Topics

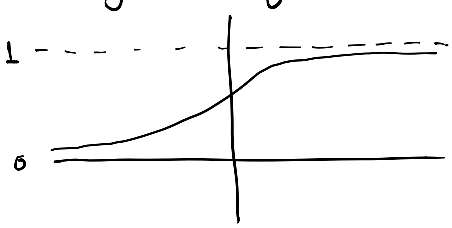
Perceptron

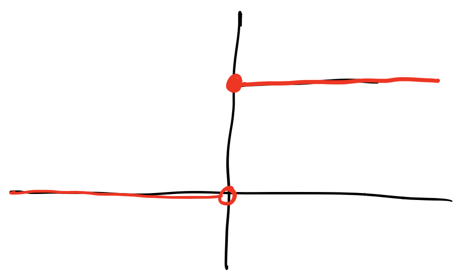
Exponential Family

Generalized Linear Models

Softmax Regression
(Multiclass Classification)

## Logistic Regression



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

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

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

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

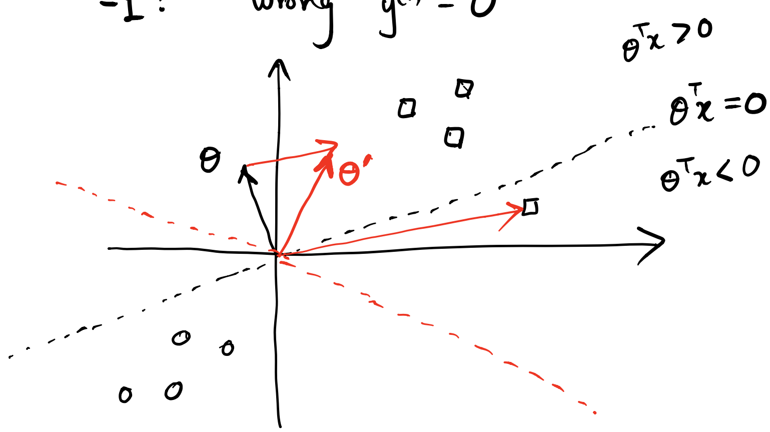
$$\theta_j := \theta_j + \alpha \left( y^{(i)} - \underbrace{h_\theta(x^{(i)})}_{\text{different}} \right) x_j^{(i)}$$

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

0 : algo got it right

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

−1 : if wrong $y^{(i)} = 0$



want $\theta^T x > 0$ but $\theta^T x < 0$

$\theta' = \theta + \alpha x$ $\quad (\theta + \alpha x)^T \cdot x = \theta x + \alpha \cdot x^T x$

Exponential Families

PDF

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

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

$y$ : data

$\eta$ : natural parameter

$T(y)$ : sufficient statistic

$y$ in class

$b(y)$ : Base measure

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

$y$ : scalar

$\eta$ : vector / scalar

$T(y)$ : vector / scalar

match

$b(y)$ : scalar

Examples:

Bernoulli (Binary Data)

$\phi$ : probability of event

$$p(y;\phi) = \phi^y (1-\phi)^{1-y}$$

$$= \exp(\log(\phi^y(1-\phi)^{1-y})) = \exp(y \log \phi + (1-y)\log(1-\phi))$$

$$= \exp\Big[\underbrace{\log\Big(\frac{\phi}{1-\phi}\Big)}_{\eta}\underbrace{y}_{T(y)} + \underbrace{\log(1-\phi)}_{-a(\eta)}\Big]$$

$$b(y) = 1$$
$$T(y) = y$$
$$\eta = \log\left(\frac{\phi}{1-\phi}\right) \Rightarrow \phi = \frac{1}{1+e^{-\eta}} \quad \text{(sigmoid)}$$
$$a(\eta) = -\log(1-\phi) = -\log\left(1 - \frac{1}{1+e^{-\eta}}\right)$$
$$= \log(1+e^{\eta})$$

