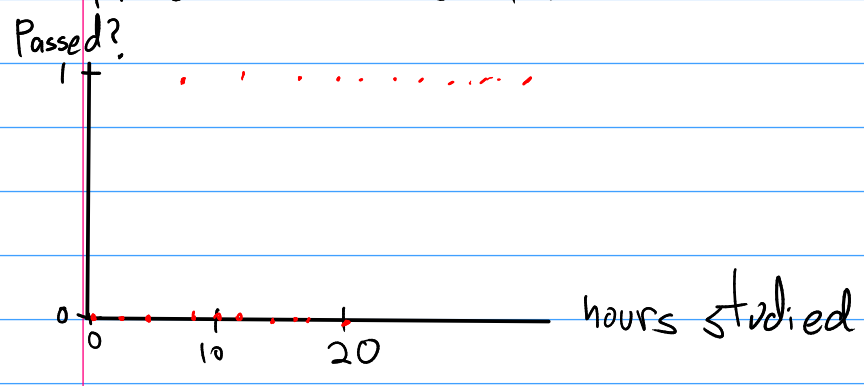
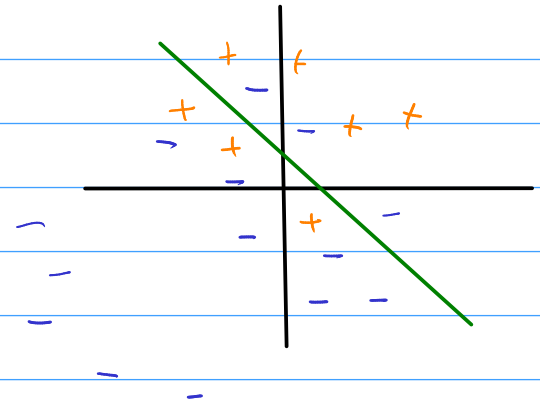
Logistic Regression

Predictions with Confidence





Bernoulli Model

$\Pr[Y=1|X,w] = p(X,w) \in [0,1]$

$\Pr[Y=0|X,w] = 1-p(X,w)$

Take 1: $p(X,w) = \langle X,w \rangle$

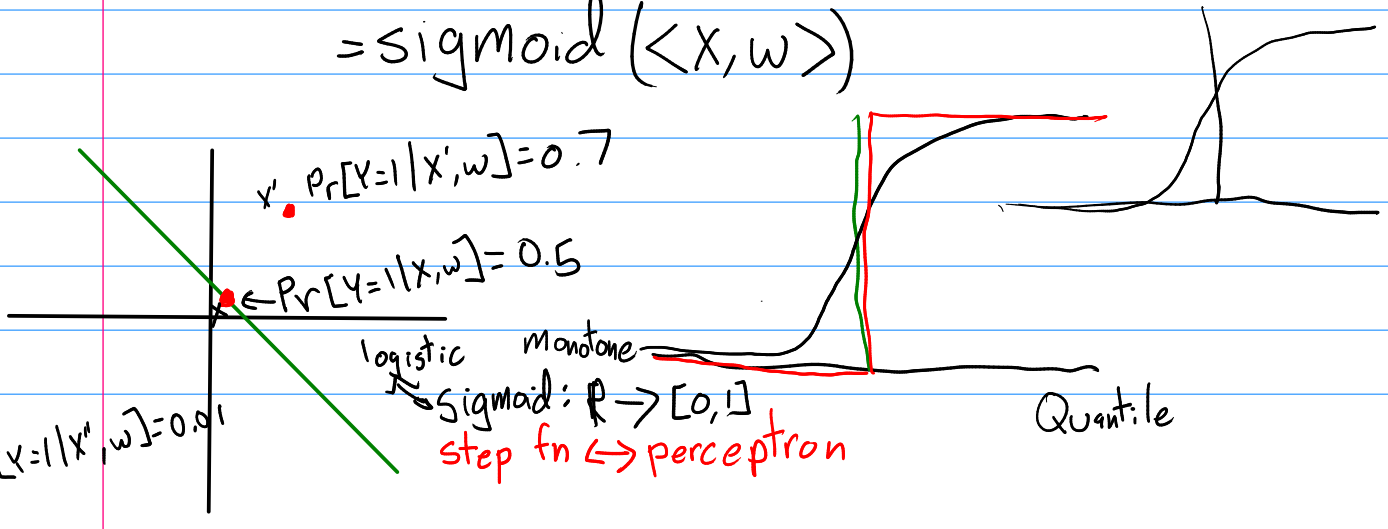
$$\log\left(\frac{p(X,w)}{1-p(X,w)}\right) = \langle X,w \rangle \Rightarrow \frac{p(X,w)}{1-p(X,w)} = \exp(\langle X,w \rangle)$$

$$p(X,w) = \exp(\langle X,w \rangle)(1-p(X,w))$$

$$\Rightarrow p(X,w)(1+\exp(\langle X,w \rangle)) = \exp(\langle X,w \rangle)$$

$$p(X,w) = \frac{\exp(\langle X,w \rangle)}{1+\exp(\langle X,w \rangle)} = \frac{1}{1+\exp(-\langle X,w \rangle)}$$

$$= \text{sigmoid}(\langle X,w \rangle)$$



$$p(X,w) = \frac{1}{1+\exp(-\langle w, X\rangle)}$$

If $p(X,w) > \frac{1}{2}$, predict $\hat{y} = 1$. Else $\hat{y} = 0$.

