

---

---

---

---

---



## CH2.

1. 各从正负 150 个

因此有  $C_{500}^{150} \times C_{500}^{150}$  种方法

2. 10 折交叉验证：正反例样本一样多，因此错误率相同。

备注：若留下为正，那么训练样本中反例数大于正例数，则预测为反；反之预测为正，错误率为 100%。

3. F1 是 P 和 R 的调和平均

BEP 是 P 和 R 相等时的取值

无法判断两者差异。

4.  $TPR = \frac{TP}{TP+FN}$  真实正例被预测为正例的比率

$FPR = \frac{FP}{FP+TN}$  真实反例被预测为正例的比率

5.  $l_{rank}$  中,  $f(x^+) < f(x^-)$  对应扫描到反例, 其参数代入向上延伸的单值表, 而  $f(x^+) = f(x^-)$  对应扫描到正例, 斜向上延伸, 因此前面乘以  $\frac{1}{2}$ , 再除以  $m^+m^-$  为一代, 正好对应了 ROC 曲线上方向而移, 因此  $AUC = 1 - l_{rank}$ .

$$6. E = \frac{(m^+ FNR + m^- FPR)}{m^+ + m^-}$$

$\curvearrowright$  正类被预测成反类       $\curvearrowright$  反类被预测成正类

7. ROC 曲线上的点为  $(FPR, TPR)$ . 对于代价平面的一条直线  $FNR = 1 - TPR$ , 即可在代价平面上绘制一条从  $(0, FPR)$  到  $(1, FNR)$  的线段, 线段下的面积即表示了该条件下的期望总体代价.

8. min-max规范化计算简单, 新增数据计算简单  
但对极端元素敏感.  
 $z$ -score 计算复杂, 新增数据需要全部重新计算.  
对极端数据不敏感.

## 9. ① 提出假设

② 计算  $\chi^2$  值.

③ 计算拒绝域

④ 根据拒绝域决定是接受还是不接受假设.

CH3

1. 从任意两个样本作差作为新的训练集，这样就能将  $y = w^T x + b$  的  $b$  除去。

2. 多元函数为凸  $\Leftrightarrow$  Hessian 矩阵半正定

$$y = \frac{1}{1 + e^{-(w^T x + b)}}$$

$$\frac{dy}{dw^T} = -\frac{e^{-(w^T x + b)} \cdot (-x)}{(1 + e^{-(w^T x + b)})^2} = x \left(1 - \frac{1}{1 + e^{-(w^T x + b)}}\right) \frac{1}{1 + e^{-(w^T x + b)}}$$
$$= x(1-y)y$$

$$\frac{d}{dw^T} \frac{dy}{dw^T} = x(1-y) \cdot x(1-y)y = x^T y (1-y)(1-y).$$

$$\therefore y \in (0, 1)$$

$$\because y \in (0, 1) \text{ 且 } y(1-y)(1-y) < 0.$$

$$\therefore \boxed{y \in (0, 1)}$$

$$l(p) = \sum_{i=1}^m (-y_i p^T x_i + \ln(1 + e^{p^T x_i})), \quad p \in (0, 1).$$

$$\frac{d}{p^T} \left( \frac{dl}{dp} \right) = x x^T p_1(x, p)(1-p_1(x, p)) \quad \text{且} \quad p_1 \geq 0$$

3.  
4. } R github.

5.

6. 可以映射到更高维度上使之线性可分

7. 对于 EOC = 线码，当码长为  $2^n$ ，至少可以使  $2^n$  个类别达到  
最优间隔，它们的海明距离为  $2^{(n-1)}$

1	1	1	1	-1	-1	-1	-1
1	1	-1	-1	1	1	-1	-1
-1	-1	1	-1	1	-1	1	-1
-1	-1	-1	-1	1	1	1	1
-1	-1	1	1	-1	-1	1	1
-1	1	-1	1	-1	1	-1	1

从中选四个，再任意添加一位码长

3.9. 由于对每个类进行了相同的处理，因此类别不平衡问题  
会抵消