# 机器学习课程　第2次作业

黄昊　　*20204205*

选择题目2.1　2.3　2.7　2.8　2.9

2.1　正、反例各抽70%的样本，那么共有$C_{500}^{150}C_{500}^{150}$种划分方式。

2.3　由公式(2.10)和$BEP$的定义可知，两者无直接关系。因为$BEP$规定了$BEP = P = R$，但$F1$对P和 R没有具体规定大小关系，不可直接比较。

2.7　已知ROC曲线和FPR与FNR=1-TPR存在一一对应的关系，绘制代价曲线时，ROC 曲线的每一个点和一对FPR、FNR一一对应，转化为代价曲线上的一条直线。将所有直线画出，其包络线（有限情况下围成一个多边形）即为代价曲线。

由于$FPR = \frac{FP}{TN+FP}, FNR = \frac{FN}{TP+FN}$，FP与FN不可能同时为0，则FPR和FNR不可能同时为0，那么根据代价曲线的绘制过程，代价曲线必然存在；若代价曲线存在，则取代价曲线上的任意一个连续可导的点做代价曲线的切线，则必然能求出所有二元组（FPR，FNR），进而能求出ROC曲线上的所有可绘制点，绘制即得ROC曲线。即可成名每一条ROC曲线均有代价曲线，反之亦然。

2.8　min-max规范化的**优点**是计算简单，当有新元素加入时，只要不是最值，均可以实现O(1)的计算；适合在线处理；**缺点**在于受最值影响大，若样本含有离群点且未作处理将会使其他正常值规范化后所得到的值不合理；

z-score规范化的**优点**在于鲁棒性强，受极端值影响小；**缺点**在于每次新加入点后都需要重新计算，复杂度为O(n)，只适合离线处理。

2.9　卡方检验在比较两个学习器的性能时的步骤如下：

首先对符号做如下约定：

| 算法B | 算法A | |
|---|---|---|
| | 正确 | 错误 |
| 正确 | $\epsilon_{00}$ | $\epsilon_{01}$ |
| 错误 | $\epsilon_{10}$ | $\epsilon_{11}$ |

使用如下统计量：

$$\tau_{\chi^2} = \frac{(|\epsilon_{01} - \epsilon_{10}| - 1)^2}{\epsilon_{10} + \epsilon_{01}}$$

该统计量服从自由度为1的卡方分布。给定显著度$\alpha$，当上述变量小于该值则不能拒绝假设，认为两个学习器的效果没有显著差异；反之则拒绝，认为有明显差异。