机器学习笔记–逻辑回归

空修菜

1 罗辑回归（Logistic Regression）

1. 逻辑回归处理二分类问题 **binary Classification**, 也就是 $y \in \{0, 1\}$.
2. 几率 (odds) 分布函数, 也称为 sigmoid 函数. $y \in \{0, 1\}$, 令

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}},$$

$z \in \mathbb{R}$, "sigmoid" or "logistic function" g can be expressed as

$$g(z) = \frac{1}{1 + e^{-z}}.$$

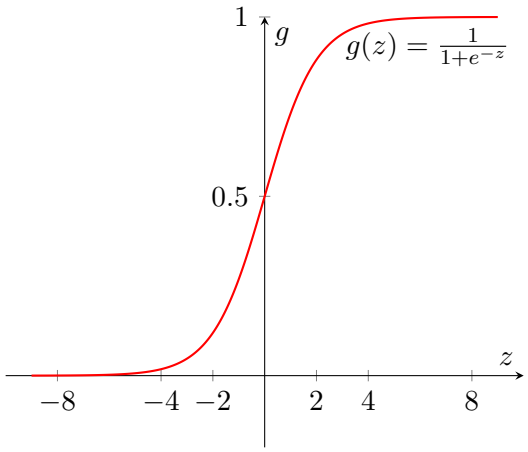

Figure 1.1: 函数 $g(z) \in (0, 1)$

3. sigmodi 函数的与它的导数的关系. $h(\theta, x) \in (0, 1)$, 所以 $h(\theta, x)$ 是一个概率分布函数, 它给出的是每个点的概率. 关于 g 的导数和 g 的关系, 有

$$g'(z) = g(z)(1 - g(z)).$$

4. 类别的条件概率. Since $y = 0$ or $y = 1$,let $p(y = 1 \mid x; \theta) = h_\theta(x)$, thus $p(y = 0 \mid x; \theta) = 1 - h_\theta(x)$. We summarize these expression as follows: $y \in \{0, 1\}$,

$$p(y \mid x; \theta) = h_\theta(x)^y (1 - h_\theta(x))^{1-y}. \tag{1.1}$$

上式表示的是一个 x 是 0 还是 1 的概率, $X = (x^{(1)}, \cdots, x^{(m)})$, $y = (y^{(1)}, \cdots, y^{(m)})$, 由独立同分布以及乘法原理有,

$$\mathcal{L}(\theta) = p(y \mid X; \theta) = \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) = \prod_{i=1}^{m} h_\theta(x^{(i)})^{y^{(i)}} (1-h_\theta(x^{(i)}))^{1-y^{(i)}},$$

5. 对数几率. 对 $\mathcal{L}(\theta)$ 取对数可得

$$\ell(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})).$$

6. 梯度升上法. 选择 θ 使得 $\ell(\theta)$ 取得最大值. 可以使用梯度上升法 **Batch gradient ascent** 是因为函数 $\log \mathcal{L}(\theta)$ 的凹性, 凹性使得最大值存在.

$$\theta_j := \theta_j + \alpha \frac{\partial}{\partial \theta_j} \ell(\theta),$$

thus

$$\theta_j := \theta_j + \alpha \Big(\underbrace{\sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) x_j^{(i)}}_{\frac{\partial}{\partial \theta_j} \ell(\theta)} \Big).$$

7. 分类模型评价指标.

   (1). 在对某个问题进行二分类时, 将主要关注的类别视为正类 (positive), 将另一类与正类相对的视为负类 (negative).

   (2). 将本就是正类的样本点归为正类, 记此时的正类个数为 TP(true positive).

   (3). 将本就是正类的样本点归为负类, 记此时被错误归类的个数为 FN(false negativate).

   (4). 将本就是负类的样本点归为负类, 记此时的负类个数为 TN(true negativate).

   (5). 将本就是负类的样本点归为正类, 记此时的负类个数为 FP(false positive).

(6). 精确率 (precision) 就是所得预测结果中, 被正确预测的正类占预测结果正类的比例.

$$P = \frac{TP}{TP + FP}.$$

(7). 召回率 (recall) 就是所得预测结果中, 被正确预测的正类占实际正类的比例. 假设原来有 10 个正类, 预测所得的正确的正类个数为 6, 则召回率为 $6/10 = 0.6$,

$$R = \frac{TP}{TP + FN}.$$

(8). 将两者结合到一起就得到 F_1,

$$F_1 = \frac{TP}{TP + \frac{FN+FP}{2}} = \frac{2TP}{2TP + FN + FP}.$$

(9). ROC 曲线是真正率 TPR 与假正率 FPR 关系的曲线, 可以理解为每增加一个假正类可以增加几个真正类;

(10). AUC 是 ROC 曲线与假正类轴之间的面积, 值在 0 与 1 之间, 值越大分类效果越好, 对角线表示分类器的分类效果是纯随机的;