

Perf-1 (周志华书) 数据集中包含 1000 个样本, 其中 500 个是正样本, 500 个负样本, 将其划分为 70% 的训练集和 30% 的测试集, 请问总计有多少种不同的划分方式。

1. 共 1000 个样本, 其中 500 正样本, 500 负样本, 即分布完全均衡。
现要从这 1000 个样本中选 700 个作训练集, 300 个作测试集。
为了保持样本分布的一致性, 则应从正负样本中各选 350 个构成训练集, 各选 150 个构成测试集。
∴ 划分方式共有 $C_{500}^{350} \times C_{500}^{350}$ 种。

Perf-2 请给出 ROC 曲线总是过 (0,0) 和 (1,1) 的理由。

2. ROC 曲线横轴是 FPR, 纵轴是 TPR, 当阈值 > 所有预测概率的最大值时, 所有样本被预测为负类, 此时 $FP=0$, $TP=0$, 导致 $FPR = \frac{FP}{FP+TN} = 0$, $TPR = \frac{TP}{TP+FN} = 0$, 此时对应点为 (0,0); 当阈值 < 所有预测概率的最小值时, 所有样本被分为正类, 此时 $FN=0$, $TN=0$, 导致 $FPR = \frac{FP}{FP+TN} = 1$, $TPR = \frac{TP}{TP+FN} = 1$, 此时对应点为 (1,1), 因此 ROC 曲线必过 (0,0) 和 (1,1)。

Perf-3 (样本不平衡情况) 数据集中包含 1000 个样本, 其中 5 个是正样本 (有病样本), 995 个负样本 (无病样本), 有一个学习机, 其 Accuracy 是 99.5%, 你认为这个学习机性能好吗 (能够用于疾病筛查吗)? 你认为用什么性能评价能够反映出这个学习机在疾病筛查上的筛查性能?

3. 该学习机的性能不好, 不能用于疾病筛查。
对于一个有 995 个负样本和 5 个正样本的数据集, 该学习机的分类 accuracy 为 99.5%, 即说明它明显倾向于预测一个样本为负样本, 这在现实的疾病筛查中是不可取的, 因为它很可能把有病样本分类成无病样本。
我认为可以用召回率去衡量。 $Recall = \frac{TP}{TP+FN}$, 即真阳样本占所有阳性样本的比例。对于上述分类器, 其 $Recall = \frac{0}{0+5} = 0$, 说明该分类器查全率为 0, 对于检测有病样本的能力很差, 符合我们的预期。

Perf-4 一个训练好的二分类器, 对于 10 个类标分别为 +++-+---- 的测试样本 (样本 1, 样本 2, ...), 若分类结果得到的样本属于正样本的概率由大到小的排序为: 样本 1, 样本 2, 样本 3, ..., 而决策门限为排在前 k 的样本判为正类样本, $k=0,1,2,\dots$, 请画出这个二分类器的 ROC 曲线, 给出其 AUC 数值。

4. ① $k=0$ 时:

true: + + + - + - + - - -

pred: - - - - - - - - - -

此时 $TP=0$, $FP=0 \Rightarrow FPR=TPR=0$, 对应 $(0,0)$

② $k=1$ 时:

true: + + + - + - + - - -

pred: + - - - - - - - - -

此时 $TP=1$, $FP=0$, $TN=5$, $FN=4 \Rightarrow FPR=0$, $TPR=\frac{1}{1+4}=0.2$, 对应 $(0,0.2)$

③ $k=2$ 时:

true: + + + - + - + - - -

pred: + + - - - - - - - -

此时 $TP=2$, $FP=0$, $TN=5$, $FN=3 \Rightarrow FPR=0$, $TPR=\frac{2}{2+3}=0.4$, 对应 $(0,0.4)$

④ $k=3$ 时:

true: + + + - + - + - - -

pred: + + + - - - - - - -

此时 $TP=3$, $FP=0$, $TN=5$, $FN=2 \Rightarrow FPR=0$, $TPR=\frac{3}{3+2}=0.6$, 对应 $(0,0.6)$

