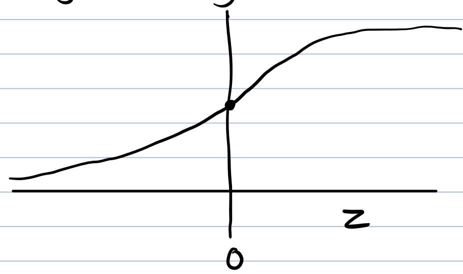
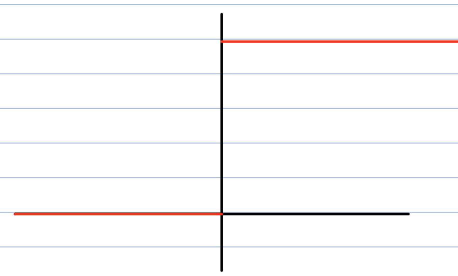
- Perception
- Exponential Family
- Generalized Linear Models
- Softmax Regression
  (Multiclass Classification)

Logistic Regression





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

$$g(z) = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0 \end{cases}$$

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

$$h_\theta(x) = g(\theta^T x)$$

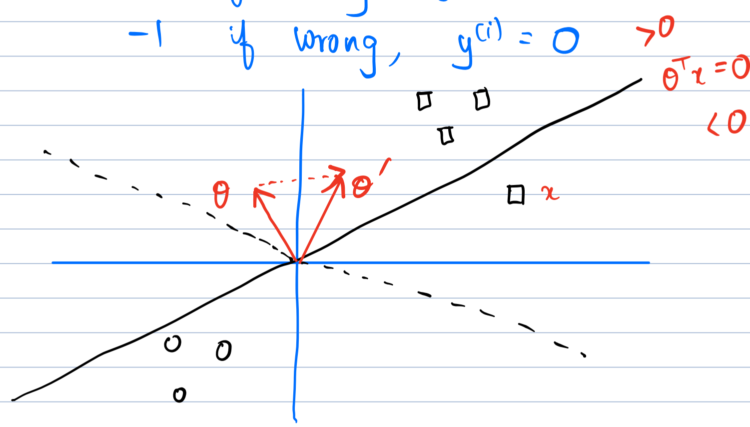
$$\theta_j := \theta_j + \alpha\left(y^{(i)} - \underbrace{h_\theta(x^{(i)})}_{\text{different}}\right) x_j^{(i)}$$

$y^{(i)} - h_\theta(x^{(i)})$

0 : algorithm got it right

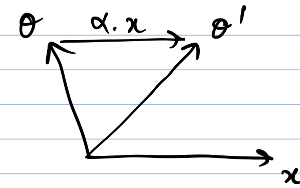
+1/-1 : +1 if wrong, $y^{(i)} = 1$

-1 if wrong, $y^{(i)} = 0$



$$\theta' = \theta + \alpha x$$

$$\theta'^T x = (\theta + \alpha x)^T x$$
$$= \theta^T x + \underbrace{\alpha . x^T x}_{>0}$$



Exponential Families

PDF

$$p(y;\eta) = b(y) \exp[\eta^T T(y) - a(\eta)]$$

y : data

η : natural parameter

$T(y)$: sufficient statistic

: y today

$b(y)$: Base measure

$a(\eta)$: log-partition function

$$p(y;\eta) = \frac{b(y) \exp(\eta^T T(y))}{e^{a(\eta)}}$$

y : scalar

η : vector / scalar

$T(y)$: —— " ——

$b(y)$: scalar

Bernoulli (Binary Data)

ϕ : probability of event

$$p(y;\phi) = \phi^y (1-\phi)^{1-y}$$
$$= \exp(\log(\phi^y (1-\phi)^{(1-y)}))$$
$$= \exp\left[\underbrace{\log\left(\frac{\phi}{1-\phi}\right)}_{\eta} \underbrace{y}_{T(y)} + \underbrace{\log(1-\phi)}_{a(\eta)} \right]$$

$b(y) = 1$

$T(y) = y$

$\eta = \log\left(\frac{\phi}{1-\phi}\right) \Rightarrow \phi = \frac{1}{1+e^{-\eta}}$ (sigmoid)

