

《大数据机器学习》第 2 次作业

姓名：刘培源 学号：2023214278

题目 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。

答：F1 值是基于特定决策阈值下的查准率 P 和查全率 R 的调和平均。而 BEP (Break-Even Point) 则是在查准率 P 和查全率 R 曲线上寻找某个决策阈值，使得查准率和查全率达到均衡。这两个指标虽然都涉及查准率 P 和查全率 R ，但是它们二者之间并没有直接的因果关系。一个具体例子如下：

考虑一组真实数据 1(+), 2(+), 3(+), 4(-), 5(-), 6(-)，其中数字代表了测试样本编号， \pm 代表了正负样例。进一步，假设学习器 A 和 B 的预测结果为：

- 学习器 A : 1(+), 2(+), 4(-), 3(-), 5(-), 6(-)
- 学习器 B : 1(+), 2(+), 4(+), 3(-), 5(-), 6(-)

其中，预测结果中的样例顺序代表对所有样本按为正例可能性的排序，括号内的 \pm 代表学习器 A 和 B 分别取某一个特定决策阈值下的预测结果。我们可以算出：

$$\text{对于学习器 } A: P = \frac{2}{2+0} = 1, R = \frac{2}{2+1} = \frac{2}{3}, F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = 0.8$$

$$\text{对于学习器 } B: P = \frac{2}{2+1} = \frac{2}{3}, R = \frac{2}{2+1} = \frac{2}{3}, F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} \approx 0.667$$

显然，在此情况下，学习器 A 的 F1 值比学习器 B 的 F1 值高。但是注意到，学习器 A 与 B 的 BEP 均在将前三个样本看成正例，后三个样本看成负例时取到。他们的 BEP 均为 $\frac{2}{2+1} = \frac{2}{3}$ 。所以在这个例子中，尽管学习器 A 的 F1 值比学习器 B 的高，他们的 BEP 值却是相等的。

真实情况	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 (真反例)

Table 1: TP, FN, FP, TN 关系表

题目 2: 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。

答: 由表1可得:

- 真正例率 $TPR = \frac{\text{预测为正例且真实为正例的数量}}{\text{真实为正例的数量}} = \frac{TP}{TP + FN}$
- 假正例率 $FPR = \frac{\text{预测为正例且真实为反例的数量}}{\text{真实为反例的数量}} = \frac{FP}{FP + TN}$
- 查准率 $P = \frac{\text{预测为正例且真实为正例的数量}}{\text{预测为正例的数量}} = \frac{TP}{TP + FP}$
- 查全率 $R = \frac{\text{预测为正例且真实为正例的数量}}{\text{真实为正例的数量}} = \frac{TP}{TP + FN}$

由上可以得出, 真正例率 TPR 等于查全率 R, 假正例率 FPR 和查准率 P 没有什么直接的数学关系。