Generalized Linear Model

K. Gibert[(1)]

[(1)]*Department of Statistics and Operation Research*

Knowledge Engineering and Machine Learning group at
Intelligent Data Science and Artificial Intelligence research center
Universitat Politècnica de Catalunya, Barcelona

Statistical Modelling

Data= Fit+Error

- Fit:
  - Structural
  - Law governing the phenomenon
  - Analytic Function
- Error:
  - Random
  - Variability arround Fit (null expectation)
  - Probabilistic model

Statistical models

- Determine the family of fits:
  - Linear
  - Quadratic
  - Exponential
  - .....

- Determine the law of error:
  - Normal
  - Poisson
  - Binomial....

Generalized Linear Model (GLMz)

Enlarges the scope of General Lineal Model

Using complex dependent variables that are

Convenient transformations of genuine response variable

Reduces to a linear model in the transformed space complex models

General Linear Model

- *Formalization:*

I=i:n observations

Y: Response variable

X_1 ..... X_K *: ExplanatoryVariables*

Find β_0... β_K such that

$$N = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_K X_K + \varepsilon$$

Where N is a transformation of Y such that

The LINK FUNCTION

$$\eta_i = g(\mu_i)$$

, being $\eta_i = E(N|X_i)$ and $\mu_i = E(Y|X_i)$

Generalized Linear Model

- Family of models based on :
  - Response variable (Y): Continuous, binary or qualitative
  - Variable to be predicted: Num transformation of Y
  - Explanatory variables: Continous or Categorical
  - Prediction Model: Linear
  - Estimation method: max likelihood in transformed space

- Particular cases of General Linear Model Family
  - Generalized lineal model (all particular cases)
  - Logistic regression
  - Loglinear models
  - Poisson regression
  - Multinomial regression

Matricial formulation

Matricial notation: $\eta = \mathbf{X}\beta$ $\qquad g(\mu_i) = x_i^T \beta$

$$\underbrace{\begin{bmatrix} \eta_1 \\ \vdots \\ \eta_n \end{bmatrix}}_{\boldsymbol{\eta}} = \underbrace{\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}}_{\boldsymbol{\beta}}$$

Solution: Solve a general linear model for a transformed Y

General Linear model

A particular case of Generalized linear model

- *The link function is the IDENTITY*

  – *Response variable Y continuous*

  – *Predicted variable N=Y*

  – *Link function* $\eta_i = g(\mu_i) = \mu_i$

  – *Error distribution* $\varepsilon \sim \mathcal{N}(0,\sigma)$

Canonical link

Canonical link

A special link function with good properties

We will see later

Logistic Regression model

A particular case of Generalized linear model

- *The link function is*

  – *Response variable Y: binary*

  – *Predicted variable* $N = \ln\left[\frac{P(Y|X)}{1-P(Y|X)}\right]$

  – *Link function: logit* $\eta_i = g(\mu_i) = \frac{\mu_i}{1-\mu_i}$

  Canonical link

  – *Error distribution* $\varepsilon \sim \text{Bin}(n,\pi)$

Alternative models for binomial

A particular case of Generalized linear model

The link function is

- *Response variable Y: binary*
- *Link function: probit* $\eta_i = g(\mu_i) = \Phi^{-1}(\mu_i)$, $\Phi^{-1} = F^{-1}(\mathcal{N}(0,1))$
- *Link function: log-log complementary*

$$\eta_i = g(\mu_i) = \log(-\log(1-\mu_i))$$

- *Error distribution* $\varepsilon \sim \text{Bin}(n,\pi)$

Poisson Regression model

A particular case of Generalized linear model

- *The link function is*

  – *Response variable: Y count*

  – *Predicted variable* $N = \ln[Y \mid X]$

  – *Link function* $\eta_i = g(\mu_i) = log(\mu_i)$

Canonical link

  – *Error distribution* $\varepsilon \sim \text{Poisson}(\lambda)$

Estimation of the model

Goodness of fit criterion: Deviance

$$D'(\mathrm{y},\mu) = 2\,\ell(\mathrm{y},\mathrm{y}) - 2\,\ell(\mu,\mathrm{y})$$

– *Maximize the likelihood (the probability that the given parameters enerate the observed data)*

$$\ell(\mu,\mathbf{y}) = \sum_{i=1}^{n} \log f(y_i,\theta_i)$$

– *f: joint distribution between observed data and model parameters*

– *The function to maximize has different expressions depending on the distribution of Y*

Estimation of the model

General equation to be solved

$$S(\beta)=D\Sigma^{-1}(Y-g(X^T\beta)) = D\Sigma^{-1}(Y-Y\hat{})$$

with $D=\frac{\partial}{\partial\beta}g(X^T\beta)$, $\Sigma=Var(Y)$

For canonical links $D=\Sigma$ and

$$S(\beta)= (Y-g(X^T\beta))$$

Expected and observed information are equal

Estimate procedures for parameters and variances simplify

Validation

- Goodness of fit

Deviance $$D'(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\,\ell(\mathbf{y}, \phi, \mathbf{y}) - 2\,\ell(\hat{\boldsymbol{\mu}}, \phi, \mathbf{y})$$

Null model: all coefficients non signifficant, prediction equal to independent term. Constant model. All X useless

Complete (maximal) model: one parameter for each observed value, perfec prediction

Deviance compares maximal model with obtained model

Perfect model: Deviance=0 (H_0), test statistics χ^2

Significat: Deviance too big : Model invalid

Deviance for different models

- Normal distribution $$D(\mathbf{y},\hat{\mu})=\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2$$
- Poisson dist. $$D'(\mathbf{y},\hat{\mu})=D(\mathbf{y},\hat{\mu})=2\sum_{i=1}^{n}\left\{y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right)-(y_i-\hat{\mu}_i)\right\}$$
- Binomial dist. $$-2\sum_{i=1}^{n}\left(y_i\log p_i+(1-y_i)\log(1-p_i)\right)$$