$a(\eta) = -\log(1-\phi) = -\log\left(1 - \frac{1}{1+e^{-\eta}}\right)$

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

Gaussian (w. fixed variance) Assume $\sigma^2 = 1$

$$p(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2}\right)$$

$$= \underbrace{\frac{1}{\sqrt{2\pi}} e^{-y^2/2}}_{b(y)} \cdot \exp\left(\underbrace{\mu}_{\eta} \underbrace{y}_{T(y)} - \underbrace{\frac{1}{2}\mu^2}_{a(\eta)}\right)$$

$b(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)$

$T(y) = y$

$\eta = \mu$

$a(\eta) = \frac{\mu^2}{2} = \frac{\eta^2}{2}$

Properties

natural parameter

ⓐ MLE w.r.t. η is concave

negative Log Likelihood (NLL) is convex

ⓑ $E[y;\eta] = \frac{\partial}{\partial\eta} a(\eta)$

ⓒ $Var[y;\eta] = \frac{\partial^2}{\partial\eta^2} a(\eta)$

GLM

Assumptions / Design Choices

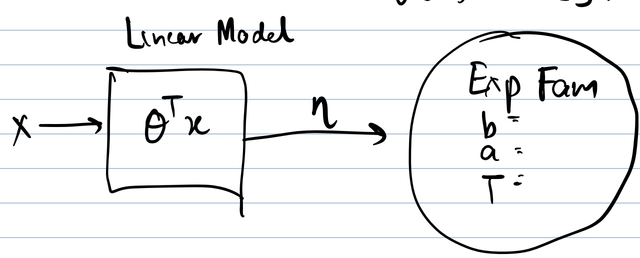
(I) $y | x ; \theta \sim$ Exponential family

Real	—	Gaussian
Binary	—	Bernoulli
Count	—	Poisson
R^+	—	Gamma, Exponential
Dist^n	—	Beta, Dirichlet

(II) $\eta = \theta^T x$ $\quad x = \mathbb{R}^d$, $\theta = \mathbb{R}^d$

(III) Test time: Output $E[y | x ; \theta]$

$h_\theta(x) = E[y | x ; \theta]$



$\rightarrow E[y ; \eta]$

$= E[y ; \theta^T x]$

$= h_\theta(x)$ (Test)

$$\max_\theta \log P(y^{(i)} ; \theta^T x^{(i)}) \qquad \text{(Train)}$$

Learning Update Rule

$$\theta_j := \theta_j + \alpha (y^{(i)} - h_\theta(x^{(i)})) \, x_j^{(i)}$$

plug in appropriate $h_\theta(x)$

Terminology: η : natural parameter

$\mu = E[y ; \eta] = g(\eta)$ $\rightarrow$ canonical response f^n

$\eta = g^{-1}(\mu)$ — " — link f^n

$g(\eta) = \frac{\partial}{\partial \eta} a(\eta)$

3 parametrizations

Model parameters — Natural param — Canonical param

$\theta \xrightarrow{\ \theta^T x\ } \eta \underset{g^{-1}}{\overset{g}{\rightleftarrows}}$ ϕ : Bernoulli, μ, σ^2 : Gaussian, λ : Poisson

$\uparrow$ learning — Design choice

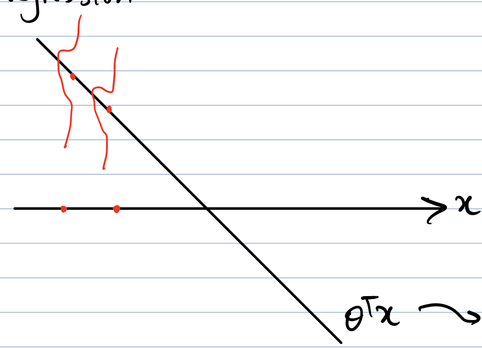
Logistic Regression

$$h_\theta(x) = E[y \mid x ; \theta] = \phi = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-\theta^T x}}$$

