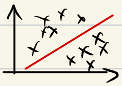
Week 6: Generalized linear models & Exponential families

Motivation:

- least squares:

$$\min \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T w)^2$$



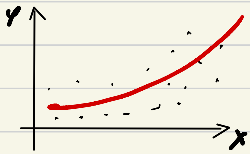
- $y = x^T w + \varepsilon$, $\varepsilon \sim \mathcal{N}(0,\sigma^2)$

$\Rightarrow y \sim \mathcal{N}(x^T w, \sigma^2 I)$

$\rightarrow$ MLE $\rightarrow \hat{w}$

- very simple model

In general we have more complicated models:



We can:
- do some feature augmentation $(x, x^2 \dots)$
- Richer class of probabilistic models (beyond Gaussian and logistic)

$\Rightarrow$ Exponential families

Exponential family:

- logistic regression:

$$\begin{cases} \mathbb{P}(y=1|x,w) = \sigma(x^T w) = \dfrac{e^\eta}{1+e^\eta} \\ \mathbb{P}(y=0|x,w) = 1-\sigma(x^T w) = \dfrac{1}{1+e^\eta} \end{cases} \qquad \eta = x^T w$$

Single form:

- $$\mathbb{P}(y|\eta) = \frac{e^{\eta y}}{1+e^\eta}$$

$$= e^{\eta y - \log(1+e^\eta)}$$

$$= e^{[\eta\phi(y) - A(\eta)]} \quad \text{with } \phi(y) = y, \quad A(\eta) = \log(1+e^\eta)$$

- η is related to $\mu = \mathbb{E}[Y]$ through a nonlin. relat.:

$$\underbrace{\eta = \log\left(\frac{\mu}{1-\mu}\right)}_{\text{link fct}} \Longleftrightarrow \mu = \sigma(\eta) \quad \{ \mu = \mathbb{E}[Y] = \mathbb{P}(y=1).1 = \sigma(\eta)$$

Exponential family:

- $y \in \mathbb{R}, \mathbb{R}_+, \{0,1\}, \mathbb{N} \dots$ can be cont. and disc.

$$\mathbb{P}(y|\eta) = h(y)\, e^{[\eta^T \phi(y) - A(\eta)]}$$

- η : natural parameter
- $h(y) \geq 0$
- $\phi(y)$: sufficient statistic
- $A(\eta)$: (normalization) cumulant, log partition

$$\int h(y)\, e^{\eta^T \phi(y) - A(\eta)} dy = 1$$

$$\Rightarrow A(\eta) = \log\left(\int h(y)\, e^{(\eta^T \phi(y))} dy \right)$$

$\Rightarrow$ degrees of freedom (h, ϕ, η)

- $N := \{\eta : \int h(y)\, e^{\eta^T \phi(y) - A(\eta)} dy < \infty\}$ Natural par. space

Example:

- Bernoulli dist. with par μ

$$\mathbb{P}(y|\mu) = \mu^y (1-\mu)^{1-y} = \frac{(1-\mu)\mu^y}{(1-\mu)^y}$$

$$= e^{\log(1-\mu) + y \log\left(\frac{\mu}{1-\mu}\right)}$$

$$= e^{\log\left(\frac{\mu}{1-\mu}\right) \cdot y + \log(1-\mu)} = e^{\eta^T \phi(y) - A(\eta)}$$

$\phi(y) = y$, $\eta = \log\left(\frac{\mu}{1-\mu}\right)$; $A(\eta) = -\log\left(1 - \frac{e^\eta}{1+e^\eta}\right) = \log(1 + e^\eta)$

$\eta = g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) \Leftrightarrow \mu = \sigma(\eta) = \frac{e^\eta}{1+e^\eta}$ $\quad | \mu = \mathbb{E}[\phi(y)]$

link fct

Gaussian: $\mathcal{N}(\mu, \sigma^2)$

$$\mathbb{P}(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)}$$

- $y \in \mathbb{R}$
- we have 2 par. (μ, σ^2)

$$\mathbb{P}(y|\mu,\sigma^2) = e^{\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right]}$$
$$= e^{\left(\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}\right)\binom{y}{y^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)}$$

$$\phi(y) = (y, y^2)^T, \quad \eta = \left(\mu/\sigma^2, -1/2\sigma^2\right)^T$$

$$A(\eta) = \mu^2/2\sigma^2 + 1/2\log(2\pi\sigma^2) = -\eta_1^2/4\eta_2 + 1/2\log(-\pi/\eta_2)$$
$$= -\eta_1^2/4\eta_2 - 1/2\log(\eta_2/\pi)$$

$h(y) = 1$

- Link function: $\eta_1 = \frac{\mu}{\sigma^2}$; $\eta_2 = -1/2\sigma^2 \Rightarrow \begin{cases} \mu = -\eta_1/2\eta_2 \\ \sigma^2 = -1/2\eta_2 \end{cases}$

Poisson with mean μ

- $y \in \mathbb{N}$

$$\mathbb{P}(y|\mu) = \frac{\mu^y}{y!} \cdot e^{-\mu}$$
$$= \frac{1}{y!} \cdot e^{y\ln(\mu) - \mu}$$
$$= h(y) \cdot e^{\eta^T\phi(y) - A(\eta)}$$

$h(y) = \frac{1}{y!}$; $\eta = g(\mu) = \ln(\mu) \Rightarrow \mu = e^\eta$

$\phi(y) = y$

⚠ Cauchy dist. is not a member of the exp. family.

