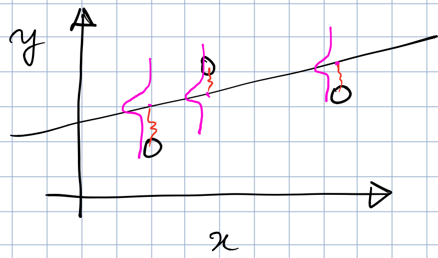
– Probabilistic view of linear regression



min (—)

OR

Assume $y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$

$\uparrow$ models the uncaptured effects or noise.

$$\varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$$

$$P(\varepsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{\varepsilon^{(i)2}}{2\sigma^2}\right)$$

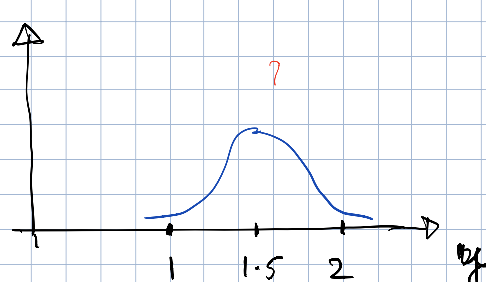
$$P(y^{(i)} \mid x^{(i)}; \theta, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

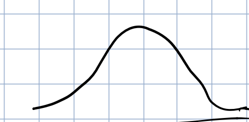
$$y^{(i)} \mid x^{(i)}; \theta, \sigma^2 \sim \mathcal{N}(\theta^T x^{(i)}, \sigma^2)$$

Likelihood

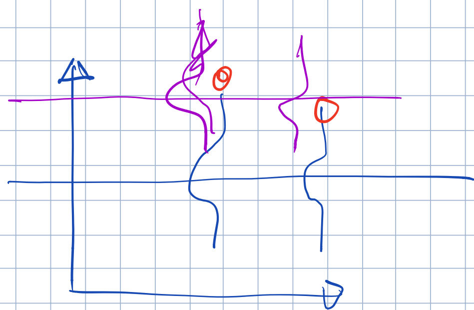
$y_1 = 1, \quad y_2 = 2, \quad y_3 = 1.5$

$y_i \sim \mathcal{N}(\mu, 1)$





$$\mathcal{L}(\mu) = P(1 \mid \mu)\, P(1.5 \mid \mu)\, P(2 \mid \mu)$$



Write down the likelihood. (I.I.D.)

$$L(\theta) = \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta, \sigma^2)$$

$$= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

Goal: choose θ to maximize $L(\theta)$.

Log-likelihood.

