Generalized Linear Models (GLM)

Logistic regression corresponds to $p(y|x,w) = Ber(y|\sigma(w^T x))$

Linear regression corresponds to $p(y|x,w) = \mathcal{N}(y|w^T x, \sigma^2)$

In both cases, the mean of the output $\mathbb{E}(y|x,w)$ is a linear function of the inputs x.

There is a broad family of models with this property, known as **generalized linear models** or **GLMs.**

Consider a family of probability distributions parameterized by $\eta \in R^K$ with fifixed support over $Y^D \subseteq R^D$. We say that the distribution $p(y|\eta)$ is in the **exponential family** if its density can be written in the following way:

$$p(y|\eta) = \frac{1}{Z(\eta)} h(y) \exp[\eta^T t(y)] = h(y) \exp[\eta^T t(y) - A(\eta)]$$

By defining $\eta = f(\phi)$,

$$p(y|\phi) = h(y) \exp[f(\phi)^T t(y) - A(f(\phi))]$$

If the mapping from ϕ to η is nonlinear, we call this a curved exponential family. If $\eta = f(\phi) = \phi$, the model is said to be in canonical form. If, in addition, $t(y) = y$, we say this is a natural exponential family or NEF. In this case,

$$p(y|\eta) = h(y) \exp[\eta^T y - A(\eta)]$$

Bernoulli distribution

$$\begin{aligned}
\mathrm{Ber}(y|\mu) &= \mu^y (1-\mu)^{1-y} \\
&= \exp[y \log(\mu) + (1-y) \log(1-\mu)] \\
&= \exp[\mathbf{t}(y)^\top \boldsymbol{\eta}]
\end{aligned}$$

where $\mathbf{t}(y) = [\mathbb{I}(y=1), \mathbb{I}(y=0)]$, $\boldsymbol{\eta} = [\log(\mu), \log(1-\mu)]$, and μ is the mean parameter.

Categorical distribution

the discrete distribution with K categories

$$
\begin{aligned}
\mathrm{Cat}(y|\boldsymbol{\mu}) &= \prod_{k=1}^{K} \mu_k^{y_k} = \exp\left[\sum_{k=1}^{K} y_k \log \mu_k\right] \\
&= \exp\left[\sum_{k=1}^{K-1} y_k \log \mu_k + \left(1 - \sum_{k=1}^{K-1} y_k\right) \log(1 - \sum_{k=1}^{K-1} \mu_k)\right] \\
&= \exp\left[\sum_{k=1}^{K-1} y_k \log\left(\frac{\mu_k}{1 - \sum_{j=1}^{K-1} \mu_j}\right) + \log(1 - \sum_{k=1}^{K-1} \mu_k)\right] \\
&= \exp\left[\sum_{k=1}^{K-1} y_k \log\left(\frac{\mu_k}{\mu_K}\right) + \log \mu_K\right]
\end{aligned}
$$

where $\mu_K = 1 - \sum_{k=1}^{K-1} \mu_k$. We can write this in exponential family form as follows:

$$
\begin{aligned}
\mathrm{Cat}(y|\boldsymbol{\eta}) &= \exp(\boldsymbol{\eta}^\top \mathbf{t}(\mathbf{y}) - A(\boldsymbol{\eta})) \\
\boldsymbol{\eta} &= [\log \frac{\mu_1}{\mu_K}, \ldots, \log \frac{\mu_{K-1}}{\mu_K}] \\
A(\boldsymbol{\eta}) &= -\log(\mu_K) \\
\mathbf{t}(y) &= [\mathbb{I}\,(y=1)\,, \ldots, \mathbb{I}\,(y=K-1)] \\
h(y) &= 1
\end{aligned}
$$

Univariate Gaussian / Multivariate Gaussian

The univariate Gaussian is usually written as follows:

$$
\begin{aligned}
\mathcal{N}(y|\mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp[-\frac{1}{2\sigma^2}(y-\mu)^2] \\
&= \frac{1}{(2\pi)^{\frac{1}{2}}} \exp[\frac{\mu}{\sigma^2} y - \frac{1}{2\sigma^2} y^2 - \frac{1}{2\sigma^2}\mu^2 - \log\sigma]
\end{aligned}
$$

We can put this in exponential family form by defining

$$
\begin{aligned}
\boldsymbol{\eta} &= \begin{pmatrix} \mu/\sigma^2 \\ -\frac{1}{2\sigma^2} \end{pmatrix} \\
\mathbf{t}(y) &= \begin{pmatrix} y \\ y^2 \end{pmatrix} \\
A(\boldsymbol{\eta}) &= \frac{\mu^2}{2\sigma^2} + \log\sigma = \frac{-\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2) \\
h(y) &= \frac{1}{\sqrt{2\pi}}
\end{aligned}
$$

$$\frac{dA}{d\eta} = \mathbb{E}(t(y))$$

$$\frac{d^2A}{d\eta^2} = var(t(y))$$

Generalized linear models (GLMs)

$p(y_n|x_n, w, \sigma^2) = \exp[\frac{y_n \eta_n - A(\eta_n)}{\sigma^2} + \log h(y_n, \sigma^2)]$

where $\eta_n = w^T x_n$ is the natural parameter.

$\mathbb{E}[y_n|x_n, w, \sigma^2] = A'(\eta_n)$

$\mathbb{V}[y_n|x_n, w, \sigma^2] = A''(\eta_n)\sigma^2$

Maximum likelihood estimation

The negative log-likelihood:

$-\log p(D|w) = -\frac{1}{\sigma^2}\Sigma_{n=1}^N l_n$

where $l_n = \eta_n y_n - A(\eta_n)$

The gradient for a single term:

$\frac{\partial l_n}{\partial w} = (y_n - A'(\eta_n))x_n = (y_n - \mu_n)x_n$

where $\mu_n = f(w^T x)$, and f is the inverse link function that maps from canonical parameters to mean parameters. The gradient can be used in SGD.