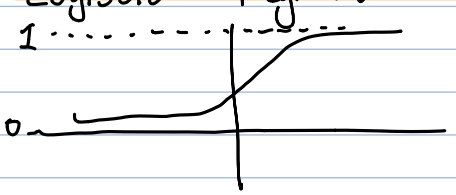
Topics

- Perceptron
- Exponential Family
- Generalized Linear Methods
- Softmax Regression
  (Multiclass classification)

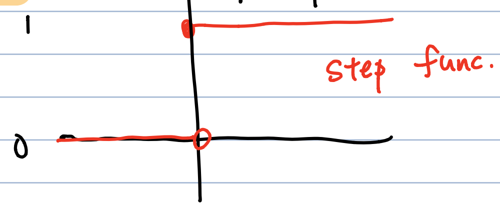
Logistic Regression (Recap)



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

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

Perceptron



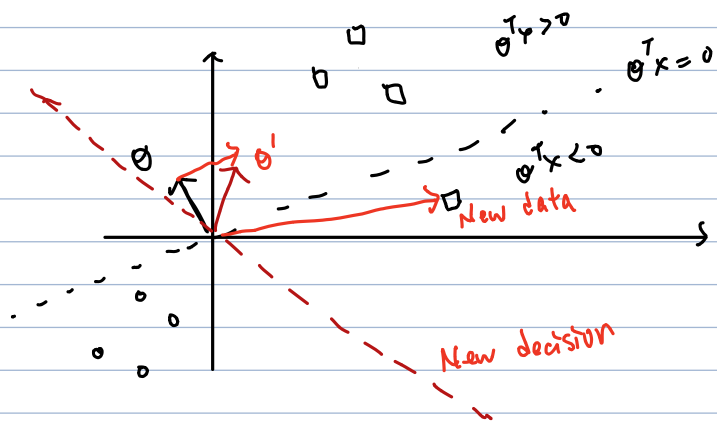
$$g(z) = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0 \end{cases}$$

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

$$\theta_j := \theta_j + \alpha \big(\underbrace{y^{(i)} - h_\theta(x^{(i)})}_{\text{diff.}}\big)\, x_j^{(i)}$$

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

$$\begin{cases} 0: & \text{algo got it right} \\ +1: & \text{wrong} \quad y^{(i)} = 1 \\ -1: & \text{wrong} \quad y^{(i)} = 0. \end{cases}$$



#Note
The order of the data matters.

Exponential Families

PDF (of y para by η)

$$P(y;\eta) = b(y)\exp[\eta^T T(y) - a(\eta)] = \frac{b(y)\exp(\eta^T T(y))}{\underbrace{e^{a(\eta)}}_{\text{Normalizing func.}}}$$

y: data
η: natural parameter
$T(y)$: sufficient statistic
y in class
$b(y)$: Base measure
$a(\eta)$: log-partition function

y: scalar
η: vector / scalar
$T(y)$: vector / scalar
match
$b(y)$: scalar

Examples:

· Bernoulli: Binary Data
ϕ: Probability of event

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

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

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

· Gaussian (w/ fix variance) $\sigma^2 = 1$

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

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

nature param.

Properties

(a) MLE w.r.t. η is concave

negative log likelihood (NLL) is convex

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

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

GLM (Generalized Linear Model)

Assumption / Design Choices

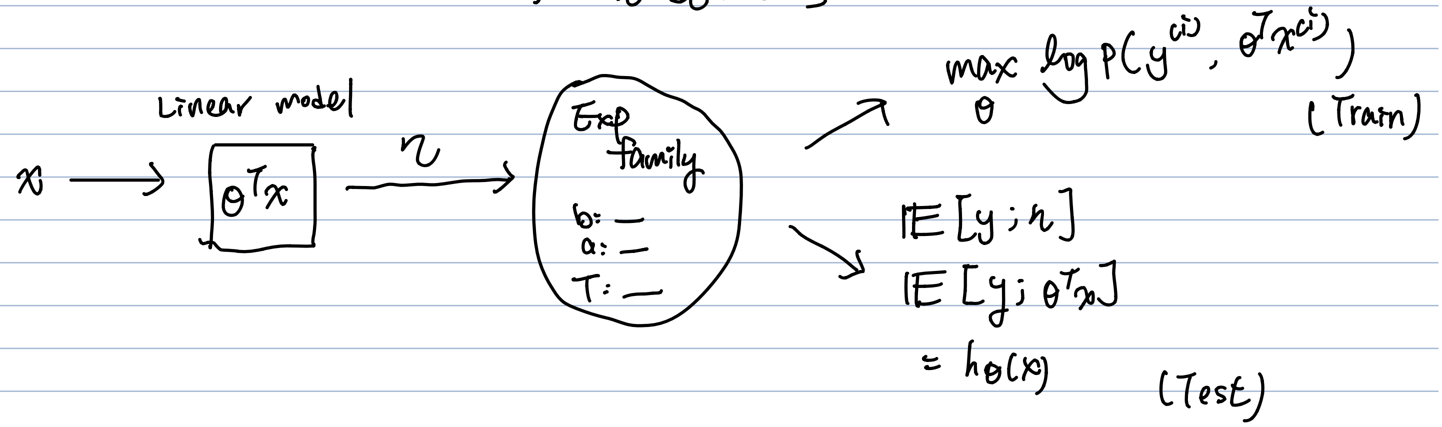
(i) $y \mid x;\theta \sim$ Exponential family

Real	–	Gaussian
Binary	–	Bernoulli
Count	–	Poisson
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 $\mathbb{E}[y \mid x;\theta]$

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



Learning Update Rule.

