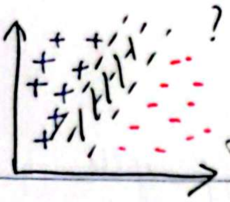


fig1

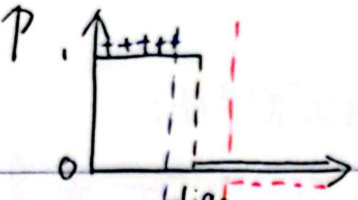


fig2

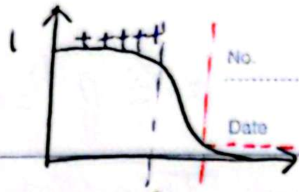


fig3

Logistic Regression

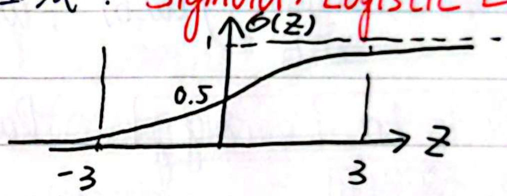
Recall: Perceptron 是简单二分模型，但它无法处理不可分的数据；同时，Perceptron 无法建模 uncertainty：如fig1，超平面划分有很多种，但若有点在"\"区域呢？Perceptron 被迫给出±1，但直观上，给出概率可能更合理！

如fig2&fig3；同样如fig中的sample point，在红蓝中间处，perceptron 只能硬性找一个地方然后作分界；但感性上，希望如fig3这样建模

用怎样的函数能很好建模fig3曲线？Sigmoid / Logistic 函数

$$\sigma(z) = \frac{1}{1+\exp(-z)}$$

fig4:

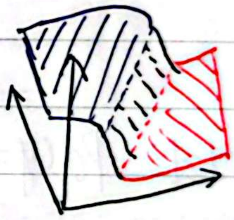


$\Downarrow$

$$g(x) = \sigma(w^T x + b) = \frac{1}{1+\exp\{-(w^T x + b)\}}$$

如何预测: predict +1 if probability > 0.5

$\Rightarrow \sigma(w^T x + b) > 0.5$

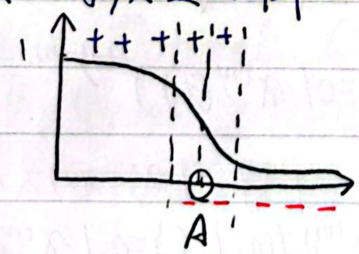


$$\Rightarrow \frac{1}{1+\exp\{-(w^T x + b)\}} > 0.5 \Rightarrow \exp\{-(w^T x + b)\} < 1$$

$\Rightarrow w^T x + b > 0$，可见依然是要训w.b参数，预测时，$w^T x + b > 0 \Rightarrow +1$

但！Sigmoid 建模能引入 Uncertainty，并且在数据不可分时，建模仍有质量保障。*

*



虽然A点处有两种 sample point 但因为给出的是 probabity distribution 因此有 quality guarantees

How do we learn a classifier?

*: 令 $g^{(i)}=\sigma(\omega^T x^{(i)}+b)$

$$p(\text{data}) = \prod_i p(\text{data point } i) = \prod_i \begin{cases} g^{(i)}, & \text{if } y^{(i)}=+1 \\ 1-g^{(i)}, & \text{else} \end{cases}$$

$$= \prod_i \left\{ (g^{(i)})^{1\{y^{(i)}=+1\}} (1-g^{(i)})^{1\{y^{(i)}\neq+1\}} \right.$$

$$\Rightarrow Loss = \frac{1}{n}\sum_{i=1}^{n} -\left(1\{y^{(i)}=+1\}\log g^{(i)} + 1\{y^{(i)}\neq+1\}\log(1-g^{(i)})\right)$$

取负对数 (negative log likelihood loss) (g for guess, a for actual)

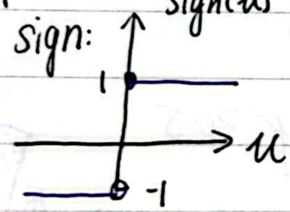
$$-L_{nll}(g,a) = 1\{a=+1\}\log(g) + 1\{a\neq+1\}\log(1-g)$$

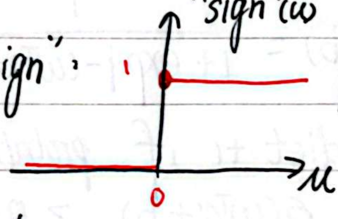
目标便是: $J(\omega,b) = \frac{1}{n}\sum_{i=1}^{n} L_{nll}(\sigma(\omega^T x^{(i)}+b), y^{(i)})$

找 $\underset{\omega,b}{\operatorname{argmin}} J(\omega,b)$, where: $-L_{nll}(g,a) = 1\{a=+1\}\log g + 1\{a\neq+1\}\log(1-g)$

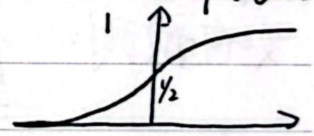
$\boxed{\text{Back to Learning}}$: Data: $D=\{\vec{x}^{(i)}, y^{(i)}\}_{i=1}^{N}$, $\vec{x}\in\mathbb{R}^M$, $y\in\{0,1\}$.