Assumptions

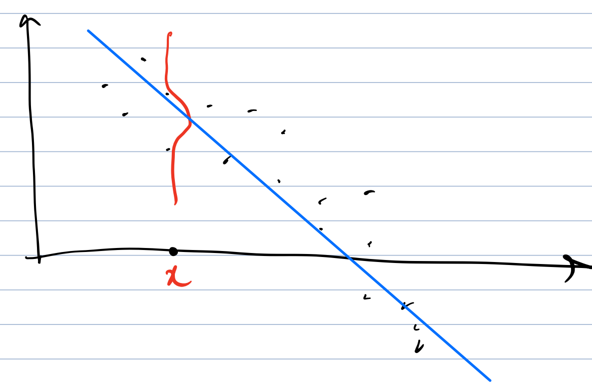
Regression

Assumption: y for any x distributed as gaussian w. mean $\mu = \theta^T x$



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

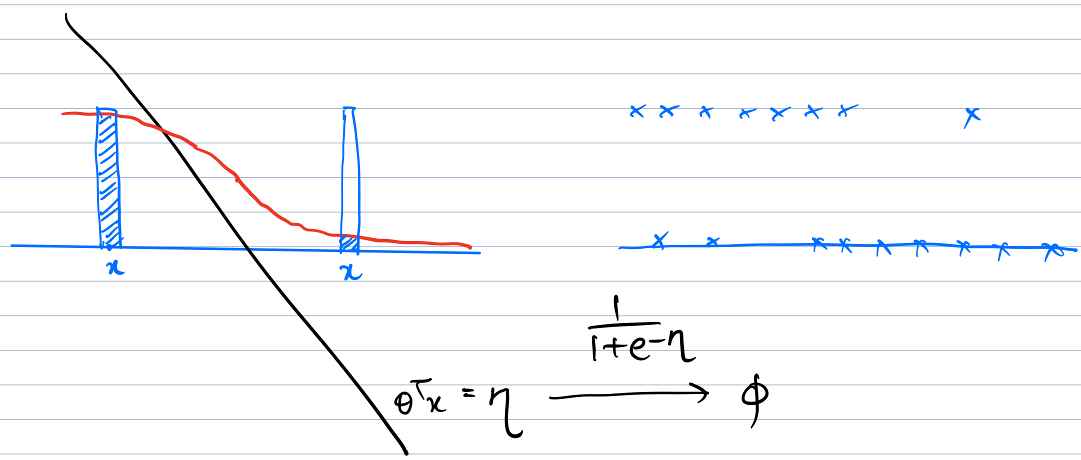
Data generation $\Longrightarrow$



Find θ

$\Longleftarrow$

Classification



$$\theta^T x = \eta \xrightarrow{\frac{1}{1+e^{-\eta}}} \phi$$

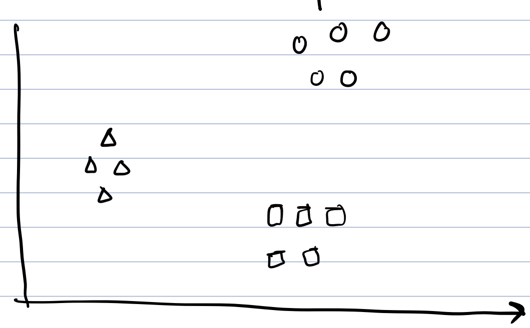
$\Longleftarrow$ θ ?