Basic properties of $A(\eta)$

- $A(\eta)$ is convex
- $\nabla_\eta A(\eta) = \mathbb{E}[\phi(y)]$
- $\nabla_\eta^2 A(\eta) = \mathbb{E}[\phi(y)\phi(y)^T] - \mathbb{E}[\phi(y)]\,\mathbb{E}[\phi(y)]^T$

Hoelder inequality:

$\|f\|_p = \left(\int |f|^p dy\right)^{1/p}$

$$\|fg\|_1 \leq \|f\|_p \|g\|_q$$

with $1/p + 1/q = 1$

i.e Logistic Regression:

$A(\eta) = \log(1+e^\eta)$

$\nabla A(\eta) = \frac{e^\eta}{1+e^\eta} = \sigma(\eta)$

$\nabla^2 A(\eta) = \sigma(\eta)(1-\sigma(\eta)) \geq 0$

$\nabla A(\eta) = \sigma(\eta) = \mathbb{E}[\phi(y)] = \mathbb{E}[y]$

$\nabla^2 A(\eta) = \sigma(\eta)(1-\sigma(\eta)) = \mu(1-\mu) = \mathrm{Var}(y)$

Link function

Is this fct g s.t.: $\eta = g(\mathbb{E}[\phi(y)])$

- Parameter estimation
  - Fixed family (h, ϕ) + data $(y_i)_{i=1}^N$ i.i.d
  - $\rightarrow$ Recover η
- MLE:

$$\mathbb{P}(y|\eta) = h(y)\, e^{\eta^T\phi(y) - A(\eta)}$$

$$\mathcal{L}(\eta) = -\ln(\mathbb{P}(y|\eta)) = \sum_{i=1}^N -\ln(\mathbb{P}(y_i|\eta))$$

$$= \sum_{i=1}^N [-\ln(h(y_i)) - \eta^T\phi(y_i) + A(\eta)]$$

$$\nabla\mathcal{L}(\eta) = \sum_{i=1}^N -\phi(y_i) + N\nabla A(\eta)$$

$$= -\sum_{i=1}^N \phi(y_i) + N\,\mathbb{E}[\phi(y)]$$

$$\nabla\mathcal{L}(\eta) = 0 \iff \mathbb{E}[\phi(y)] = \frac{1}{N}\sum_{i=1}^N \phi(y_i)$$

with a link fct g:

$$\eta = g(\mathbb{E}[\phi(y)]) \Rightarrow \boxed{\eta = g\left(\frac{1}{N}\sum_{i=1}^N \phi(y_i)\right)}$$

Generalized linear model:

(x, y) iid

$$\mathbb{P}(y \mid w, x) = h(y) \cdot e^{x^T w \, \phi(y) - A(x^T w)}$$

for $\eta = x^T w$: the linear prediction

$S_{Train} = (x_i, y_i)_{i=1}^{N}$, Goal: Find w

How: MLE

Negative log likelihood:

$$\mathcal{L}(w) = - \sum_{i=1}^{N} \log \left(\mathbb{P}(y_i \mid x_i, w) \right)$$

$$= - \left[\sum_{i=1}^{N} \log(h(y_i)) + x_i^T w \, \phi(y_i) - A(x_i^T w) \right]$$

$\mathcal{L}$ is convex.

- $$\nabla \mathcal{L}(w) = \sum_{i=1}^{N} - x_i \, \phi(y_i) + A'(x_i^T w) \, x_i$$

$$= \sum_{i=1}^{N} - x_i \, \phi(y_i) + \mathbb{E}[\phi(y_i)] \, x_i$$

$$= \sum_{i=1}^{N} - x_i \, \phi(y_i) + g^{-1}(x_i^T w) \, x_i$$

$$\nabla \mathcal{L}(w) = 0 \iff g^{-1}(x_i^T w) x_i = \sum_{i=1}^{N} x_i \, \phi(y_i)$$

$$\iff g^{-1}(x_i^T w) x_i - \sum_{i=1}^{N} x_i \, \phi(y_i) = 0 \qquad \cancel{(LS: g = id)}$$

$$\iff X^T \left[g^{-1}(Xw) - y \right] = 0 \qquad \cancel{(LR: g^{-1} = \sigma)}$$

Recap:

- linear Model: $y = X^T w + \varepsilon$ $\rightsquigarrow$ LS estimator
- Logistic regression: $\mathbb{P}(y = 1 \mid x, w) = \sigma(x^T w)$
- Exponential family: $\mathbb{P}(y \mid \eta) = h(y) \, e^{\eta^T \phi(y) - A(\eta)}$
  - h, ϕ you can decide
  - η : nature par.
  - A : log-partition (cumulant)
    - $A(\eta)$ is convex
  - $\nabla A(\eta) = \mathbb{E}[\phi(y)]$

- G.L.M.

$$\mathbb{P}(y \mid x, w) = h(y) \, e^{(x^T w \, \phi(y) - A(x^T w))}$$

$\rightsquigarrow$ with MLE find $\hat{w}$