原先预测用:

sign: 

"sign": 

先在: sigmoid: $P_\theta(y=1|\vec{x}) = \frac{1}{1+\exp(-\theta^T\vec{x})}$



Model: $P_\theta(y=1|x) = \frac{1}{1+\exp(-\theta^T x)} = g(x) \Rightarrow P(y|x,\theta) = \begin{cases} g(x) & \text{if } y=1 \\ 1-g(x) & \text{if } y=0 \end{cases}$

Learning: objective:

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} -\log p(y^{(i)}|x^{(i)},\theta)$$

欲 $\theta^* = \underset{\theta}{\operatorname{argmin}}\, \ell(\theta)$, $\ell(\theta) = -\frac{1}{N}\log\prod_{n=1}^{N} P(y^{(n)}|\vec{x}^{(n)},\theta)$

$$= -\frac{1}{N}\log\prod_{n=1}^{N} P(Y=1|x^{(n)},\theta)^{y^{(n)}} (P(y=0|x^{(n)},\theta))^{1-y^{(n)}}$$ *:

不同.

*: 若 $y^{(n)}=1$, 欲 $P(Y=1|x^{(n)},\theta)$大; 反之, 欲 $P(y=0|x^{(n)},\theta)$大; 对称

$$= -\frac{1}{N}\sum_{n=1}^{N} y^{(n)}\log P(Y=1|x^{(n)},\theta) + (1-y^{(n)})\log P(Y=0|x^{(n)},\theta)$$

$$= -\frac{1}{N}\sum_{n=1}^{N} y^{(n)}\theta^T x^{(n)} - \log(1+\exp(\theta^T x^{(n)}))$$

No.

Date

Gradient: Target: $J(\theta) = -\frac{1}{N}\sum_{n=1}^{N} y^{(n)}\theta^T x^{(n)} - \log(1+\exp(\theta^T x^{(n)}))$

$$\nabla_\theta J(\theta) = -\frac{1}{N}\sum_{n=1}^{N}\left[y^{(n)}\nabla_\theta(\theta^T x^{(n)}) - \nabla_\theta \log(1+\exp(\theta^T x^{(n)}))\right]$$

$$= -\frac{1}{N}\sum_{n=1}^{N}\left[y^{(n)}x^{(n)} - \frac{\exp(\theta^T x^{(n)})}{1+\exp(\theta^T x^{(n)})}x^{(n)}\right] \rightarrow \frac{1}{1+\exp(-\theta^T x^{(n)})}$$

$$= \frac{1}{N}\sum_{n=1}^{N} x^{(n)}\left(P(Y=1|x^{(n)},\theta) - y^{(n)}\right)$$

Then for learning with Gradients given. we can GD/SGD

☆: No close form solution! (无closed form解 for MLE参数)

附: 旧 LMS (Least Mean Square), $h_\theta(x) = \theta^T x$,

则 $J_\theta = \frac{1}{n}\sum_{i=1}^{n}(\theta^T x^{(i)} - y^{(i)})^2$, $\boxed{\text{stochastically}}$; (针对一个sample)

$\nabla_\theta J(\theta)^{(i)*} = 2(\theta^T x^{(i)} - y^{(i)})x^{(i)}$, 则 $\theta_k \leftarrow \theta_k + \nabla_\theta J(\theta)$

$= \theta x + \lambda(h_\theta(x^{(i)}) - y^{(i)})$; $\nabla_\theta J(\theta)^{(i)*} = x^{(i)}\cdot(P(Y=1|x^{(n)},\theta) - y^{(i)})$

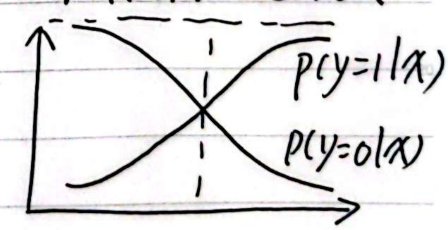
发现 Linear Regression 的 SGD 也是: $\theta_k \leftarrow \theta_k + \lambda(h_\theta(x^{(i)}) - y^{(i)})$

*△: 再次recall: SGD: for $i \in$ shuffle($\{1,2,\cdots,N\}$) do:

$\theta \leftarrow \theta - \gamma\nabla_\theta J^{(i)}(\theta)$ ☆ 注意这里是'-', 且 $\gamma > 0$!

可见, 之前的 $\theta_k \leftarrow \theta_k + \lambda(h_\theta(x^{(i)}) - y^{(i)})$ 中的 λ 应是负数

Bayes Optimal Classifier:



若 $p^*(y|x)$ 这个先验信息被告知呢?

如: 测核酸时. $P(阳) << P(阴)$

则: $l(y,\hat{y}) = \begin{cases} 1000000, & y \neq \hat{y}, y = 阳 \\ 1000, & y \neq \hat{y}, y = 阴 \\ 0, & y = \hat{y} \end{cases}$

No.

Date

则可考虑：Bayes Decision Rule：$\hat{y} = h(x) = \begin{cases} 1, & \text{if } P^*(y=1|x) \geq \alpha \\ 0, & \text{otherwise} \end{cases}$

若 $\alpha = 0.5$，则 $\ell(y, \hat{y}) = \mathbb{1}(y \neq \hat{y})$

这样在 $\alpha \neq 0.5$ 下，可以设计 asymmetric loss

带这种思想，能用于修改 Log-Loss：

原：$J(\theta) = \frac{-1}{n} \sum_{i=1}^{n} y^{(i)} \log P(Y=1|x^{(i)}, \theta) + (1-y^{(i)}) \log P(Y=0|x^{(i)}; \theta)$

如例中，若 $y=1$，但 $P(Y=1|x^{(i)}; \theta)$ 很小，则 Loss 更大

则赋以权重：$J(\theta) = \frac{-1}{n} \sum_{i=1}^{n} w_1 y^{(i)} \log P(Y=1|x^{(i)}, \theta) + w_2 (1-y^{(i)}) \log P(Y=0|x^{(i)}; \theta)$

$w_1/w_2 > 1$，则 $y=1$ 预测错的后果更严重！i.e., $\nabla_\theta J(\theta) \uparrow \Delta\theta \uparrow$