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

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

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

Properties

natural params

ⓐ MLE w.r.t $\eta$ 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

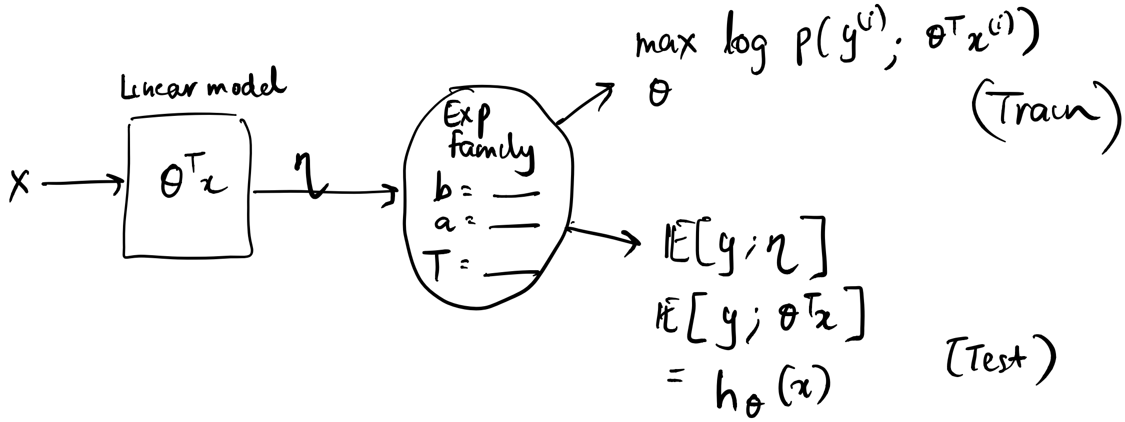
(i) $y \mid x; \theta \sim$ Exponential family

Real — Gaussian
Binary — Bernoulli
Count — Poisson
$\mathbb{R}^+$ — Gamma, Exponential
Dist$^n$ — Beta, Dirichlet (Bayesian ML/stats)

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

(iii) Test time: output $\mathbb{E}[y \mid x; \theta]$

$$\Longrightarrow h_\theta(x) = \mathbb{E}[y \mid x; \theta]$$




## Learning Update Rule

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

## Terminology

$\eta$ : natural parameter

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

$\eta = g^{-1}(\mu)$ canonical link f$^n$

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

3 paramatrizations

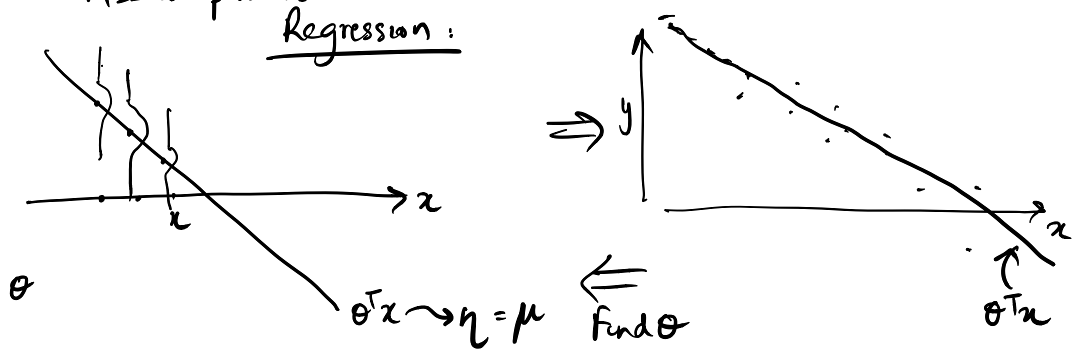
| Model parameters | | Natural param | | Canonical param |
|---|---|---|---|---|
| $\theta$ | $\xrightarrow{\theta^T x}$ | $\eta$ | $\xrightarrow{g}$ / $\xleftarrow{g^{-1}}$ | $\phi$ : Bernoulli; $\mu, \sigma^2$ : Gaussian; $\lambda$ : Poisson |
| $\uparrow$ learning | Design choice | | | |

Logistic Regression

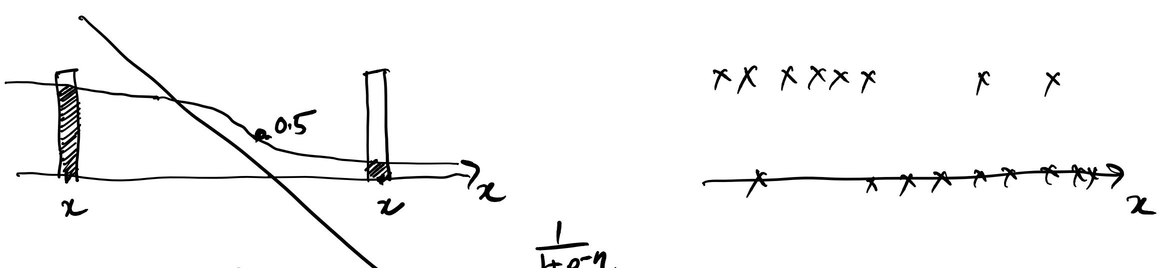
$$h_\theta(x) = \mathbb{E}[y \mid x; \theta] = \phi = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-\theta^T x}}$$

Assumptions:

Regression:




