Generalized Linear Models

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

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

$$\varepsilon \sim N(0, \sigma^2)$$

$$\approx \quad Y \sim N(X^T\beta, \sigma^2)$$

X wing span

Y

~~body mass ~ $N(\mu, \sigma^2)$~~

Lin beak length ~ $N(\mu, \sigma^2)$

GLM Populationsize ~ $Pois(\lambda)$

Line talon size ~ $N(\mu, \sigma^2)$

Species

10cm hummingbird ~ $Bern(p)$

20cm seagull

$Bin(n=1, p)$

30cm duck

1m Conor ~ $Multinomial(p_k)$

Can predictors explain the outcome?

Logistic Reg

GLM

$$E(Y) = g^{-1}(X^T \beta)$$

$X = 0.85m$

$x = 25mm$

$x = 0.25m$

likelihood

$$\ell(\beta) + \lambda \beta^2$$

Prediction vs Explanation

Bayes Classifiers

Naive Bayes

Linear Discrim Analysis

k Nearest Neighbors

Learning Outcomes

- Exponential Family of Distributions
- Generalized Linear Models

g ?

Exponential Family of Distributions

$$E(Y) = g^{-1}(X^T\beta)$$

Exponential Family of Distributions

An exponential family of distributions are random variables that allow their probability density function to have the following form:

$$f(y;\theta,\phi) = a(y,\phi)\exp\left\{\frac{y\theta-\kappa(\theta)}{\phi}\right\}$$

$a(y\ \phi) =$ normalizing constant

$\phi =$ dispersion parameter function

$\theta =$ Canonical link parameter/function

$\kappa(\theta) =$ log-cumulant function

Canonical Parameter

The canonical parameter represents the relationship between the random variable and the $E(Y) = \mu$

$$\eta = X^T \beta = g(\mu)$$

Normal Distribution

$f(x) = a(x;\phi)\, e^{\frac{x\theta - k(\theta)}{\phi}}$

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$E(X) = \mu$

$a(x;\phi) \propto \frac{1}{\sqrt{2\pi\sigma^2}}$ $\phi = \sigma^2$

$f(x) = a(x;\phi)\, e^{\frac{x\theta - k(\theta)}{\phi}}$

$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}\left[x^2 - 2x\mu + \mu^2\right]}$

$e^{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}}$

$f = a(\cdot)\, e^{\frac{x\theta - k(\theta)}{\phi}}$

$\frac{1}{\sqrt{2\pi\sigma^2}}\, e$

$a(x;\phi) = \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \cdot e^{\frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}}$$

$$f = a(\cdot)\, e^{\frac{x\theta - b(\theta)}{\phi}}$$

$$\frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \quad e^{\frac{x\mu - \mu^2/2}{\sigma^2}}$$

$$E(y) = \beta_0 + \beta_1 x \dots$$

$$\theta = \mu = X^T\beta \qquad E(y) = X^T\beta$$

$$E(y) = X^T\beta$$

Binomial Distribution

$$f_{(x)} = a(x;\phi)\, e^{\frac{x\theta - k(\theta)}{\phi}}$$

$$f(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$E(y) = p$$

$$a(x;\phi) \propto \binom{n}{x}$$

$$y = e^{\ln y}$$

$$\binom{n}{x} e^{\ln\left(p^x (1-p)^{n-x}\right)}$$

$$\binom{n}{x} e^{\ln(p^x) + \ln\left[(1-p)^{n-x}\right]}$$

$$f = a(\cdot)\, e^{\frac{x\theta - k(\theta)}{\phi}}$$

$$\binom{n}{x} e^{x\ln p + (n-x)\ln(1-p)}$$

$$\binom{n}{x} e^{x(\ln p - \ln(1-p)) + n\ln(1-p)}$$

$$\binom{n}{x} e^{\frac{x\ln\left(\frac{p}{1-p}\right) + n\ln(1-p)}{1}}$$

$$f = a(\cdot)\, e^{\frac{x\theta - k(\theta)}{\phi}}$$

$$\binom{n}{x} e^{\frac{x\ln\left(\frac{p}{1-p}\right) + n\ln(1-p)}{1}}$$

$$\phi = 1$$

logit function $\rightarrow \ln\left(\frac{p}{1-p}\right) = X^T\beta$

$E(Y)$

odds

$$\frac{(P=1)}{(P=0)} \quad \frac{\text{success}}{\text{failure}}$$

$$E(Y) = p = \frac{e^{X^T\beta}}{1+e^{X^T\beta}}$$

Logistic Regression

$Y = {1 \atop 0}$

$X: X_1, \ldots, X_p$

Poisson Distribution

$f(x) = a(x;\phi)\, e^{\frac{x\theta - k(\theta)}{\phi}}$

$$f(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

$E(y) = \lambda$

$a \propto \frac{1}{x!} e^{-\lambda} e^{\ln(\lambda^x)}$

$\frac{1}{x!} e^{-\lambda + \ln(\lambda^x)} = \frac{1}{x!} e^{\frac{x\ln(\lambda) - \lambda}{1}}$

$\theta = \ln(\lambda) = X^T\beta$

Common Distributions and Canonical Parameters

Random Variable	Canonical Parameter	θ
Normal	μ	Identity link
Binomial	$\log\left(\frac{\mu}{1-\mu}\right)$	logit
Negative Binomial	$\log\left(\frac{\mu}{\mu+k}\right)$	logit
Poisson	$\log(\mu)$	log ln
Gamma	$-\frac{1}{\mu}$	inverse
Inverse Gaussian	$-\frac{1}{2\mu^2}$	$\sim$ double inverse $\frac{1}{\mu^2}$

Random Variable	Canonical Parameter

Generalized Linear Models

Generalized Linear Models

A generalized linear model (GLM) is used to model the association between an outcome variable (of any data type) and a set of predictor values. We estimate a set of regression coefficients $\boldsymbol{\beta}$ to explain how each predictor is related to the expected value of the outcome.

$$X^T\beta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$E(Y) = g^{-1}(X^T\beta)$$

Generalized Linear Models

A GLM is composed of a systematic and random component.

Sys: $X^T\beta$

Random: Distribution

Random Component

The random component is the random variable that defines the randomness and variation of the outcome variable.

Systematic Component

The systematic component is the linear model that models the association between a set of predictors and the expected value of Y:

$$g(\mu) = \eta = \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}$$

Function RV $X \sim$ Normal

$Y = X^2$ what is Y?

$X \sim$ Pois

$Var(Y)$ $Y = 2X + 5$

Central Limit Theorem $n \to \infty$

Sampling Distribution $\bar{X}$? S^2 ?

$X \sim$ Normal

MLE $\leftarrow$

Linear Regression $\leftarrow$ no math

GLM's Link Functions.

Estimation $\leftarrow$