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

plug-in appropriate h_θ

Terminology

η: natural parameter

μ: $E[y;\eta] = g(\eta)$ canonical response fn.

$\eta = g^{-1}(\eta)$ canonical link fn.

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

3 parameters

Model parameters — Natural param. — Canonical param.

$\theta \xrightarrow{\theta^T x} \eta \underset{g^{-1}}{\overset{g}{\rightleftarrows}}$

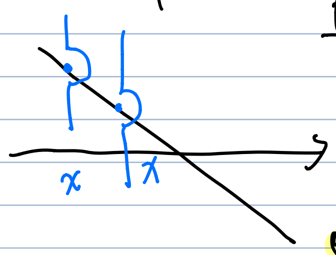
$\uparrow$ learning; $\theta^T x$: Design choice

ϕ: Bernoulli
μ, σ^2: Gaussian
λ: Poisson

Logistic Regression

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

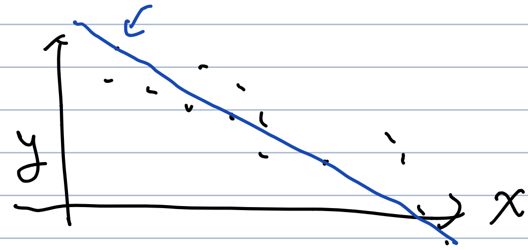
Assumptions: Linear Regression



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

corresponds to $\Rightarrow$

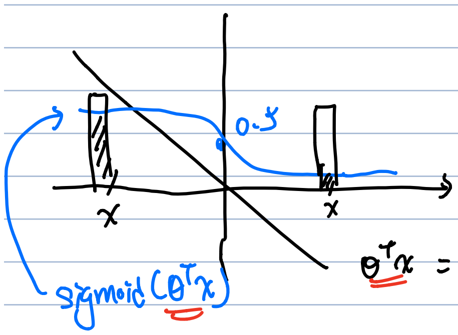
$\Leftarrow$ Find θ



$\forall x$. corresponding y is a gaussian with mean μ.

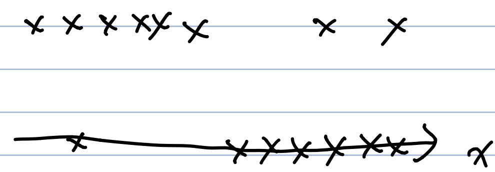
Choose a line, that the means of the distributions lies on the line

Classification



θ? $\Longleftarrow$

Data

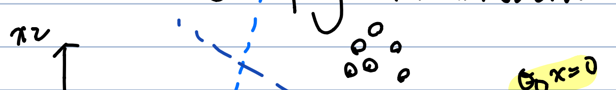


$\theta^T x = \eta \xrightarrow[g]{\frac{1}{1+e^{-\eta}}} \phi$.

e.g. bernoulli(p). $p = \phi$

Softmax Regression
(Multiclass regression)

Cross Entrophy minimization

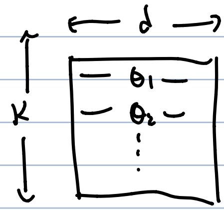


$x^{(i)} \in \mathbb{R}^d$

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

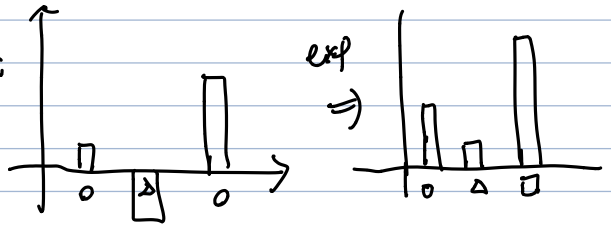
$\theta_{class} \in \mathbb{R}^d$

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



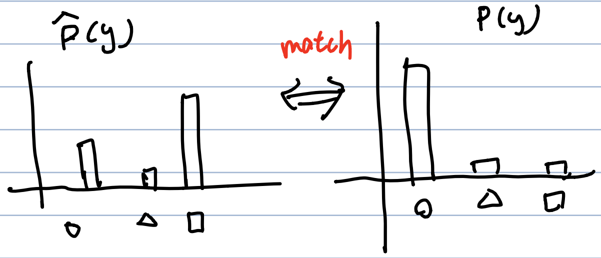
Goal x

$\theta_c^T x$



$\dfrac{e^{\theta_i^T x}}{\sum_{i \in \{\ldots\}} e^{\theta_i^T x}} \Longrightarrow$

$\hat{p}(y)$ match $\Longleftrightarrow$ $p(y)$



goal: min distance between 2 distributions

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

$$= -\log \hat{p}(y_\circ)$$

$$= -\log \frac{e^{\theta_\circ^T x}}{\sum_{c \in \{\}} e^{\theta_c^T x}}$$

Find $\theta_\circ, \theta_\triangle, \theta_\square$ through gradient descent.