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"

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

(y: DATA; η: 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 y

$\Rightarrow$ it MAKES SURE p is probability function

$y, a(\eta), b(y)$ ARE SCALARS

$\eta, T(y)$ ARE SAME DIMENSION

Examples

Bernoulli ϕ 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:

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

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

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

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

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

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

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

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

Why do we care about this form?

Inference is "easy"

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

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

Learning is "well defined"

MLE wrt to η is CONCAVE

(so negative log likelihood is convex)

Generalized Linear Models (GLM)

Design choices $\Rightarrow$ Assumptions.

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

Binary $\rightarrow$ Bernoulli

Real $\rightarrow$ Gaussian

Counts $\rightarrow$ Poisson

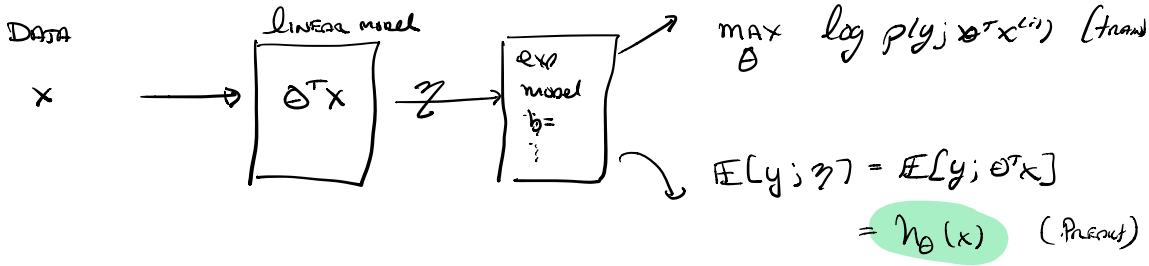
$\mathbb{R}^+$ $\rightarrow$ Gamma, Exponential

Distribution $\rightarrow$ Dirichlet

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

(iii) Define @ test time

Output $E[y|x;\theta]$ i.e. $h_\theta = E[y|x;\theta]$



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
				μ, σ^2 : Gaussian
				λ : Poisson

train on these

g is called the Canonical response function

g^{-1} " " the link function

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

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

logistic regression (Bernoulli)

Canonical Natural Model

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

Use for classification?

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

o.w. $\Rightarrow$ No 0

linear regression (Gaussian fixed variance)

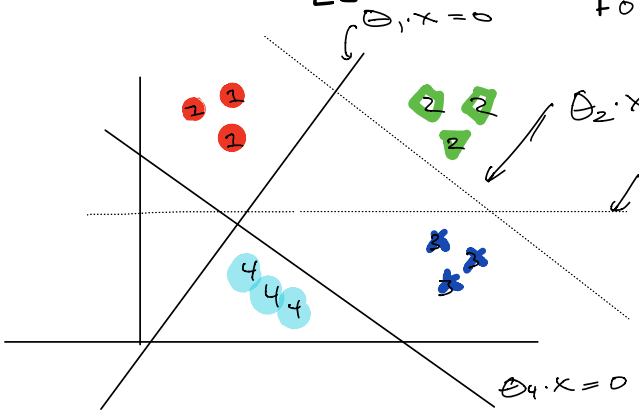
$$h_\theta(x) = \mathbb{E}[y \mid x; \theta] = \mu = \eta = \theta^T x$$ as before

Multiclass via Softmax (Multinomial)

① Discrete values 1 to K $\quad \{\text{cat}, \text{dog}, \text{car}, \text{bus}\}$ $K=4$.

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)



$\Theta \in \begin{bmatrix}\text{—}\ \theta_1\ \text{—}\\ \text{—}\ \theta_2\ \text{—}\\ \text{—}\ \theta_3\ \text{—}\end{bmatrix}$ (K rows, d columns)

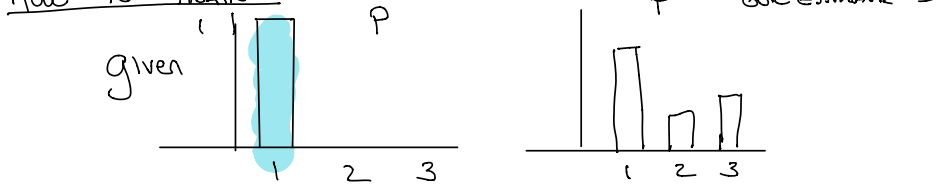
$x, \theta_i \in \mathbb{R}^d \quad i = 1 \dots K$

E.g.

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

$$p(y=k \mid x; \theta) = \frac{\exp(\theta_k \cdot x)}{\sum_{y=1\dots K} \exp(\theta_j \cdot x)}$$

How to train?



"the label is 1"

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

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

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

Just do gradient descent