

机器学习课程 第2次作业

黄昊 20204205

选择题目 2.1 2.3 2.5 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。

2.5 考虑式(2.20)与式(2.21)

2.7

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	ϵ_{00}	ϵ_{01}
错误	ϵ_{10}	ϵ_{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 α ，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。