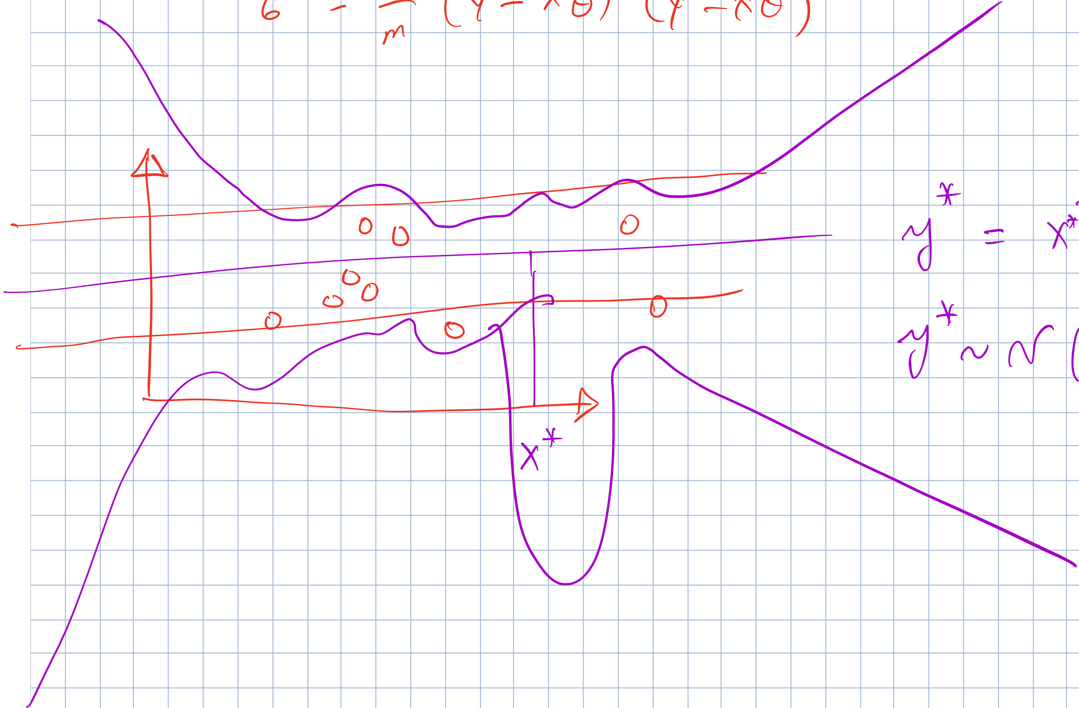
$$\left.\begin{aligned} l(\theta) &= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \\ &= m \log \frac{1}{\sqrt{2\pi}\,\sigma} + \sum_{i=1}^{m} \left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \end{aligned}\right] \max$$

$$\text{neg } l(\theta) = -m \log \frac{1}{\sqrt{2\pi}\,\sigma} + \boxed{\frac{1}{2\sigma^2}\sum_{i=1}^{m} \left(y^{(i)} - \theta^T x^{(i)}\right)^2} \Bigg] \min$$

$$\frac{\partial \text{ neg } l(\theta, \sigma)}{\partial \theta} = 0 \quad \text{and} \quad \frac{\partial \text{ neg } l(\theta, \sigma)}{\partial \sigma} = 0$$

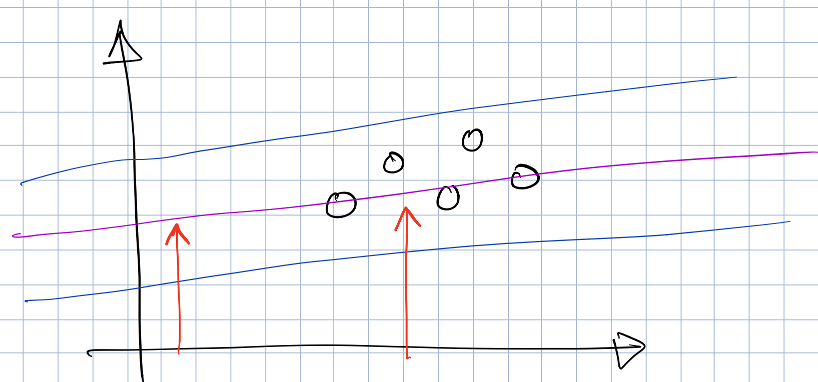
$$\theta_{ML} = (X^T X)^{-1} X^T Y$$

$$\sigma^2 = \frac{1}{m}(y - X\theta)^T(y - X\theta)$$

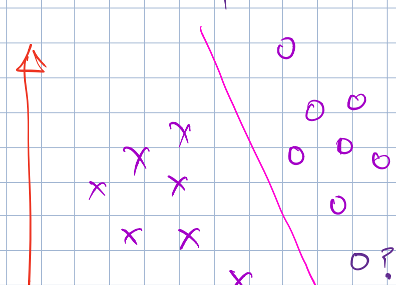


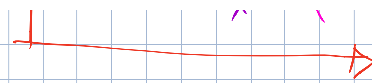
$$y^* = x^{*T}\theta$$

$$y^* \sim \mathcal{N}(x^{*T}\theta, \sigma^2)$$



Logistic Regression (classification problem)

