

2. Log-linear Models

Exponential Family

$$p(x|\theta) = \frac{1}{Z(\theta)} h(x) \exp(\theta \cdot \phi(x))$$

partition function
determines supports (zeros)
canonical parameters
sufficient statistics.

Good because:

- Only family with finite sufficient statistics
- Only family with conjugate priors
- Corresponds to maximum entropy distributions.

Example 1: Beta distribution

$$p(x|\alpha, \beta) = \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad \alpha > 0, \beta > 0, \text{ support is } [0, 1]$$

$$\begin{aligned}
 p(x|\alpha, \beta) &= \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)} & - \phi(x) &= [\log(x), \log(1-x)] \\
 &= \frac{1}{B(\alpha, \beta)} \cdot x^{\alpha-1} \cdot (1-x)^{\beta-1} & - \theta &= [\alpha-1, \beta-1] \\
 &= \frac{1}{B(\alpha, \beta)} \cdot \exp(\log(x^{\alpha-1} \cdot (1-x)^{\beta-1})) & - \xi(\alpha, \beta) &= B(\alpha, \beta) \\
 &= \frac{1}{B(\alpha, \beta)} \cdot \exp(\log(x^{\alpha-1}) + \log((1-x)^{\beta-1})) & - h(x) &= 1 \\
 &= \frac{1}{B(\alpha, \beta)} \cdot \exp((\alpha-1) \log(x) + (\beta-1) \log(1-x))
 \end{aligned}$$

Example 2: Chi-square distribution

$$p(x, k) = \frac{x^{k/2-1} \exp(-x/2)}{2^{k/2} \Gamma(k/2)}, \quad k \in \mathbb{N}^+, \text{ support} = \begin{cases} (0, \infty), & k=1 \\ [0, \infty), & \text{otherwise} \end{cases}$$

$$\begin{aligned}
 p(x, k) &= \frac{x^{k/2-1} \exp(-x/2)}{2^{k/2} \Gamma(k/2)} \\
 &= \frac{1}{2^{k/2} \Gamma(k/2)} \cdot x^{k/2-1} \exp(-x/2) \\
 &= \frac{1}{2^{k/2} \Gamma(k/2)} \cdot \exp(\log(x^{k/2-1} \exp(-x/2))) & - \theta &= k/2 - 1 \\
 &= \frac{1}{2^{k/2} \Gamma(k/2)} \cdot \exp(\log(x^{k/2-1}) - \frac{x}{2}) & - \phi(x) &= \log(x) \\
 & & - \xi(\theta) &= 2^{\theta+1} \Gamma(\theta+1)
 \end{aligned}$$

$$= \frac{1}{2^{k/2} \Gamma(k/2)} \cdot \exp\left(\left(\frac{k}{2} - 1\right) \cdot \log(x) - \frac{x}{2}\right)$$

$$= \frac{1}{2^{k/2} \Gamma(k/2)} \cdot \exp\left(-\frac{x}{2}\right) \cdot \exp\left(\left(\frac{k}{2} - 1\right) \cdot \log(x)\right)$$

$$- \mathcal{Z}(k) = 2^{k/2} \Gamma(k/2)$$

$$- h(x) = -x/2$$

Example 3: Uniform distribution

$$p(x, b) = \begin{cases} \frac{1}{b}, & x \in [0, b] \\ 0, & \text{otherwise} \end{cases}, \quad b \in (0, \infty), \text{ support is } [0, b]$$

$$p(x, b) = \mathbb{1}\{x \in [0, b]\} \cdot \frac{1}{b}$$

$$= \exp(\log(\mathbb{1}\{x \in [0, b]\} \cdot \frac{1}{b}))$$

$$= \exp(\log(\mathbb{1}\{x \in [0, b]\}) + \log(\frac{1}{b}))$$

$$= \exp(\underbrace{\log(\mathbb{1}\{x \in [0, b]\})}_{\phi(x)} - \log(b))$$

Not exponential family!!

$\phi(x)$ not only depends on x , but also b .

Log-linear Models

$$p(y|x, \theta) = \frac{1}{\mathcal{Z}(\theta)} \exp(\theta \cdot f(x, y)), \quad \mathcal{Z}(\theta) = \sum_{y' \in Y} \exp(\theta \cdot f(x, y')) \quad \text{* } O(|Y|) \text{ compute}$$

Inputs: $x \in X$ Feature function: $f: X \times Y \rightarrow \mathbb{R}^k$
 Outputs: $y \in Y$ Parameters: $\theta \in \mathbb{R}^k$

Softmax

$$\text{softmax}(h, y, T) = \frac{\exp(h_y/T)}{\sum_{y' \in Y} \exp(h_{y'}/T)}, \quad h_y = \theta \cdot f(x, y) \quad \text{* can be a blackbox}$$

Temperature: $T \in \mathbb{R}$, $T \rightarrow \infty$ uniform, $T \rightarrow 0$ argmax (annealing)

Log-linear models & Logistic Regression

Let $Y \in Y = \{0, 1\}$, $\theta \in \mathbb{R}^{2d}$, $f(x, y) \in \mathbb{R}^{2d}$, consider $p(y|x, \theta) = \frac{1}{\mathcal{Z}(\theta)} \exp(\theta \cdot f(x, y))$.

Assume we have d base features, i.e., $f(x, y) = [f_1(x) \wedge y=1, \dots, f_1(x) \wedge y=0]$. With $\bar{\theta} =$

$[\theta_1 - \theta_{d+1}, \dots, \theta_d - \theta_{2d}]$ and $\bar{f}(x) = [f_1(x), \dots, f_d(x)]$. Show $p(y=1|x, \theta) = \sigma(\bar{\theta}, \bar{f}(x))$.

$$p(y=1|x, \theta) = \frac{\exp(\theta \cdot f(x, 1))}{\sum_{y' \in Y} \exp(\theta \cdot f(x, y'))}$$

$$= \frac{\exp(\theta \cdot f(x, 1))}{\exp(\theta \cdot f(x, 0)) + \exp(\theta \cdot f(x, 1))}$$

$$= \frac{\cancel{\exp(\theta \cdot f(x, 1))} / \cancel{\exp(\theta \cdot f(x, 1))}}{\frac{\exp(\theta \cdot f(x, 0))}{\cancel{\exp(\theta \cdot f(x, 1))}} + \frac{\exp(\theta \cdot f(x, 1))}{\cancel{\exp(\theta \cdot f(x, 1))}}} \quad \text{1}$$

$$= \frac{\exp(\theta \cdot f(x, 0))}{\exp(\theta \cdot f(x, 0)) + \exp(\theta \cdot f(x, 1))} \quad \text{1}$$

sigmoid function

$$\begin{cases} \theta \cdot f(x, 0) = \sum_{i=1}^d \theta_i f_i(x) \\ \theta \cdot f(x, 1) = \sum_{i=1}^d \theta_{d+i} f_i(x) \end{cases}$$

$$\rightarrow \sum_{i=1}^d (\theta_{d+i} - \theta_i) f_i(x) = \sum_{i=1}^d (\theta_i - \theta_{d+i}) f_i(x)$$

$$= \frac{1}{1 + \exp(\theta \cdot f(x, 0) - \theta \cdot f(x, 1))}$$

* Because we only have 0, 1 $\bar{f}(x)$

$$= \frac{1}{1 + \exp(-\bar{\theta} \cdot \bar{f}(x))} = \sigma(\bar{\theta}, \bar{f}(x))$$