Advanced Machine Learning, 5 Sep 2019

Sourish

- Linear regression, logistic regression to Gaussian process regression
- Helps explainability — regulated industries

Natural Exponential family

$y \in \mathbb{R}$, random variable

$$f(y) = c(y) \exp\{\theta y - b(\theta)\}$$

θ — natural parameter

Gaussian
Binomial
Poisson

] Can all be written in this form

All very different types of values

Biniary distribution

$$f(y) = p^y (1-p)^{1-y} \qquad y = 0,1 \; ; \; p \in (0,1)$$

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

$$\exp\left(\log\left(p^y (1-p)^{1-y}\right)\right)$$

$$y \underbrace{\log\left(\frac{p}{1-p}\right)}_{\theta} + \underbrace{\log(1-p)}_{\psi(\theta)}$$

range $[-\infty, \infty]$

Features: $x_i = (x_{i1}, \ldots, x_{ip})$ $\qquad y = f(x)$ $= 0,1$

$$E(y) = \mu = g(x\beta)$$

↳ link function

↳ p

Write g as $p = \frac{e^{\theta}}{1+e^{\theta}}$ $\theta = x_i^T \beta$

$\log \frac{p}{1-p} = \theta$

$= \frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}}$ ← sigmoid, aka logit

Logistic regression!

Natural Exponential Family

(1) $f(y) = c(y) \exp\{\theta y - p(\theta)\}$

(2) $E(y) = g(\theta)$

(3) $\theta = x^T \beta$

Everything can be "converted" into a regression

Poisson regression

e.g. $y = \#$ insurance claims $= 0, 1, 2, \ldots$

(1) $f(y) = e^{-\lambda} \cdot \dfrac{\lambda^y}{y!}$ $\quad \lambda \in \mathbb{R}^+$

- rewrite as $\exp\{ \ldots \}$

(2) $\theta = \log \lambda$ $\quad$ [work out!]

(3) $\log \lambda = x_i^T \beta$

Generalized Linear Models

Data Set

$$\begin{bmatrix} y_1 & x_{11} & \cdots & x_{1p} \\ y_2 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \\ y_n & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

$$f(y_i) = p_i^{y} (1-p)^{y_i}$$

$$= \exp\left\{ y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i) \right\}$$

$$p_i = \frac{e^{\theta_i}}{1+e^{\theta_i}} \qquad \theta_i = x_i \beta$$

Each y_i has different p_i — e.g. different risk profile

If we estimate β, $x_i\beta$ gives $\theta_i \Rightarrow p_i$

What is β?

Likelihood function

$$L(\beta \mid y, x) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

rows = observations are independent

$$p_i = \frac{e^{\theta_i}}{1+e^{\theta_i}} = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right)^{y_i} \left(\frac{1}{1+e^{x_i^T\beta}}\right)^{1-y_i}$$

Fn of x_i, y_i, β

Take -log & use stochastic gradient descent

loglikelihood = $\log L = \ell =$

$$l = \sum_{i=1}^{n} y_i x_i^T\beta - y_i \log\left(1+e^{x_i^T\beta}\right) - (1-y_i)\log\left(1+e^{x_i^T\beta}\right)$$

If optimizer maximizes, use l

If optimize minimizes, use $-l$

$$y_i \sim N\left(\mu_i, \sigma_i^2\right)$$

$$E(y_i) = \mu_i = x_i^T\beta$$

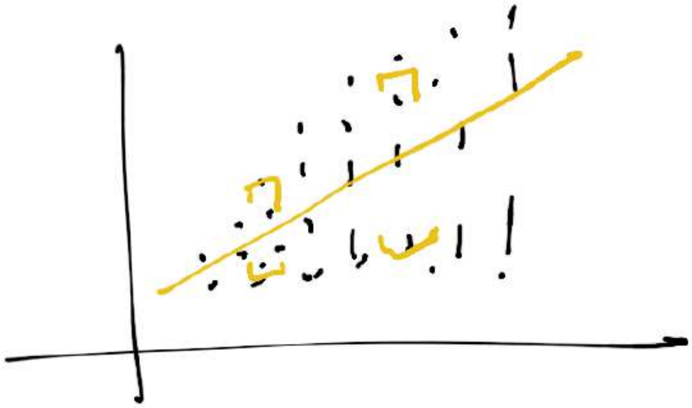
$$x_i^T = (x_{i1}, \ldots, x_{ip})$$

$$\mu_i = x_i^T\beta$$

$$\sigma_i^2 = g(\underline{\underline{x_i^T\beta}}) = \exp(\underline{\underline{x_i^T\beta}})$$

OK if we assume uniform $V(y_i) = \sigma^2 \ \forall i$

- Same confidence interval for all x from μ_i



σ_i^2 increases with x

Confidence interval changes with x

$$\sigma_i^2 = g(x_i^T w) = \exp(x_i^T w)$$

independent parameter!

$\therefore$ More computation to converge

Gaussian Process Regression

$$Y = X\beta + \varepsilon$$

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \\ x_{1n} & & x_{np} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Typical assumption

$$E(\varepsilon) = 0$$

$$V(\varepsilon) = \sigma^2 I_n$$

Normality assumption

Only needed for hypothesis testing / confidence intervals

$$\begin{bmatrix} \sigma^2 & & 0 \\ & \sigma^2 & \\ 0 & & \ddots & \sigma^2 \end{bmatrix}$$

Covariance matrix — rows are independent

Generalize, assume symmetry, $\sigma_{ij} = \sigma_{ji}$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ & \sigma_{22} & \cdots & \sigma_{2n} \\ & & \ddots & \\ & & & \sigma_{nn} \end{bmatrix} \begin{array}{l} - n \text{ parameters} \\ - n-1 \quad " \\ \vdots \\ - 1 \text{ parameter} \end{array}$$

Totally $\dfrac{n(n+1)}{2}$ parameters

i.i.d. assumption on input avoids this parameter explosion

Can we do something else to reduce parameters ?

Assume $p=1$ for simplicity

$$\Sigma = ((\sigma_{ij})) = \left(\left(\sigma^2 \exp\{-\rho \underbrace{\|x_i - x_j\|}_{d_{ij}}\}\right)\right)$$

distance (x_i, x_j)

As $d_{ij} \to \infty$, $\sigma_{ij} \to 0$

$d_{ij} \to 0$, $\sigma_{ij} \to \sigma^2$

This is always a positive definite matrix

Now only 2 parameters fix covariance matrix

σ^2, ρ