(I)

Canonical form:

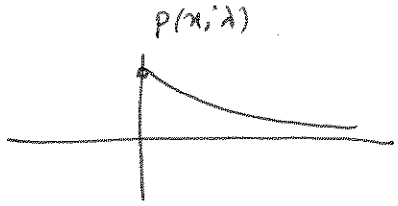
$$p(x;\theta) = \underbrace{\frac{1}{Z(\theta)}}_{\text{Partion function}} \exp(\theta^T \underbrace{T(x)}_{\text{Sufficient Stat}}) \underbrace{h(x)}_{\text{base measure}} = \exp(\theta^T T(x) - \underbrace{A(\theta)}_{\log Z(\theta)}) h(x)$$

if θ is function of something else, say η, then $p(x;\eta)$ is called "curved exponential family" ($\dim(\eta) < \dim(\theta(\eta))$)

low dim rep of θ

Ex 1: Exponential Distribution

$$p(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \le 0 \end{cases}$$



$$p(x;\theta) = \exp(\underbrace{-\lambda}_{\theta} x + \log\lambda) h(x)$$

$T(x) = x$

$\theta = -\lambda$

$h(x) = \mathbb{1}_{(0,\infty)}(x)$

$A(\theta) = \log(+\frac{1}{\lambda})$

Ex 2: → one-hot representation $x = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$ → it means $x = k$

$$\text{Cat}(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k} = \exp\left(\sum_{k=1}^{K} (\log \mu_k) x_k\right)$$

$\mu \in \Delta^K$

$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_K \end{bmatrix}$

$$= \exp\left(\sum_{k=1}^{K-1} (\log \mu_k) x_k + \left(1 - \sum_{k=1}^{K-1} x_k\right) \log\left(1 - \sum_{k=1}^{K-1} \mu_k\right)\right)$$

$$= \exp\left(\sum_{k=1}^{K-1} x_k \left[\log \mu_k - \log\left(1 - \sum_{j=1}^{K-1} \mu_j\right)\right] + \log\left(1 - \sum_{k=1}^{K-1} \mu_k\right)\right)$$

(11)

$$= \exp\left[\sum_{k=1}^{K-1} x_k \log \frac{\mu_k}{\mu_K} + \log \mu_K\right]$$

$$\theta = \left[\log \frac{\mu_1}{\mu_K}, \ldots, \log \frac{\mu_{K-1}}{\mu_K}\right]$$

$$T(x) = \begin{bmatrix} \mathbb{1}(x=1) \\ \vdots \\ \mathbb{1}(x=K-1) \end{bmatrix}$$

$$A(\theta) = -\log \mu_K$$

$$\mu_k = \frac{e^{\theta_k}}{1 + \sum_{j=1}^{K-1} e^{\theta_j}}$$

$$A(\theta) = \log\left(1 + \sum_{k=1}^{K-1} e^{\theta_k}\right)$$

softmax function

$$X_i \overset{i.i.d}{\sim} P(x;\theta)$$

joint dist

$$P(X_1, \ldots, X_n) = \left(\prod_{j=1}^{n} h(x_i)\right) \exp\left(\theta^T \left(\sum_{j=1}^{n} T(x_j)\right) - n A(\theta) \right)$$

joint distr is also exp. family

(11) Dirichlet Distribution:

- Multivariate Gaussian:

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{k}{2}} \det(\Sigma)^{\frac{1}{2}}} \exp\left\{-\tfrac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

$$= \frac{1}{(2\pi)^{\frac{k}{2}}} \exp\left[-\tfrac{1}{2}x^T\Sigma^{-1}x + \underbrace{x^T\Sigma^{-1}\mu}_{\text{already inner product}} - \tfrac{1}{2}\mu^T\Sigma^{-1}\mu - \log\det\Sigma\right]$$

$$-\tfrac{1}{2}\operatorname{tr}(\Sigma^{-1}xx^T)$$

also represented $\langle\Sigma^{-1}, xx^T\rangle$ (inner product)

$$T(x) = \begin{bmatrix} x \\ \operatorname{vec}(xx^T) \end{bmatrix} \qquad A(\theta) = \log\det\Sigma + \tfrac{1}{2}\mu^T\Sigma^{-1}\mu$$

$$\theta = \begin{bmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2}\operatorname{vec}(\Sigma^{-1}) \end{bmatrix}$$

→ let's call $\theta_1 = \Sigma^{-1}\mu$

→ " $\theta_2 = -\frac{1}{2}\Sigma^{-1}$

$h(x) = (2\pi)^{-\frac{k}{2}}$

then $A(\theta_1,\theta_2) = -\tfrac{1}{2}\operatorname{tr}(\theta_2^{-1}\theta_1\theta_1^T) - \tfrac{1}{2}\log(-2\theta_2)$

(IV)

Sufficient statistics:

Def: $t = T(x)$ is sufficient for underlying parameter θ precisely if Conditional probability of data x does not depends on the parameter θ

$$P(x|t,\theta) = P(x|t)$$

→ you already have all info

Alternative way:

$$P(\theta, x|t) = P(x|t,\theta)\, P(\theta|t) = P(x|t)\, P(\theta|t)$$

→ chair rule

$P(x|t)$: Sufficient stat

→ you can easily see

$$P(\theta|t,x) = P(\theta|t)$$

(V)