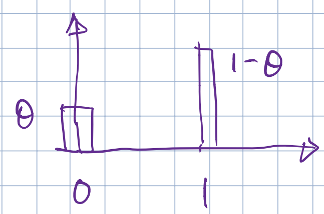




① Bernoulli Random Variable:

x takes values in $\{0, 1\}$

$$p(x|\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases}$$



$$p(x|\theta) = \theta^{x}(1-\theta)^{1-x}$$

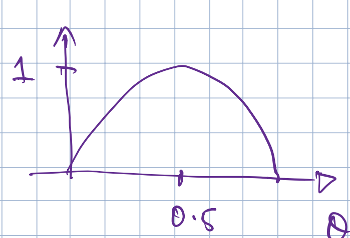
② Entropy. measure of uncertainity associated with a random variable.

$$H(x) = -\sum_{x} p(x|\theta) \log p(x|\theta)$$

Entropy of Bernoulli R.V.

$$H(x) = -\sum_{x=0}^{1} \theta^{x}(1-\theta)^{1-x} \log \theta^{x}(1-\theta)^{1-x}$$

$$= -\left[(1-\theta)\log(1-\theta) + \theta \log \theta\right]$$



③ Binary Classification

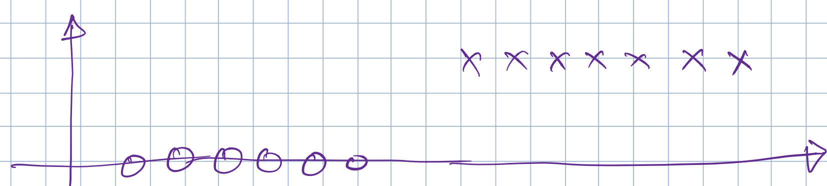
learn a model $h_\theta(x) \longrightarrow$ assign labels $y \in \{0,1\}$

Use Bernoulli distribution for y.

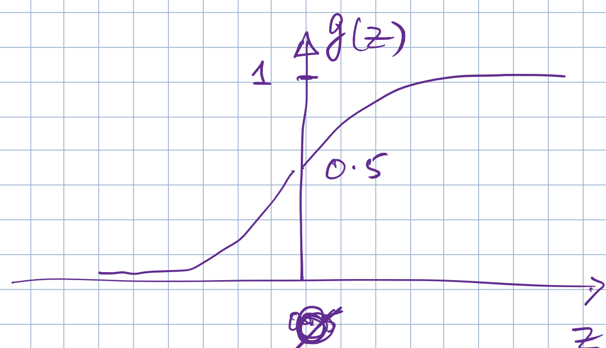
$$Pr(y=1) = h_\theta(x)$$

$$Pr(y=0) = 1 - h_\theta(x)$$

$$Pr(y) = \underline{\underline{h_\theta(x)}}^{y} \left(1 - \underline{\underline{h_\theta(x)}}\right)^{1-y}$$



Sigmoid (logistic) Function



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

Likelihood of target $y^{(i)}$?

$$P(y^{(i)} \mid x; \theta) = h_\theta(x)^{y^{(i)}} \left(1 - h_\theta(x)\right)^{1-y^{(i)}}$$

$$P(y \mid x; \theta) = \prod_{i=1}^{m} h_\theta(x)^{y^{(i)}} \left(1 - h_\theta(x)\right)^{1-y^{(i)}} = L(\theta)$$

$$\theta : \theta - \eta \nabla \left(-\ell(\theta)\right)$$

Generalized Linear Models

$P(y \mid x; \theta)$

$y \in \mathbb{R}$ Gaussian (linear regression)

$y \in \{0,1\}$ Bernoulli (logistic regression)

Gaussian and Bernoulli $\rightarrow$ Exponential Distributions Family.

$$P(y : \eta) = b(y) \exp\left(\eta^T T(y) - a(\eta)\right)$$

η = natural parameter

$T(y)$ = Sufficient statistics

$T(y) = y$ in many cases.

