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

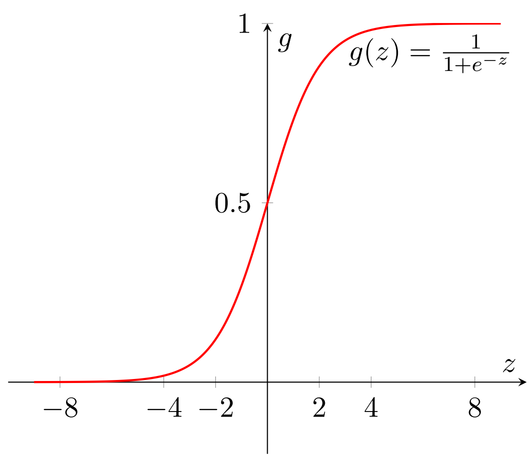

Figure 1.1: 函数 $g(z) \in (0, 1)$

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

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

7. 矩阵表示. 设 $X=(x^{(1)},\cdots,x^{(m)})$, $y=(y^{(1)},\cdots,y^{(m)})$, 其中 $x^{(i)}=(x_0^{(i)},\cdots,x_n^{(i)})$. $\theta=(\theta_1,\cdots,\theta_n)$, θ 的迭代公式可以写为:

$$\begin{aligned}(\theta_1,\cdots,\theta_n):&=(\theta_1,\cdots,\theta_n)+\alpha\Big(y^{(1)}-h_\theta(x^{(1)}),\cdots,y^{(n)}-h_\theta(x^{(n)})\Big)\begin{pmatrix}x^{(1)}\\ \vdots\\ x^{(m)}\end{pmatrix}\\ &=\theta+\alpha\Big(y^{(1)}-h_\theta(x^{(1)}),\cdots,y^{(n)}-h_\theta(x^{(n)})\Big)\begin{pmatrix}x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)}\\ x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)}\\ \vdots & \vdots & \cdots & \vdots\\ x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)}\end{pmatrix}\end{aligned}$$

8. 分类模型评价指标.

(1). 对二分类问题, 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). 精确率 (precision) 是被正确预测的正类占预测的所有正类的比例. 精确率的分母是预测结果中正类的个数.

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

(3). 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. 精确率的分母是预测开始前所有正类的个数.

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

(4). 将两者结合到一起就得到 F_1,

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

- 一个好的分类模型是正确率和召回率都比较高的. 一方高一方低的模型不是好模型;
- 将所有样本点都归为正类, 则召回率为 100%, 同时, 也把负类预测为了正类, 因此精确率就低, F_1 是一个综合分反映指标.

(5). ROC 曲线.

- 真正类率 TPR 就是召回率 (Recall).
- 假正率 FPR 是预测结果中, 被错误归为正类的负类个数占实际负类个数的比,

$$FPR = \frac{FP}{TN + FP}.$$

- 对同一个模型, 选择不同的阈值产生不同的 TPR 和 FPR 对, 将这些数对标记到平面上就得到了 ROC 曲线.
- 曲线越靠近 TPR 的轴, 模型对分类效果就越好.

(6). 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;