Partion function :

$$\int \exp(\theta^T T(x)) h(x)\, dx = e^{A(\theta)}$$

$$\nabla_\theta \int \exp(\theta^T T(x))\, h(x)\, dx = \nabla_\theta A(\theta)\, e^{A(\theta)}$$

$$\int T(x) \exp(\theta^T T(x))\, h(x)\, dx = \nabla_\theta A(\theta)\, e^{A(\theta)}$$

$$\int T(x) \underbrace{\exp(\theta^T T(x) - A(\theta))}_{p(x;\theta)} h(x)\, dx = \nabla_\theta A(\theta)$$

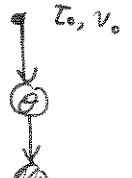
$$\mathbb{E}[T(x)] = \nabla_\theta A(\theta)$$

VI

Posterior estimation

$T \in \mathbb{R}^d$

$$p(x|\theta) = h(x)\exp(\theta^T T(x) - A(\theta))$$



$$p(\theta; \tau, \nu_0) = \exp\left(\nu_0 \tau^T T'(\theta) - B(\tau, \nu_0)\right) h'(\theta)$$

τ: a vector of param; ν_0: scalar param

$T'(\theta)$: a different sufficient stat, $T' \in \mathbb{R}^{d'}$

~~There exists a conjugate prior~~

There exists a distribution s.t. $T'(\theta) = \begin{bmatrix} T(\theta) \\ A(\theta) \end{bmatrix}$ ~~$h(\theta) \propto \exp(\nu_0 A(\theta))$~~ $h'(\theta) = 1$ $(d' = d+1)$

$$p(\theta | \underbrace{\{x_1, \dots, x_n\}}_{D}) = p(\theta|D) \propto \underbrace{p(\theta, D)}$$

$$= \underbrace{p(D|\theta)}\, p(\theta)$$

$$\propto \left(\prod_{i=1}^n h(x_i)\right) \left\{ \exp\left(\theta^T \overbrace{\sum_{i=1}^n T(x_i)}^{\bar{T}} - n A(\theta) \right) \times \right.$$

$$\left. \exp\left(\nu_0 \theta^T \tau_0 - \nu_0 A(\theta)\right) \right.$$

$$\propto \underbrace{\exp\left(\theta^T \left(N\bar{T} + \nu_0 \tau_0 \right) - A(\theta)(\nu_0 + n) \right)}$$

$$= \exp\left((\nu_0 + n) \left(\frac{N\bar{T} + \nu_0 \tau_0}{\nu_0 + n} \right)^T \theta - A(\theta)(\nu_0 + n) \right)$$

$$= p\left(\theta; \frac{N\bar{T} + \nu_0 \tau_0}{\nu_0 + n}, \nu_0 + n\right)$$

Let's assume y is scalar. Remember, the exponential family:

$$p(y; \theta) = \exp\left[\theta y - A(\theta)\right] h(y)$$

θy → we are assuming $T(y) = y$

I am going to generalize it:

$$p(y; \theta, \sigma^2) = \exp\left[\frac{\theta y - A(\theta)}{\sigma^2}\right] h(y, \sigma^2)$$

Remember:

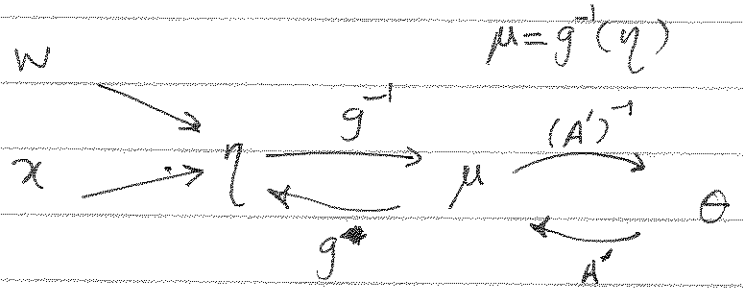
(mean) $\mu = \mathbb{E}[y] = A'(\theta)$

Let's assume σ is known.

We want to predict the using a linear function of features x. In other words $w^T x$. We can also assume any invertible function of mean. Let's call that function g^{-1} (mean function)

$$\mu = g^{-1}(w^T x)$$

g^{-1} is called mean function and $g(\cdot)$ is called link function. Let's define $\eta = w^T x$



Let's pick g to be $(A')^{-1}$

$$g \equiv (A')^{-1}$$ (Canonical Link function)

then $\mu_i = g^{-1}(\eta) \Rightarrow \theta = \eta$

Now, let's assume we have pairs of (x_i, y_i).

Th ML estimation:

$$\max_w \log p(D; w) = \sum_{i=1}^{n} \log p(y_i; w) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \theta_i y_i - A(\theta_i)$$

note that I have indexed θ_i with i b/c it's a funct. of x_i

Let's define

$$l_i \triangleq \theta_i y_i - A(\theta_i)$$

$$\frac{\partial l_i}{\partial w} = \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial w}$$

$\frac{\partial l_i}{\partial \theta_i} = y_i - A'(\theta_i) = y_i - \mu_i$

$\frac{\partial \eta_i}{\partial w} = x_i$

$$\frac{\partial l_i}{\partial w} = (y_i - \mu_i) \frac{\partial \theta}{\partial \mu} \frac{\partial \mu}{\partial \eta} x_i$$

For canonical link function $\theta = \eta$ it become

$$\frac{\partial l_i}{\partial w} = (y_i - \mu_i) x_i$$

$$\nabla_w \mathcal{L} = \frac{1}{\sigma^2} \left[\sum_{i=1}^{N} (y_i - \mu_i) x_i \right]$$