# Newton method & GLM

Monday, 19 March 2018 10:28

Recall: $P(y=1|x;\theta) = h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$

$l(\theta) = \sum y^{(i)} \log h(x^{(i)}) + (1-y^{(i)}) \log(1-h(x^{(i)}))$

$\theta_j := \theta_j + \alpha(y^{(i)} - h(x^{(i)})) \cdot x_j^{(i)}$

Newton method:

$\frac{f(\theta^{(0)})-0}{\theta^{(0)}-0} = f'(\theta_0)$

$\Rightarrow \theta^{(t+1)} = \theta^{(t)} - \frac{f(\theta^{(t)})}{f'(\theta^{(t)})}$

to maxmize $l(\theta)$, want $l'(\theta)=0$

When $\theta$ is a vector:

$$\theta^{(t+1)} = \theta^{(t)} - H^{-1} \nabla_\theta l(\theta)$$

H is hessian matrix

$$H_{ij} = \frac{\partial l}{\partial \theta_i \partial \theta_j}$$

# Short coming: need to inverse Hessian.

Generalized Linear Model:

$P(y|x;\theta)$

$y \in \mathbb{R}$ : Gussian $\to$ least square.

$y \in \{0,1\}$ : Bernouli $\to$ Logistic function

Bernouli $(\phi)$ : $P(y=1;\phi) = \phi$.

exponential function

$$P(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

$\eta$: natural parametric

$T$: sufficient - statistic.

$(a,b,T)$ can determine a special instance of exp family.

Ber $(\phi)$ $P(y=1;\phi) = \phi$

$P(y;\phi) = \phi^y (1-\phi)^{1-y}$

$= \exp(\log \phi^y (1-\phi)^{1-y})$

$= \exp(y\log\phi + (1-y)\log(1-\phi))$

$= \exp(\underbrace{\log\frac{\phi}{1-\phi}}_{\eta} \underbrace{y}_{T(y)} + \underbrace{\log(1-\phi)}_{-a(\eta)})$

$\phi = \frac{1}{1+e^{-\eta}}$

$$\begin{cases} a(\eta) = -\log(1-\phi) = \log(1+e^{\eta}) \\ T(y) = y \\ b(y) = 1 \end{cases}$$

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

Set $\sigma^2 = 1$ since it will take no effect on $\theta$.

$\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y-\mu)^2)$

$\Rightarrow$: $\underbrace{\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2)}_{b(y)} \exp(\mu y - \frac{1}{2}\mu^2)$

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

Gama, Possion, exponential and $\beta$, Dirchlet, Wishart distribution is exponential destribution family.

GLM:

Assume:

1): $y|x;\theta \sim$ ExpFamily$(\eta)$

2): Given $x$: goal is to output $E[T(y)|x]$

want $h(x) = E[T(y)|x]$

3): $\eta = \theta^T x$ // $\eta_i = \theta_i^T x$

Bernouli:

$y|x;\theta \sim$ ExpFamily$(\mu)$

For fixed, $x, \theta$, algorithm output

$h_\theta(x) = E[y|x;\theta] = P(y=1|x;\theta)$

$= \phi$

$= \frac{1}{1-e^{-\mu}}$

$= \frac{1}{1-e^{-\theta^T x}}$

$g(\eta) = E[y;\eta] = \frac{1}{1+e^{-y}}$

canonical response function

$g^{-1}$ canonical link function.

example:

Multinomial : $y \in \{1, \dots k\}$

Parameter: $\phi_1, \phi_2 \dots \phi_k$.

$P(y=i) = \phi_i$.

$\phi_k = 1-(\phi_1 + \dots + \phi_{k-1})$

parameter $(\phi_1, \phi_2 \dots \phi_{k-1})$

$T(y) \neq y$!

$T(1) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{k-1}$ $T(2) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ $\in \mathbb{R}^{k-1}$

$T(k) = [0, 0 \dots 0]^T$

$1\{true\} = 1$ / $1\{false\} = 0$

$T(y)_i = 1\{y=i\}$.

the ith element of $T(y)$

$P(y) = \phi_1^{1\{y=1\}} \phi_2^{1\{y=2\}} \dots \phi_k^{1\{y=k\}}$

$= \phi_1^{T(y)_1} \phi_2^{T(y)_2} \dots \phi_k^{1-\sum_{j=1}^{k-1} T(y)_j}$

$= b(y)\exp(\eta^T T(y) - a(\eta))$

where $\eta = \begin{bmatrix} \log(\phi_1/\phi_k) \\ \vdots \\ \log(\phi_{k-1}/\phi_k) \end{bmatrix}_{k-1}$ $a(\eta) = -\log(\phi_k)$

$b(y) = 1$

$\downarrow$

$\phi_i = \frac{e^{\eta_i}}{1+\sum_{i=1}^{n} e^{\eta_i}}$ $(i = 1, \dots k-1)$

$= \frac{e^{\theta_i^T x}}{1+\sum_{j=1}^{k-1} e^{\theta_j^T x}}$

$h_\theta(x) = E[T(y)|x;\theta]$

$= E\left[\begin{bmatrix} 1\{y=1\} \\ \vdots \\ 1\{y=k-1\} \end{bmatrix} \middle| x;\theta\right] = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{k-1} \end{bmatrix}$

$= \begin{bmatrix} e^{\theta_1^T x} / (1+\sum_{j=1}^{k-1} e^{\theta_j^T x}) \\ \vdots \\ e^{\theta_{k-1}^T x} / (1+\sum_{j=1}^{k-1} e^{\theta_j^T x}) \end{bmatrix}$

Soft max regression: $(2 \to k)$

train set: $(x^{(1)}, y^{(1)}) \dots (x^{(m)}, y^{(m)})$ $y \in \{1, \dots k\}$

$L(\theta) = \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta)$

$= \prod_{i=1}^{m} \phi_1^{1\{y^{(i)}=1\}} \dots \phi_k^{1\{y^{(i)}=k\}}$

$\downarrow$

$\frac{e^{\theta_i^T x}}{1+\sum_{j=1}^{k-1} e^{\theta_j^T x}}$