Statistical Data Analysis

Dr. Jana de Wiljes

15. Dezember 2021

Universität Potsdam

Generalized Linear Models

Setting: $y_i = x_i^\top \beta + \epsilon, \quad i = 1, \dots, n$

Up till now:

- $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$,
- $\mu = x_i^\top \beta$, $i = 1, \dots, n$
- $\mu_i = \mathbb{E}[y_i | x_i]$

Assumtion:

$$f(y_i | x_i, \theta_i, \phi, w_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi, w_i)\right) \tag{1}$$

where

- θ_i is the natural parameter of the family,
- ϕ is scale or dispersion parameter,
- $b(\cdot)$ $c(\cdot)$ $a(\cdot)$ are specific functions corresponding to the type of the family

How does the normal distribution fit in the picture?

$$
\begin{aligned}
f(y_i|x_i, \theta_i, \phi) &= \frac{1}{\sqrt{2\pi\sigma}} \exp\Big(-(y_i - \mu)^2/2\sigma^2 \Big) \\
&= \exp\Big(\frac{y_i\mu - \mu^2/2}{\sigma^2} - (y_i^2/\sigma^2 + \log(2\pi\sigma^2))/2 \Big)
\end{aligned}
$$

i.e.,

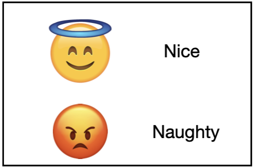
- $\theta_i = \mu$
- $a(\phi) = \sigma^2$
- $b(\theta_i) = \theta_i^2/2$
- $c(\cdot) = -\frac{1}{2}(y_i^2/\sigma^2 + \log(2\pi\sigma^2))$

Mean and Variance?

Mean and Variance?

Link mean to linear predictor

Naughty or nice ?



Elves need system to determine if child naughty or nice

Idea:

- Blood measurements of
  1. Serontonixi
  2. Oximontiuous
- observation of a test group of kids for a year to identify label:
  1. naughty
  2. nice

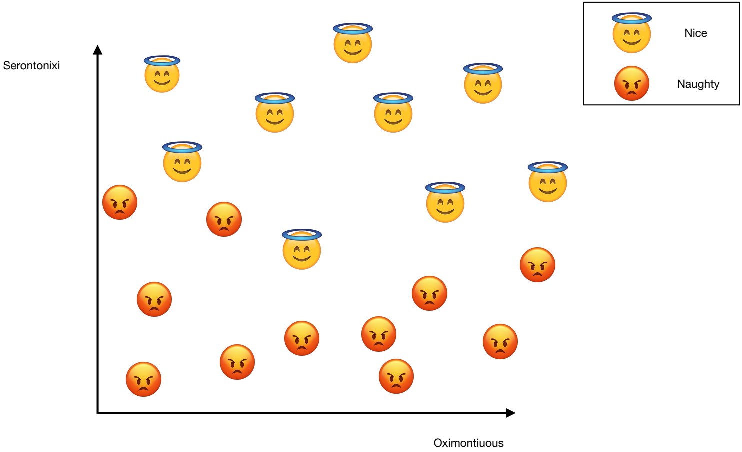
Logistic regression

Logistic regression

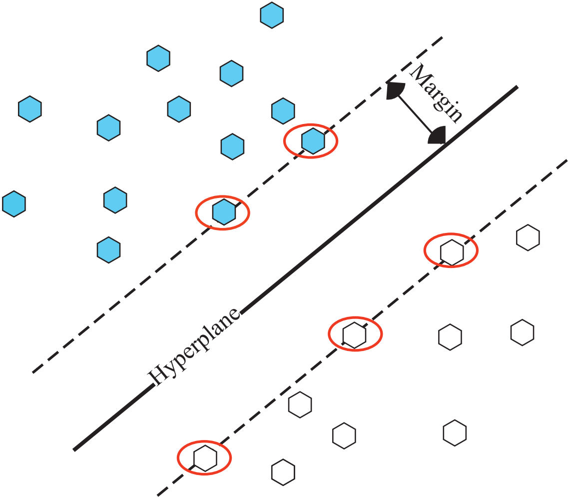
Different variations

	Normal	Poisson	Binomial	Gamma
Notation	$N(\mu_i, \sigma^2)$	$\text{Pois}(\mu_i)$	$\text{Bin}(n_i, \pi_i)$	$G(\mu_i, \nu)$
Range of y_i	$(-\infty, \infty)$	$[0, \infty)$	$[0, n_i]$	$(0, \infty)$
Dispersion, ϕ	σ^2	1	$1/n_i$	ν^{-1}
Cumulant: $b(\theta_i)$	$\theta_i^2/2$	$\exp(\theta_i)$	$\log(1 + e^{\theta_i})$	$-\log(-\theta_i)$
Mean function, $\mu(\theta_i)$	θ_i	$\exp(\theta_i)$	$1/(1 + e^{-\theta_i})$	$-1/\theta_i$
Canonical link: $\theta(\mu_i)$	identify	log	logit	reciprocal
Variance function, $V(\mu_i)$	1	μ	$\mu(1 - \mu)$	μ^2

Naughty or nice?



Support vector machine



Support vector machine