4*

Generalized Linear Models

★ The exponential family

⇒ We say that a class of distributions is in the exponential family if it can be written in the form:

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

→ η is called the natural parameters

→ $T(y)$ is the Sufficient statistic (mostly $T(y)=y$)

→ $a(\eta)$ is the log partition function

→ $b(y)$ is called base measure

⇒ A fixed choice of T, a and b defines a family of distributions that is parameterized by η.

↳ As we vary η, we then get different distributions within this family.

⇒ Bernoulli and Gaussian distributions are example of exponential family distribution

① Bernoulli distribution

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

★ **Properties of Exponential family**

① MLE is Convex

② $E[y;\eta] = \frac{\delta}{\delta \eta} a(\eta)$

③ $Var[y;\eta] = \frac{\delta^2}{\delta \eta^2} a(\eta)$

$$P(y;\phi) = \exp\left(\log\left(\frac{\phi}{1-\phi}\right) y + \log(1-\phi) \right)$$

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

② Gaussian distribution (fixed variance $\sigma^2 = 1$)

$$P(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\mu)^2\right)$$

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

- $b(y) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y^2}{2}\right)$
- $T(y) = y$
- $\eta = \mu$
- $a(\eta) = \frac{\eta^2}{2}$

⇒ There are many other distributions that are members of the exponential family:

- multinomial
- Poisson
- gamma
- exponential
- beta
- Dirichlet
- ...etc...

Date
Page

★ Constructing GLMs

⇒ Consider a classification or regression problem where we would like to predict the value of some random variable y as a function of x.

⇒ To derive a GLM for this problem, we will make the following three assumptions about the conditional distribution of y given x and about our model:

① $y|x;\theta \sim$ Exponential Family (η)

② Given x, our goal is to predict the expected value of $T(y)$ given x.
   - → In most cases, $T(y) = y$
   - → So this means we would like the prediction $h(x)$ output by our learned hypothesis h to satisfy.

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

③ The natural parameter η and the ~~output~~ input x are related linearly:

$$\eta = \theta^T x$$

⇒ These three assumptions / design choices will allow us to derive a very elegant class of learning algorithms, namely GLMs.

★ Ordinary Least Square

$y|x;\theta \sim N(\mu, \sigma^2)$

$$\eta = \mu = \theta^T x$$

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

$$= \mu$$

$$\boxed{h_\theta(x) = \theta^T x}$$

★ Logistic Regression

$y|x;\theta \sim \text{Bernoulli}(\phi)$

$$\Rightarrow \eta = \ln\left(\frac{\phi}{\phi - 1}\right)$$

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

$$E[y|x;\theta) = \phi = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-\theta^T x}}$$

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

Date
Page

⇒ To introduce a little more terminology, the

→ The function g given the distribution mean as a function of natural parameter

$$g(\eta) = E[T(y); \eta]$$

is the Canonical response function.

→ Its inverse g^{-1} is called the Canonical link function.

★ Softmax Regression

⇒ Consider a classification problem in which the response variable y can take on any one of K values, so $y \in \{1, 2, \cdots K\}$

⇒ We will thus model it as distributed according to a multinomial distribution.

⇒ Lets derive a GLM for modelling this type of multinomial data.

⇒ To parameterize a multinomial over K possible outcomes, one could use K parameters $\phi_1, \cdots \phi_K$ specifying the probability of each of the outcomes.

⇒ However, these parameters would be redundant, since knowing any $K-1$ of the ϕ_s uniquely determine the last one.

$$\sum_{i=1}^{K} \phi_i = 1 .$$

⇒ So, we will instead parameterize the multinomial with only $K-1$ parameters

$$\phi_1, \ldots \phi_{K-1}$$

$$\phi_i = p(y=i; \phi)$$

$$p(y=K; \phi) = 1 - \sum_{i=1}^{K-1} \phi_i$$

⇒ To express the multinomial as an exponential family distribution, we will define $T(y) \in R^{K-1}$ as follows

$$T(1) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad T(2) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \cdots T(K) \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \quad T(K) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

⇒ Let $1\{\cdot\}$ be an indicator function that takes on value 1 if argument is true, and 0 otherwise.

⇒ So $(T(y))_i = 1\{y=i\}$

$\Rightarrow$ Further we have $E[(T(y))_i] = P(y=i) = \phi_i$

$$P(y;\phi) = \phi_1^{1\{y=1\}} \cdot \phi_2^{1\{y=2\}} \cdots \phi_K^{1\{y=K\}}$$

$T(y)_1 \qquad T(y)_2 \qquad 1 - \sum_{i=1}^{K-1} 1\{y=i\}$

$T(y)_i$

$$\Rightarrow \exp\Big(T(y)_1 \log(\phi_1) + T(y)_2 \log(\phi_2) + \cdots + \Big(1 - \sum_{i=1}^{K-1} T(y)_i\Big) \log(\phi_K) \Big)$$

$$\Rightarrow \exp\Big(T(y)_1 \log\Big(\frac{\phi_1}{\phi_K}\Big) + T(y)_2 \log\Big(\frac{\phi_2}{\phi_K}\Big) + \cdots + T(y)_{K-1} \log\Big(\frac{\phi_{K-1}}{\phi_K}\Big) + \log(\phi_K) \Big)$$

$$\Rightarrow P(y;\phi) = b(y)\exp\left(\eta^T T(y) - a(\eta)\right)$$

Where,
$$\eta = \begin{bmatrix} \log(\phi_1/\phi_K) \\ \log(\phi_2/\phi_K) \\ \vdots \\ \log(\phi_{K-1}/\phi_K) \end{bmatrix}$$

$$a(\eta) = -\log(\phi_K)$$

$$b(y) = 1$$

$\Rightarrow$ The link function is given (for $i = 1, \cdots K$) by

$$\boxed{\eta_i = \log\frac{\phi_i}{\phi_K}}$$

$$e^{\eta_i} = \frac{\phi_i}{\phi_K}$$

$$\phi_K e^{\eta_i} = \phi_i$$

$$\phi_K \sum_{i=1}^{K} e^{\eta_i} = \sum_{i=1}^{K} \phi_i = 1$$

$$\phi_K = \frac{1}{\sum_{i=1}^{K} e^{\eta_i}}$$

$\Rightarrow$ $$\boxed{\phi_i = \frac{e^{\eta_i}}{\sum_{i=1}^{K} e^{\eta_i}}}$$ {Response function}

⇒ Lastly, lets discuss parameter fitting

$$\ell = \sum_{i=1}^{m} \log P(y^{(i)} | x^{(i)}; \theta)$$

$$= \sum_{i=1}^{m} \log \prod_{l=1}^{K} \left(\frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=1}^{K} e^{\theta_j^T x^{(i)}}} \right)^{1\{y^{(i)} = l\}}$$

⇒ We can now obtain the maximum likelyhood estimate of the parameters by maximizing $\ell(\theta)$ in terms of θ, using a method such as gradient ascent or Newton's method.