GLM :

Assumptions or design choices.

- $y \mid x; \theta \sim$ Exponential family.
- Given x, my goal is to output $E[T(y) \mid x]$

  i.e., $h_\theta(x) = E[T(y) \mid x]$

- The relationship between η and x is linear.

$$\eta = \theta^T x$$

$$y \mid x ; \theta \sim \text{ExpFamily}(\eta)$$

$$h_\theta(x) = E[y \mid x;\theta] = p(y=1 \mid x;\theta) = \phi.$$

$\text{Ber}(\phi)$

$$\begin{aligned} p(y;\phi) &= \phi^y (1-\phi)^{1-y} \\ &= \exp\left(\log \phi^y (1-\phi)^{1-y}\right) \\ &= \exp\left(y \log \phi - y \log(1-\phi) + \log(1-\phi)\right) \\ &= \exp\Big(\underbrace{y}_{T(y)} \underbrace{\log\left(\frac{\phi}{1-\phi}\right)}_{\eta} + \underbrace{\log(1-\phi)}_{-a(\eta)}\Big) \end{aligned}$$

$$\eta = \log\left(\frac{\phi}{1-\phi}\right)$$

$\downarrow$ invert this fn.

$$\boxed{\phi = \frac{1}{1+e^{-\eta}}}$$

using $a(\eta) = -\log(1-\phi)$

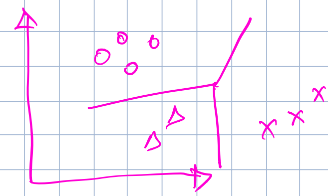
$= -\log(1+e^{\eta})$.

$T(y) = y$

$b(y) = 1$

Multinomial Distribution

$y \in \{1, \dots, k\}$ $\quad$ k possible values



$P(y=i) = \phi_i$

Total # parameters: $\quad \phi_1, \dots, \phi_k \quad \Big| \sum_{i=1}^{k} \phi_i = 1$

$\therefore \quad \phi_1, \dots, \phi_{k-1}, \quad 1 - \sum_{i=1}^{k-1} \phi_i$

$$\begin{cases} \phi_1 & \text{if } y=1 \\ \phi_2 & y=2 \\ \vdots & \\ \phi_k & y=k \end{cases}$$

$$\boxed{\theta^y (1-\theta)^{1-y}}$$

Indicator Variables

$\mathbb{1}\{\text{true}\} = 1$ $\qquad$ $\mathbb{1}\{3=3\} = 1$

$\mathbb{1}\{\text{false}\} = \phi$

$$P(y) = \phi_1^{\mathbb{1}\{y=1\}} \phi_2^{\mathbb{1}\{y=2\}} \cdots \phi_k^{\mathbb{1}\{y=k\}}$$

Goal: Multinomial distribution ~~belong~~ belongs to the Exp Family.

$T(y)$ is a vector.

$$T(1) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad T(2) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \cdots \quad T(k) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

$$= \phi_1^{T(y)_1} \phi_2^{T(y)_2} \cdots \phi_{k-}^{T(y)_{k-1}}$$

$$\Downarrow$$

$$\eta = \begin{bmatrix} \log \phi_1/\phi_k \\ \vdots \\ \log \phi_{k-1}/\phi_k \end{bmatrix}$$

$$a(\eta) = -\log(\phi_k)$$

$$b(y) = 1$$

$$\phi_i = \frac{e^{\eta_i}}{1 + \sum_{j=1}^{k-1} e^{\eta_j}} \qquad i = 1, \cdots, k-1$$

$$= \frac{e^{\theta_i^T x}}{1 + \sum_{j=1}^{k-1} e^{\theta_j^T x}} \qquad \eta_i = \theta_i^T x$$