Exponential family models

- Definition & MOTIVATION
- Examples
- Softmax (Multiclass Classification

Unify INFERENCE & LEARNING for MANY MODELS

Exponential family

PDF. IDEA "If p has special form $\Rightarrow$ some questions for free"

DATA

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

natural parameters

$T(y)$ is called sufficient statistic (we'll use $T(y) = y$ in class)

is same dim as η

$b(y)$ is called base measure. Does not depend on η

$a(\eta)$ is called log partition function. Does not depend on η

$\Rightarrow$ it makes sure p is probability function

$y, a(\eta), b(y)$ are scalars

$\eta, T(y)$ are same dimension

Examples

Bernoulli $\quad \phi$ is probability of an event

$$p(y;\phi) = \phi^y (1-\phi)^{1-y}$$
$$= \exp\left(y \log \phi + (1-y) \log (1-\phi) \right)$$
$$= \exp\left(\log \frac{\phi}{1-\phi} \cdot y + \log (1-\phi) \right)$$

Check fits into form:

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

$T(y) = y \qquad \eta = \log \frac{\phi}{1-\phi} \qquad b(y) = 1$

Claim: $-a(\eta) = \log(1-\phi)$

Observe: $\eta = \log \frac{\phi}{1-\phi} \Rightarrow \phi = \frac{1}{1+e^{-\eta}}$

Here, $1 - \phi = \frac{e^{-\eta}}{1+e^{-\eta}} = \frac{1}{1+e^{\eta}}$ so $-\log(1-\phi) = \log(1+e^{\eta})$ $\square$.

Example #2 Gaussian (w/ fixed variance) $\sigma^2 = 1$

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

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

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

$\eta = \mu$ $\quad T(y) = y$ and $\quad a(\eta) = \frac{1}{2}\eta^2$ ✓

Why do we care about this form?

Inference is "easy"

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

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

Learning is "well defined"

MLE wrt to η is <u>concave</u>

(so negative log likelihood is convex)

Generalized Linear Models (GLM)

Design choices $\Rightarrow$ Assumptions.

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

Binary $\longrightarrow$ Bernoulli

Real $\longrightarrow$ Gaussian

Counts $\longrightarrow$ Poisson

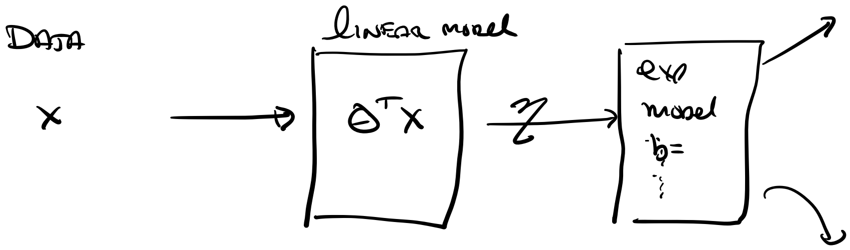
$\mathbb{R}^+$ $\longrightarrow$ Gamma, Exponential

Distribution $\longrightarrow$ Dirichlet

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

(iii) Inference @ test time

Output $\mathbb{E}[y \mid x; \theta]$ i.e. $h_\theta = \mathbb{E}[y \mid x; \theta]$



$\max_\theta \log p(y; \theta^T x^{(i)})$ (train)

$\mathbb{E}[y; \eta] = \mathbb{E}[y; \theta^T x]$

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

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

Terminology

Model parameter		Natural parameter		Canonical
θ	$\xrightarrow{\theta^T x}$	η	$\xrightarrow{g}$ / $\xleftarrow{g^{-1}}$	ϕ : Bernoulli
train on these				μ, σ^2 : Gaussian
				λ : Poisson

g is called the canonical response function

g^{-1} " " the link function

$$\mu = \mathbb{E}[y;\eta] \triangleq g(\eta)$$

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

Logistic regression (Bernoulli)

$$h_\theta(x) = \mathbb{E}[y \mid x; \theta] = \underbrace{\phi}_{\text{canonical}} = \underbrace{\frac{1}{1+e^{-\eta}}}_{\text{natural}} = \underbrace{\frac{1}{1+e^{-\theta^T x}}}_{\text{model}} \in [0,1]$$

Use for classification?

