

留一法：每次取 99 个样本作为训练集，1 个样本作为测试集。如果测试集样本为正例，则训练集样本分布为正例 49，反例 50，则模型会把测试集的样本预测为反例，预测错误。如果测试集样本为反例，同理，模型会把测试集样本预测为正例。因此总体预测错误率为 100%。

2.3

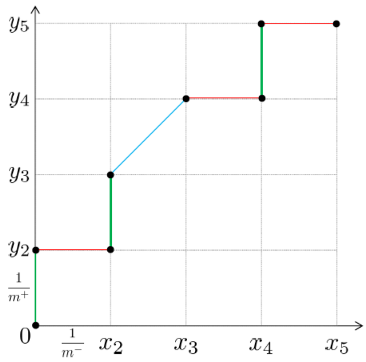
根据书中定义：BEP 就是 P-R 曲线中 P=R 的取值，F1=$\frac{2P\times R}{P+R}$。因此$F1_{P=R} = BEP$。因为$\mathrm{F1_A} > \mathrm{F1_B}$，所以$\mathrm{BEP_A} > \mathrm{BEP_B}$。

2.4

参考四者的公式可以发现 R=TPR。

2.5

参考南瓜书的公式推导：图 2.20 公式推导，ROC 绘制曲线图如下图，因此 AUC 可以用梯形公式计算。即绿色、蓝色、红色线下面的面积。



要证明 AUC = 1-l_{rank}，即证明 l_{rank} 是绿色、蓝色、红色线上面的面积。即 l_{rank} 同样为梯形面积的和。对 l_{rank} 公式转换，过程如下(摘自南瓜书)：

$$\begin{aligned}
\ell_{rank} &= \frac{1}{m^+m^-}\sum_{\boldsymbol{x}^+\in D^+}\sum_{\boldsymbol{x}^-\in D^-}\left(\mathbb{I}\left(f(\boldsymbol{x}^+)<f(\boldsymbol{x}^-)\right)+\frac{1}{2}\mathbb{I}\left(f(\boldsymbol{x}^+)=f(\boldsymbol{x}^-)\right)\right)\\
&= \frac{1}{m^+m^-}\sum_{\boldsymbol{x}^+\in D^+}\left[\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f(\boldsymbol{x}^+)<f(\boldsymbol{x}^-)\right)+\frac{1}{2}\cdot\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f(\boldsymbol{x}^+)=f(\boldsymbol{x}^-)\right)\right]\\
&= \sum_{\boldsymbol{x}^+\in D^+}\left[\frac{1}{m^+}\cdot\frac{1}{m^-}\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f(\boldsymbol{x}^+)<f(\boldsymbol{x}^-)\right)+\frac{1}{2}\cdot\frac{1}{m^+}\cdot\frac{1}{m^-}\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f(\boldsymbol{x}^+)=f(\boldsymbol{x}^-)\right)\right]\\
&= \sum_{\boldsymbol{x}^+\in D^+}\frac{1}{2}\cdot\frac{1}{m^+}\cdot\left[\frac{2}{m^-}\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f(\boldsymbol{x}^+)<f(\boldsymbol{x}^-)\right)+\frac{1}{m^-}\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f(\boldsymbol{x}^+)=f(\boldsymbol{x}^-)\right)\right]
\end{aligned}$$

其中，$\sum_{x^+\in D^+}$即为梯形的求和部分，$\frac{1}{m^+}=y_{i+1}-y_i$即为梯形的高。

梯形的上底为$\frac{1}{m^-}$乘以预测值比x^+大的假正例的个数，即为

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

下底为$\frac{1}{m^-}$乘以预测值大于等于 x^+的假正例的个数，即为

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

2.6

错误率就是预测错误的概率，因此根据混淆矩阵可得：

错误率$E=\frac{FP+FN}{m^++m^-}=\frac{FP+m^+-TP}{m^++m^-}=\frac{m^-*\frac{FP}{m^-}+m^+*(1-\frac{TP}{m^+})}{m^++m^-}=\frac{m^-*FPR+m^+*(1-TPR)}{m^++m^-}$，在样本确定的情况下$m^+$和$m^-$为定值，则 *FPR* 变大同时 *TPR* 变小时，错误率变大；同时根据 ROC 曲线定义，整条曲线向下偏移，AUC 变小，学习器性能变差。

2.7

根据代价敏感曲线的绘制过程，ROC 曲线的每个点对应代价敏感曲线的一个线段。即 ROC 的点(TPR, FPR)和代价敏感曲线的线段[(0,FPR)(1,FNR)]是一一对应的。因此 ROC 曲线与代价曲线也是一一对应的。

2.8

Min-Max 规范化适用于最大最小值已知的情形。缺点在于当有新数据输入时，可能导致 max 和 min 的变化，需要重新定义。

z-score 规范化是把数据变为了标准正态分布，适用于最大值或最小值未知的情况，或有超出取值范围的离群数据的情况，并且要求样本数量较大。

2.9 (卡法统计量的公式有很多写法，此处 step-2 正确性有待验证)

已知：对算法进行 k 次测试的，得到 k 个测试错误率$\hat{\epsilon}_1,\hat{\epsilon}_2\cdots\hat{\epsilon}_k$。先对假设$\mu=\epsilon_0$进行验证

Step-1：根据式 2.28、2.29 计算均值和方差。

Step-2：计算卡方统计量$\frac{(\mu-\epsilon_0)^2}{\sigma}$。

Step-3：若卡方统计量超过卡方临界值χ_{α^2}，则拒绝该假设。

2.10

2.34 公式为卡方分布，而 2.35 是 F 分布。因此两个公式的区别就是两个分布的区别，即书中所说的 2.34 相较于 2.35 过于保守。过于保守的意思是：卡法分布只体现数据出现频数的差异，而 F 分布则体现了数据本身的差异。

举个例子，假设三个算法 A、B、C 通过比较后的平均序值为 1.2、1.4、1.6，现在对“三个算法性能相同”这个假设进行检验：

如果我们设定了卡方分布值域为{1,2}，则卡法分布会接受这个假设，而 F 分布会拒绝这个假设。