CS 307 Lecture 7

Logistic Regression:

Training Set $= \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$

$x^{(i)} \in \mathbb{R}^{d+1}$, $y^{(i)} \in \{0,1\}$

Goal: Find $\theta \in \mathbb{R}^{d+1}$ such that

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} l_{nll}(\theta, x^{(i)}, y^{(i)}) \text{ is minimized}$$

$$l_{nll}(\theta, x, y) = -\left[y \log h_\theta(x) + (1-y)\log(1-h_\theta(x))\right]$$

Maximum Likelihood Estimation

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

Intuitively $P[y=1 \mid x; \theta] = h_\theta(x)$.

Gradient descent update

$$\theta_j = \theta_j - \alpha \frac{1}{n}\sum_{i=1}^{n}\left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)}$$

The "same" as linear regression

Exponential Family of Distributions

$$p(y;\eta) = b(y)\, e^{(\eta^T T(y) - a(\eta))}$$

$p(y;\eta)$: Distribution over y; η: parameter

p.d.f / p.m.f

η — natural parameter

T — sufficient statistic

a — log partition.

Generalized Linear Models

(i) $P(y \mid x; \theta) \sim$ Exponential Family (η)

(ii) Goal of the learning algorithm is to

$$E[\underbrace{T(y)}_{=y} \mid x; \theta]$$

(iii) $\eta = \theta^T x$ when $\eta \in \mathbb{R}$

$\eta_i = \theta^{(i)T} x$ when $\eta \in \mathbb{R}^k$.

We get θ by maximized likelihood

maximizing log likelihood

minimizing negative log likelihood

Logistic Regression

$$P[y=1 \mid x; \theta] \sim \text{Bernouli}(\phi)$$

ϕ → depends on x & θ.

P.m.f of Bernouli (ϕ)

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

→ $\exp(z) = e^z$

$$= \exp[y \ln \phi + (1-y) \ln(1-\phi)]$$

$$= \exp\left[y \ln \frac{\phi}{1-\phi} + \ln(1-\phi)\right]$$

$b(y) = 1$, $T(y) = y$, $\eta = \ln \frac{\phi}{1-\phi}$ $\quad a(\eta) = \ln(1-\phi)$

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

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

$$\phi = \text{sigmoid}(\eta)$$

$$a(\eta) = \ln(1-\phi) = \ln\left[1 - \frac{1}{1+e^{-\eta}}\right]$$

Learning algorithm of Logistic Regression found

$$\rightarrow h_\theta(x) = p[y=1 \mid x;\theta]$$

$$\stackrel{?}{=} E[y \mid x;\theta]$$

$$= \phi$$

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

$z \sim$ Bernouli (ϕ)

$$E(z) = \sum_{z=i} i \cdot p(z=i)$$

$$= 1 \cdot p(z=1) + 0 \cdot p[z=0]$$

$$= p(z=1) = \phi$$

Softmax Regression

~~Outp~~ Output $\in \{1, 2, \ldots k\}$.

$$p[y=i \mid x;\Theta] \sim \text{Multinomial Dist}(\phi_1, \phi_2 \ldots \phi_k)$$

prob that output is 2.

$$= \phi_i$$

$$\phi_k = 1 - \sum_{i=1}^{k-1} \phi_i$$

$$p[y=i \mid x;\Theta] \sim \text{Multinomial Dist}(\phi_1 \ldots \phi_{k-1})$$

One-shot encoding.

$$T(y) = \begin{bmatrix} 1\{y=1\} \\ 1\{y=2\} \\ \vdots \\ 1\{y=i\} \\ \vdots \\ 1\{y=k-1\} \end{bmatrix} \in \{0,1\}^{k-1}$$

Example: $k = 4$.

$$T(2) = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$T(4) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$p[y \mid x; \Theta] = \phi_1^{1\{y=1\}} \phi_2^{1\{y=2\}} \cdots \phi_{k-1}^{1\{y=k-1\}} \phi_k^{1\{y=k\}}$$

$$\left[\phi_k = 1 - \sum_{i=1}^{k-1} \phi_i \right]$$

$$= \phi_1^{T(y)_1} \phi_2^{T(y)_2} \cdots \phi_{k-1}^{T(y)_{k-1}} \phi_k^{1 - \sum_{i=1}^{k-1} T(y)_i}$$

$$= \exp\left[T(y)_1 \ln \phi_1 + T(y)_2 \ln \phi_2 + \cdots T(y)_{k-1} \ln \phi_{k-1} + \left(1 - \sum T(y)_i\right) \ln \phi_k \right]$$

$$= \exp\left[\underbrace{T(y_1) \ln \frac{\phi_1}{\phi_k} + T(y_2) \ln \frac{\phi_2}{\phi_k} + \cdots T(y)_{k-1} \ln \frac{\phi_{k+1}}{\phi_k}}_{= \eta^T T(y)} + \ln \phi_k \right]$$

$b(y) = 1$, $T(y)$

$$\eta = \begin{bmatrix} \ln \frac{\phi_1}{\phi_k} \\ \vdots \\ \ln \frac{\phi_{k-1}}{\phi_k} \end{bmatrix} \qquad a(\eta) = -\ln(\phi_k)$$

$$\eta_i = \ln \frac{\phi_i}{\phi_k} \Rightarrow \phi_i = \phi_k e^{\eta_i} \qquad [\eta_k = 0$$

$$1 = \sum \phi_i = \phi_k \sum_{i=1}^{k} e^{\eta_i} , \quad \eta_k = 0 .$$

$$\phi_k = \frac{1}{1 + \sum_{i=1}^{k-1} e^{\eta_i}}$$

$$\phi_i = \frac{e^{\eta_i}}{1 + \sum_{i=1}^{k-1} e^{\eta_i}}$$

$$T(y) = \begin{bmatrix} 1(y=1) \\ 1(y=2) \\ \vdots \\ 1(y=k-1) \end{bmatrix}$$

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

$$= \begin{bmatrix} E[1(y=1)] \\ \vdots \\ E[1(y=k-1)] \end{bmatrix} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{k-1} \end{bmatrix}$$

$$\eta_i = \theta^{(i)T} x .$$

$$= \begin{bmatrix} \frac{e^{\eta_1}}{1 + \sum e^{\eta_i}} \\ \vdots \\ \frac{e^{\eta_{k-1}}}{1 + \sum e^{\eta_i}} \end{bmatrix} = \begin{bmatrix} \frac{e^{\theta^{(1)T} x}}{1 + \sum e^{\theta^{(i)T} x}} \\ \vdots \\ \frac{e^{\theta^{(k-1)T} x}}{1 + \sum e^{\theta^{(i)T} x}} \end{bmatrix}$$