⑤ $k=4$ 时:

true: + + + - + - + - - -

pred: + + + + - - - - - -

此时 $TP=3$, $FP=1$, $TN=4$, $FN=2 \Rightarrow FPR=0.2$, $TPR=0.6$, 对应 $(0.2,0.6)$

⑥ $k=5$ 时:

true: + + + - + - + - - -

pred: + + + + + - - - - -

此时 $TP=4$, $FP=1$, $TN=4$, $FN=1 \Rightarrow FPR=0.2$, $TPR=0.8$, 对应 $(0.2,0.8)$

⑦ $k=6$ 时:

true: + + + - + - + - - -

pred: + + + + + + - - - -

此时 $TP=4$, $FP=2$, $TN=3$, $FN=1 \Rightarrow FPR=0.4$, $TPR=0.8$, 对应 $(0.4,0.8)$

⑧ $k=7$ 时:

true: + + + - + - + - - -

pred: + + + + + + + - - -

此时 $TP=5$, $FP=2$, $TN=3$, $FN=0 \Rightarrow FPR=0.4$, $TPR=1$, 对应 $(0.4,1)$

⑨ $k=8$ 时:

true: + + + - + - + - - -

pred: + + + + + + + + - -

此时 $TP=5$, $FP=3$, $TN=2$, $FN=0 \Rightarrow FPR=0.6$, $TPR=1$, 对应 $(0.6,1)$

⑩ $k=9$ 时:

true: + + + - + - + - - -

pred: + + + + + + + + -

此时 $TP=5$, $FP=4$, $TN=1$, $FN=0 \Rightarrow FPR=0.83$, $TPR=1$, 对应 $(0.83, 1)$

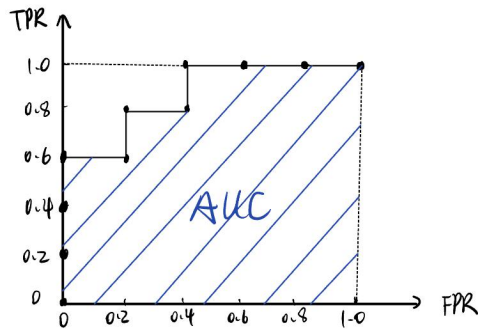
⑪ $k=10$ 时:

true: + + + - + - + - - -

pred: + + + + + + + + +

此时 $TP=5$, $FP=5$, $TN=0$, $FN=0 \Rightarrow FPR=1$, $TPR=1$, 对应 $(1, 1)$

综上, ROC 曲线为:



$$AUC = 1.0 - 0.2 \times 0.2 \times 3 = 0.88$$

Perf-5 一个分类器的测试性能由如下的 confusion matrix 给出:

	Predicted Positives	Predicted Negatives
True Positives	75	25
True Negatives	15	85

请计算这个分类器的如下性能: Accuracy, Error rate; Precision, Recall; F1.

$$5. \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{75 + 85}{75 + 85 + 15 + 25} = 0.8$$

$$\text{Error rate} = 1 - \text{Accuracy} = 0.2$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{75}{75 + 15} = 0.83$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{75}{75 + 25} = 0.75$$

$$F1\text{-score} = \frac{2 \cdot P \cdot R}{P + R} = \frac{2 \times \frac{5}{6} \times \frac{3}{4}}{\frac{5}{6} + \frac{3}{4}} = 0.789$$

Perf-5 (bias and variance) 什么叫 biase-variance dilemma? 学习的目标 (泛化能力),

反映在 bias 和 variance 上, 应该是怎样的?

5. bias-variance dilemma: 若要降低在训练集上的偏差(bias), 则须增加模型的复杂度, 但这易导致过拟合, 使泛化能力下降, 即高方差(variance)。因此, 我们是无法同时使 bias 和 variance 降低的, 也就导致了 bias-variance dilemma。
- 泛化能力反映在 bias 和 variance 上是 bias 较高而 variance 较低。

Perf-6 (选) (System debugging) 在学习出的学习机性能不满足要求情况下, 是认为学习机的结构不够合理而需要改变学习机的结构 (调整超参及增加或者删减特征) 还是认为训练学习机的样本量不足而需要等待新数据的收集, 你应该依据什么来进行选择?

- b. ① 绘制 training error 和 validation error 随超参数变化的曲线图。若在当前超参数下这两个 error 都较大, 则需调整超参数使 validation error 达到最小值。
- ② 绘制 training error 和 validation error 随训练集大小变化的曲线图。若发现随着训练集增大, validation error 下降得很慢, 甚至与 training error 共同趋于某值, 则说明增大训练集是没什么用的。若发现当训练集增大时, validation error 显著下降, 且 validation error 和 training error 之间有较大 "gap", 则说明增大训练集是有帮助的。