$\hat{y} = \text{sign}(\langle w, X\rangle)$

Maximum Likelihood Estimation

$$\hat{w} = \arg\max_w \prod_{i=1}^{n} p(x_i, w)^{y_i} \left(1 - p(x_i, w)\right)^{(1-y_i)}$$

$$p(x_i, w) = \frac{1}{1 + \exp(-\langle w, x_i \rangle)} = p_i$$

$$\arg\max_w \sum \log\left(p_i^{y_i} (1-p_i)^{(1-y_i)}\right)$$

$$= \arg\max_w \sum y_i \log p_i + (1-y_i) \log(1-p_i)$$

Suppose $y_i = 1$. $\log p_i = \log\left((1 + \exp(-\langle w, x_i \rangle))^{-1}\right)$

$$= -\log(1 + \exp(-\langle w, x_i \rangle))$$

$y_i = 0$. $\log(1-p_i) = \log\left(\frac{\exp(-\langle w, x_i \rangle)}{1 + \exp(-\langle w, x_i \rangle)}\right) = -\log(1 + \exp(\langle w x_i \rangle$

$$= \arg\max_w \sum \left(-\log\left(\exp(-y_i \langle w, x_i \rangle) + \exp((1-y_i)\langle w, x_i \rangle)\right)\right)$$

$$= \arg\min_w \frac{1}{n} \sum \log\left(\exp(-y_i \langle w, x_i \rangle) + \exp((1-y_i)\langle w, x_i \rangle)\right)$$

$$\tilde{y}_i = \begin{cases} +1 & \text{if } y_i = 1 \\ -1 & \text{if } y_i = 0 \end{cases}$$

$$= \arg\min_w \frac{1}{n} \sum \overbrace{\log\left(1 + \exp(-\tilde{y}_i \langle w, x_i \rangle)\right)}^{l_w(x_i, y_i)}$$

$$\arg\min_w \mathop{E}_{x,y \sim D} [l_w(x, y)] \approx \arg\min_w \frac{1}{n} \sum l_w(x_i, y_i)$$

$$= \arg\min_w \frac{1}{n} \sum \log(1 + \exp(-\tilde{y}_i \langle w, x_i \rangle))$$

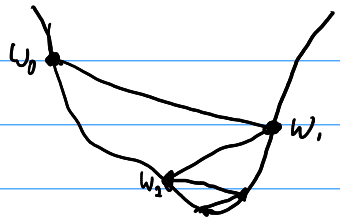
Claim: $\ell_w(x_i, y_i)$ is convex

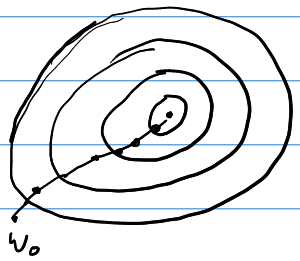
Goal: Find $\hat{w}$ where $\frac{1}{n}\sum \nabla_w \ell(x_i, y_i) = 0$

Claim: $\nabla_w \ell_w(x_i, y_i) = (p(x_i, w) - y_i)x_i$

No closed form for $\hat{w}$!

Optimization





Initialize w_0

For $t = 1, 2, \ldots,$

Choose: direction d_t

step size η_t

$w_t = w_{t-1} - \eta_t d_t$

How to pick: - η_t: constant, decaying $\left(\frac{1}{\sqrt{t}}\right)$, adaptive.

Gradient Descent (GD): $d_t = \frac{1}{n}\hat{\sum_{i=1}}\nabla_w \ell_{w_{t-1}}(x_i, y_i)$ (=0 at opt).

Stochastic GD (SGD): Pick set $B \subseteq [n]$ randomly

$d_t = \frac{1}{|B|}\sum_{i \in B} \nabla_w \ell_{w_{t-1}}(x_i, y_i)$

Newton's Method: $d_t = \underbrace{\left(\frac{1}{n}\sum \nabla_w^2 \ell_{w_{t-1}}(x_i, y_i)\right)^{-1}}_{\text{Inverse Hessian}} \underbrace{\left(\frac{1}{n}\sum \nabla_w \ell_{w_{t-1}}(x_i, y_i)\right)}_{\text{Gradient}}$

Multiclass

$$Pr[Y=k \mid X, W] = \frac{\exp(\langle w_k, X \rangle)}{\sum_{l=1}^{c} \exp(\langle w_l, X \rangle)}$$