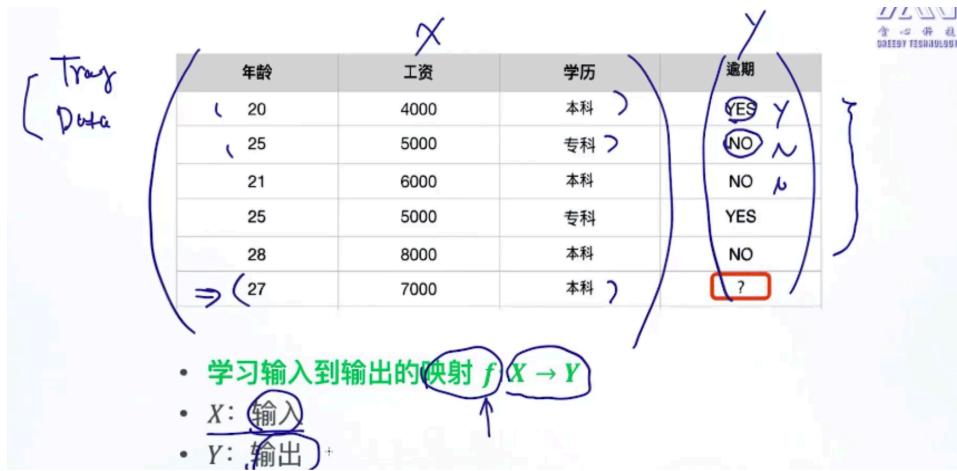


1. Logistic Regression

1.1 Classification tasks:

贷款违约、广告点击、商品推荐、情感分析、疾病诊断, etc

逻辑回归是一个很好的 baseline



1.2 problem

学习输入到输出的映射 $f: X \rightarrow Y$

X : 输入

Y : 输出

- 定义条件概率: $P(Y|X) = ?$ $\Leftrightarrow (20, 4000, \text{本科}) (Y=)$
 $P(Y= | (20, 4000, \text{本科})) \uparrow$
- 假设我们明确知道条件概率 $P(Y|X)$, 怎么做分类?
 $P(Y=|x) \geq ? P(Y=|x) \text{ if } P(Y=|x) > P(Y=|x)$

核心: 怎么去设计条件概率

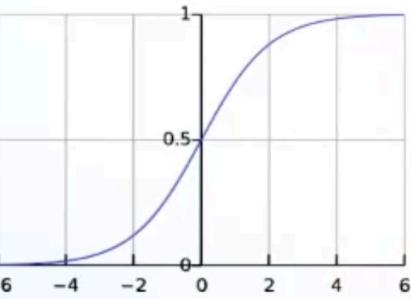
- 可不可以用线性回归来表示 $P(Y|X) = w^T x + b$? 为什么? No!

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

① $0 \leq P(y|x) \leq 1$ \otimes

② $\sum_y P(y|x) = 1$

1.3 logistic function



$$y = \frac{1}{1 + e^{-x}}$$

$x: (-\infty, +\infty)$

$y: (0, 1)$

逻辑函数 $y = \frac{1}{1+e^{-x}}$

原始条件概率 : $P(Y|X) = w^T x + b$

新条件概率 $P(Y|X) = \frac{1}{1+e^{-(w^T x + b)}}$

$$\omega = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix}, b \in \mathbb{R}$$

对于二分类问题：

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

$$p(y=0|x, w) = \frac{e^{-w^T x + b}}{1 + e^{-w^T x + b}} = 1 - p(y=1|x, w)$$

两个式子可以合并成：

$$p(y|x, w) = p(y=1|x, w)^y [1 - p(y=1|x, w)]^{1-y}$$

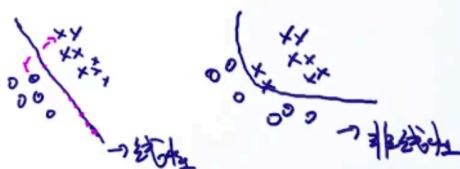
$$\text{if } y=1 \text{ 时. } P(Y=1|x, w) = p(y=1|x, w)$$

1.4 linear classifier

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

$$p(y=0|x, w) = \frac{e^{-w^T x + b}}{1 + e^{-w^T x + b}}$$

决策边界 (Decision Boundary)



1.5 objective function

假设我们拥有数据集 $D = \{(x_i, y_i)\}_{i=1}^n \quad x_i \in R^d, \quad y_i \in \{0, 1\}$

而且我们已经定义了：

$$p(y|x, w) = p(y=1|x, w)^y [1 - p(y=1|x, w)]^{1-y}$$

我们需要最大化目标函数：

$$\hat{w}_{MLE}, \quad \hat{b}_{MLE} = \operatorname{argmax}_w \prod_{i=1}^n p(y_i|x_i, w, b)$$

$$\begin{aligned} \hat{w}_{MLE}, \quad \hat{b}_{MLE} &= \operatorname{argmax}_{w,b} \prod_{i=1}^n p(y_i|x_i, w, b) \quad \textcircled{1} \\ &= \operatorname{argmax}_{w,b} \log \left(\prod_{i=1}^n p(y_i|x_i, w, b) \right) \\ &= \operatorname{argmax}_{w,b} \sum_{i=1}^n \log p(y_i|x_i, w, b) \end{aligned}$$