Classification

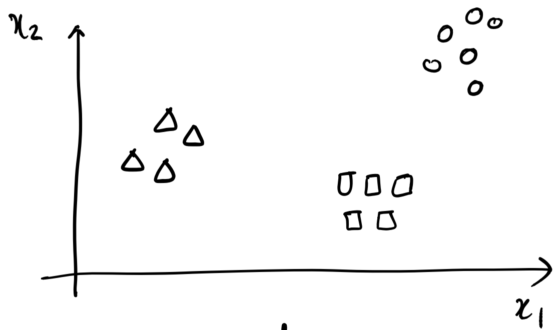



sigmoid($\theta^T x$) $\quad \theta^T x = \eta \xrightarrow{\frac{1}{1+e^{-\eta}}} \phi$

$\Longleftarrow \theta$ ?

# Softmax Regression
(Multiclass Classification)

Cross Entropy minimization

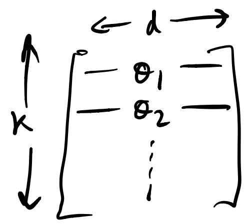


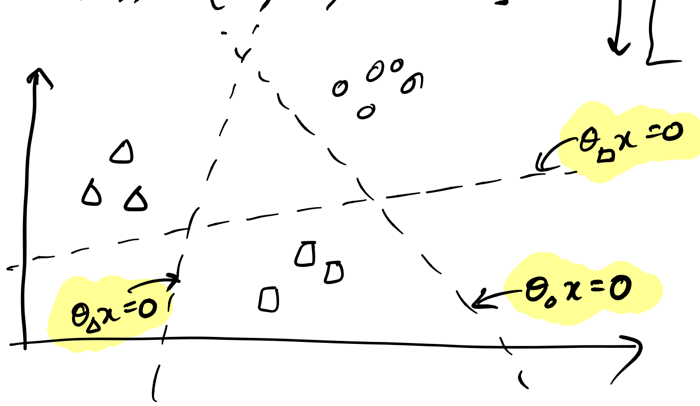

$x^{(i)} \in \mathbb{R}^d$

$y^{(i)} \in \{0,1\}^K$ "one-hot" vector $[0,0,1,0]$

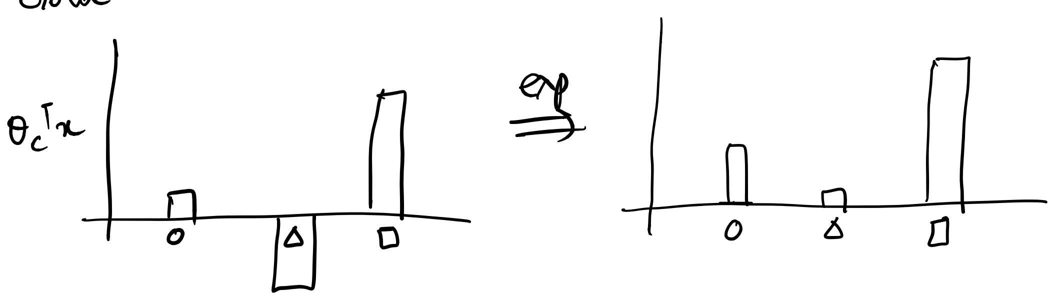
$\theta_{class} \in \mathbb{R}^d$

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



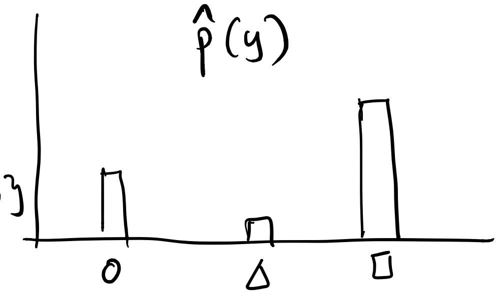




Goal $x$

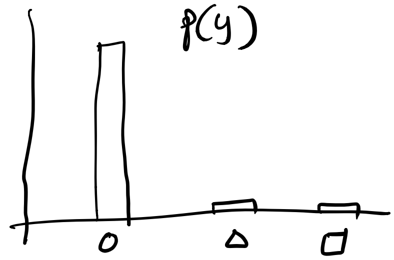



$\longrightarrow$ normalize

$$\frac{e^{\theta_i^T x}}{\sum\limits_{i\in\{\Delta,\square,o\}} e^{\theta_i^T x}}$$




$\longleftrightarrow$




goal: min distance between 2 distrbtn

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

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

$$= -\log \frac{e^{\theta_o^T x}}{\sum\limits_{c\in\{\Delta,\square,o\}} e^{\theta_c^T x}}$$

Find $\theta_o, \theta_\Delta, \theta_\square$ $\uparrow$ Gradient Descent