Softmax Regression

Member of GLM family

Cross Entropy Minimization

Multiclass Classification



$x^{(i)} \in \mathbb{R}^d$

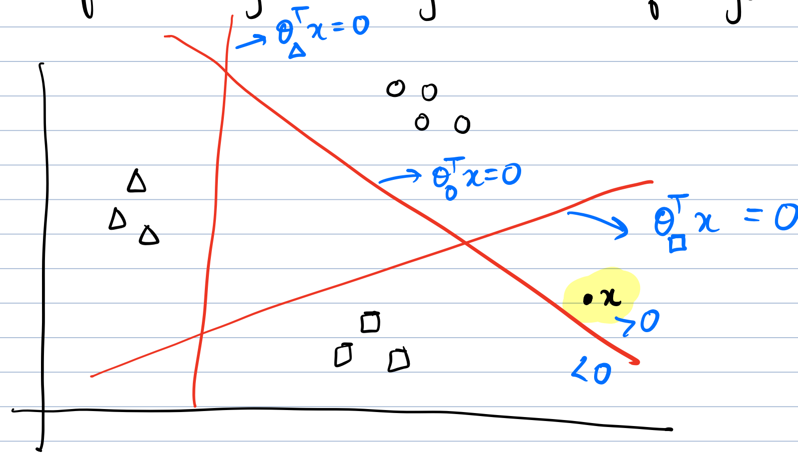
Label $y \in \{0, 1\}^K$ eg. $[0, 0, 1, 0]$ "one hot vector"

$\theta_{class} \in \mathbb{R}^d$

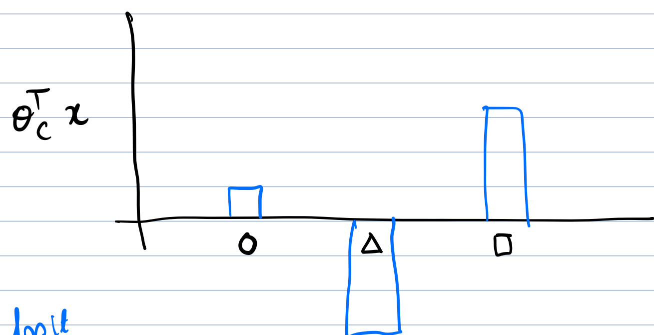
$class \in \{\triangle, o, \square, \ldots\}$

$$
\begin{array}{cc}
 & d \\
k & \begin{bmatrix} \text{—} \; \theta_1 \; \text{—} \\ \text{—} \; \theta_2 \; \text{—} \\ \\ \end{bmatrix}
\end{array}
$$

Softmax regression generalization of logistic regression

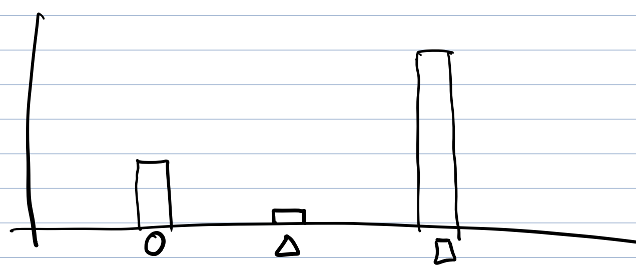


Given x



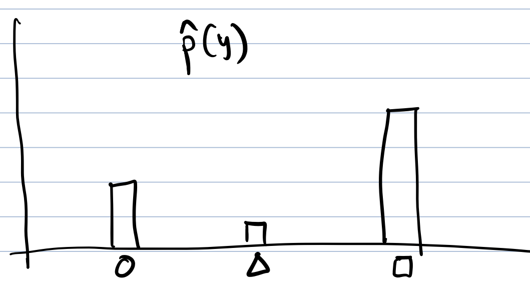
logit space

$\xrightarrow{\text{exp}}$

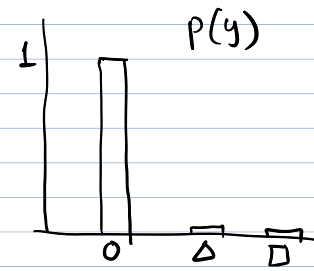


$\xrightarrow{\text{normalize}}$

$$\frac{e^{\theta_i^T x}}{\sum e^{\theta_i^T x}}$$



$\longleftrightarrow$



goal: min distance between distrib^n

min cross entropy

$$\text{Cross Entropy}(p, \hat{p}) = -\sum_{y \in \{o, \triangle, \square\}} p(y) \log \hat{p}(y)$$

$$= -\log \hat{p}(y_o)$$

$$= -\log \frac{e^{\theta_o^T x}}{\sum_{i \in \{\triangle, \square, o\}} e^{\theta_i^T x}}$$

↑

Gradient descent