$\log(xyz) = \log x + \log y + \log z$

Objective Function

$$\begin{aligned} \operatorname{argmax}_{w,b} &\sum_{i=1}^n \log p(y_i|x_i, w, b) \\ \operatorname{argmin}_{w,b} &- \sum_{i=1}^n \log p(y_i|x_i, w, b) \\ &= \operatorname{argmin}_{w,b} - \sum_{i=1}^n \log \left[p(y_i|x_i, w, b)^y \cdot [1 - p(y_i|x_i, w, b)]^{1-y} \right] \\ &= \operatorname{argmin}_{w,b} - \sum_{i=1}^n y \cdot \log p(y_i|x_i, w, b) + (1-y) \log [1 - p(y_i|x_i, w, b)] \end{aligned}$$

$\log(a^y \cdot b^x) = \log(a^y) + \log(b^x) = y \log a + x \log b$

2. 梯度下降

2.1 求解函数的最小值

求使得 $f(w)$ 值最小的参数 w

- 是否凸函数
- 最优化算法

1) $f(w) = 0 \Rightarrow w = ?$

2) Iterative

- GD (Gradient Descent)
- SGD (Stochastic GD)

Global Optimal vs Local Optimal

Non-Convex (非凸)

2.2 梯度下降

求使得 $f(w)$ 值最小的参数 w

初始化 w^1
for $t = 1, 2, \dots$:
 $w^{t+1} = w^t - \eta \nabla f(w^t)$

Gradient Descent

求使得 $f(w)$ 值最小的参数 w

初始化 w^1
for $t = 1, 2, \dots$:
 $w^{t+1} = w^t - \eta \nabla f(w^t)$

例子：求解函数 $f(w) = 4w^2 + 5w + 1$ 的最优解

$\eta = -\frac{b}{2a} = -\frac{5}{8}$

$w^1 = 0, f(w) = 8w + 5, \eta = 0.1$
 $w^2 = w^1 - \eta \cdot (8 \cdot 0 + 5) = 0 - 0.5 = -0.5$
 $w^3 = w^2 - \eta \cdot (8 \cdot (-0.5) + 5) = -0.5 - 0.5 = -1.0$
 $w^4 = w^3 - \eta \cdot (8 \cdot (-1.0) + 5) = -1.0 - 0.5 = -1.5$
 $w^5 = w^4 - \eta \cdot (8 \cdot (-1.5) + 5) = -1.5 - 0.5 = -2.0$

0.625 左右所以我们通过这种计算的方法

2.3 逻辑回归的梯度下降

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

$$\operatorname{argmin}_{w,b} - \sum_{i=1}^n y_i \log p(y_i=1|x_i, w) + (1-y_i) \log(1-p(y_i=1|x_i, w))$$

$$\begin{aligned}
 &= \operatorname{argmin}_{w,b} - \sum_{i=1}^n y_i \cdot \log \sigma(w^T x_i + b) + (1-y_i) \cdot \log [1 - \sigma(w^T x_i + b)] \\
 \boxed{\frac{\partial L(w,b)}{\partial w}} &= - \sum_{i=1}^n y_i \cdot \frac{\sigma(w^T x_i + b) \cdot [1 - \sigma(w^T x_i + b)]}{\sigma(w^T x_i + b)} \cdot x_i + (1-y_i) \cdot \frac{\sigma(w^T x_i + b) \cdot [1 - \sigma(w^T x_i + b)]}{1 - \sigma(w^T x_i + b)} \cdot x_i \\
 &= - \sum_{i=1}^n y_i \cdot (1 - \sigma(w^T x_i + b)) + (1-y_i) \cdot \sigma(w^T x_i + b) \cdot x_i \\
 &\Rightarrow \sum_{i=1}^n [y_i \cdot \sigma(w^T x_i + b)] \cdot x_i = \sum_{i=1}^n [\sigma(w^T x_i + b) - y_i] \cdot x_i
 \end{aligned}$$

$$\begin{aligned}
 L(w,b) &= - \sum_{i=1}^n y_i \cdot \sigma(w^T x_i + b) + (1-y_i) \log [1 - \sigma(w^T x_i + b)] \\
 \boxed{\frac{\partial L(w,b)}{\partial b}} &= - \sum_{i=1}^n y_i \cdot \frac{\sigma(w^T x_i + b) \cdot [1 - \sigma(w^T x_i + b)]}{\sigma(w^T x_i + b)} + (1-y_i) \cdot \frac{-\sigma(w^T x_i + b) \cdot [1 - \sigma(w^T x_i + b)]}{1 - \sigma(w^T x_i + b)} \\
 &= - \sum_{i=1}^n y_i \cdot (1 - \sigma(w^T x_i + b)) + (1-y_i) \cdot (-\sigma(w^T x_i + b)) \\
 &= \sum_{i=1}^n [\sigma(w^T x_i + b) - y_i]
 \end{aligned}$$

初始化 w^0, b^0
for $t=1, 2, \dots$ learning rate
 $w^{t+1} = w^t - \eta \cdot \sum_{i=1}^n [\sigma(w^T x_i + b^t) - y_i] \cdot x_i$
 $b^{t+1} = b^t - \eta \cdot \sum_{i=1}^n [\sigma(w^T x_i + b^t) - y_i]$

Gradient Descent



Stochastic Gradient Descent for Logistic Regression