$h_\theta(x) > 0.5 \Rightarrow$ yes 1

o.w. $\Rightarrow$ No 0

Linear regression (Gaussian fixed variance)

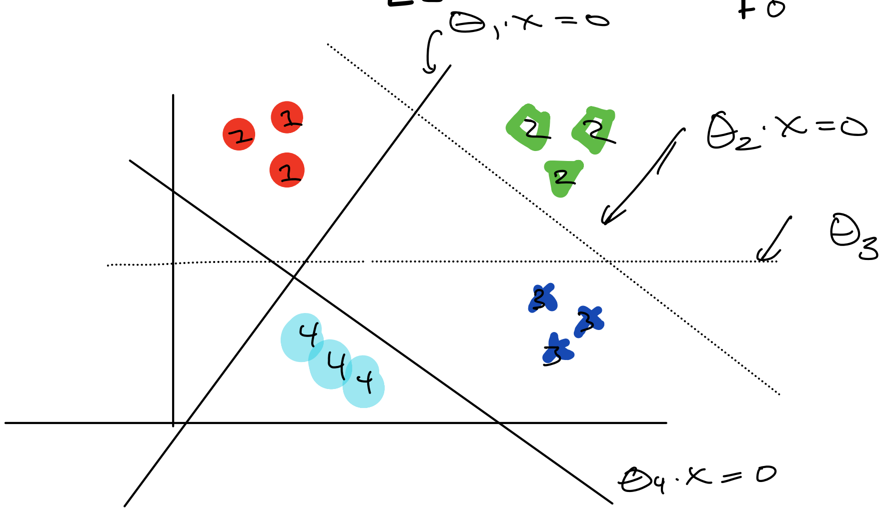
$$h_\theta(x) = \mathbb{E}[y \mid x; \theta] = \mu = \eta = \theta^T x \quad \text{as before!}$$

Multiclass via Softmax (Multinomial)

Discrete values up to k $\quad \{cat, dog, car, bus\} \quad k=4.$

Encoded as one-hot vector $\Rightarrow y \in \{0,1\}^k$

e.g. $k=3$ $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ is class 1 (cat) $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ is class 3 (car)



$\xleftarrow{\quad d \quad}$

$$\theta \in \begin{bmatrix} - \theta_1 - \\ - \theta_2 - \\ - \theta_3 \end{bmatrix} \updownarrow k$$

$x, \theta_i \in \mathbb{R}^d \quad i = 1 \ldots k$

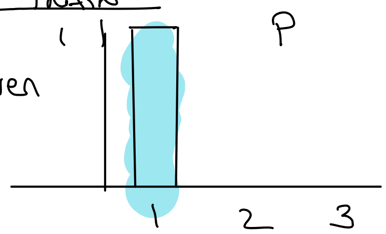
e.g.

					normalize
$\theta_1 \cdot x =$	0.7	Convert to prob.	$e^{0.7}$	≈ 2.013	0.57
$\theta_2 \cdot x =$	-0.5	$\Longrightarrow \exp \Rightarrow$	$e^{-0.5}$	≈ 0.606	$\Rightarrow 0.17$
$\theta_3 \cdot x =$	-0.1		$e^{-0.1}$	≈ 0.904	$\Rightarrow .256$

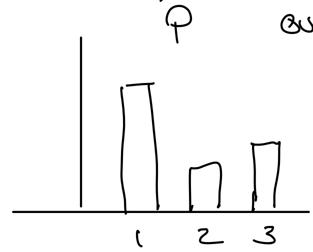
$$P(y=k \mid x; \theta) = \frac{\exp(\theta_k \cdot x)}{\sum_{j=1 \ldots k} \exp(\theta_j \cdot x)}$$

How to train?

given



$\hat{p}$ our estimate



"the label is 1"

$$\min \quad \text{Cross Entropy}(p, \hat{p}) = -\sum_{y=1}^{k} p(y) \log(\hat{p}(y))$$

ground truth

ground truth is i $\quad = -\log(\hat{p}(y_i))$

$$= -\log \frac{\exp(\theta_i \cdot x)}{\sum_{j=1} \exp(\theta_j \cdot x)}$$

Just do gradient descent