1 GLM: Logistic regression

Logistic regression: A model predicts if the patient is a vegan ($y = 1$) or not ($y = 0$) by a result of cholesterol test and gets the result x mmol/L. A binary response modeled with Bernoulli distribution $y \sim \mathcal{B}(p)$, where $p := \mathbb{P}[y = 1]$ is the probability of being a vegan.

Bernoulli distribution belongs to the exponential family; in canonical form, it is:

$$y \sim f(y|\theta) = (1 - \sigma(\theta)) \cdot e^{\theta \cdot y},$$

where $\theta = \operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right)$ is the logit function.

Fitting the parameter θ on the historical data:

$$\theta^* = \arg\max_{\theta} \sum_{(x^*, y^*) \in (X,Y)^\ell} f(y = y^* | x = x^*, \theta)$$

Then, to make a prognosis, the probability of being a vegan is calculated:

In logistic regression, the connection between the input $\boldsymbol{x}$ and the probability p is modeled as:

$$\underbrace{\log \frac{p}{1-p}}_{\operatorname{logit} p} = \beta_0 + \beta_1 \cdot x$$

The inverse of the logit function is the sigmoid function $\sigma(\theta) = \left(1 + e^{-\theta}\right)^{-1}$:

$$p = \operatorname{logit}^{-1}(\beta_0 + \beta_1 \cdot x) = \sigma(\beta_0 + \beta_1 \cdot x)$$

The parameters β_0 and β_1 are trained on the historical data:

$$\ell(\beta_0, \beta_1) = \log \prod_i p_i^{y_i} (1 - p_i)^{1 - y_i} \to \max_{\beta_0, \beta_1}.$$

Then, to make a prognosis, the probability of a bad outcome is calculated:

$$\hat{p}(\boldsymbol{x}) = \sigma(\beta_0 + \beta_1 \cdot x).$$

NB: We used distribution of $y \sim \mathcal{B}(p)$ to derive the loss function:

$$\ell = \log \prod_i \mathbb{P}[y_i = 1]^{y_i} \cdot \mathbb{P}[y_i = 0]^{1 - y_i},$$

but we ignore the distribution of y when we make a prediction, we are only interested in

Exponential family: Bernoulli distribution $\mathcal{B}(p)$ belongs to the exponential family, the exponential family parameter θ can be calculated from the Bernoulli parameter p, then

$$y \sim \operatorname{Exp}(\theta).$$

2 Generalized Linear Models (GLM)

Introduction to GLM: A generalized linear model (GLM) extends ordinary linear regression by allowing for response variables that follow any exponential family distribution. The general form is:

$$Y \sim f(y \mid \boldsymbol{\theta}) = \exp[\boldsymbol{\theta} \cdot T(y) - A(\boldsymbol{\theta}) + C(y)] \tag{1}$$

Making Predictions: To make a prediction in GLM, we estimate the conditional expectation (canonical mean parameter):

$$\hat{y}(\boldsymbol{x}) := \mathbb{E}[Y \mid X = \boldsymbol{x}] \equiv \mu \tag{2}$$

For most cases, the sufficient statistics T is trivial, and we can obtain the needed expectation from the distribution parameters:

$$\hat{y}(\boldsymbol{x}) := \mathbb{E}[Y \mid \boldsymbol{\theta}] = \mathbb{E}[Y \mid \boldsymbol{\theta} = F\beta] = \mathbb{E}[Y \mid X = \boldsymbol{x}] \tag{3}$$

For example, in logistic regression, the mean parameter corresponds to probability:

$$\mu := \mathbb{E}[Y \mid \boldsymbol{\theta}] = \mathbb{P}[Y = 1 \mid \boldsymbol{\theta}] = \mathbb{P}[Y = 1 \mid X = \boldsymbol{x}] = \hat{p} \tag{4}$$

Mean and Link Functions:

231 Mean Function

The mean function describes the expected value of the response variable Y (or sufficient statistics T(Y)) given current parameters:

$$\boldsymbol{\mu} := \mathbb{E}[T(\boldsymbol{Y}) \mid \boldsymbol{\theta}] \tag{5}$$

232 Link Function

The link function connects linear parameters ϑ = Xβ (linear predictor) to the expected value (canonical mean):

$$\boldsymbol{\mu} = \psi(\boldsymbol{\theta}) \tag{6}$$

Its inverse calculates parameters:

$$\boldsymbol{\theta} = \psi^{-1}(\boldsymbol{\mu}) \tag{7}$$

GLM as Linear + Nonlinear Transforms: GLM combines linear and nonlinear transformations:

1. Linear predictor computation:

$$\theta(\boldsymbol{x}) = \boldsymbol{\beta}^\intercal \boldsymbol{x} = M\boldsymbol{x} \tag{8}$$

$$\boldsymbol{\theta} = X\boldsymbol{\beta} \tag{9}$$

2. Link function application:

$$\psi : \quad \theta = \psi(\mu) \tag{10}$$

$$\psi(\mathbb{E}[y|\boldsymbol{x}]) = \boldsymbol{\beta}^\intercal \boldsymbol{x} \tag{11}$$

$$\psi(\mathbb{E}[y|X]) = X\boldsymbol{\beta} \tag{12}$$

3. Final prediction via inverse link function:

$$\hat{y}(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{x}] = \psi^{-1}(\boldsymbol{\beta}^\intercal \boldsymbol{x}) \tag{13}$$

Logistic Regression as GLM: Logistic regression is a special case of GLM using Bernoulli distribution:

$$y_i \sim \mathcal{B}(p), \quad \mathbb{P}[y_i = 1] = p \tag{14}$$

In canonical form:

$$y_i \sim f(y \mid \theta) = (1 - \sigma(\theta)) \cdot e^{\theta \cdot y} \tag{15}$$

The link function can be found from:

$$\psi = (\boldsymbol{\nabla}_{\boldsymbol{\theta}} A)^{-1} \tag{16}$$

Where:

$$\frac{\partial A}{\partial \theta} = \sigma(\theta) = \frac{1}{1 + e^{-\theta}} \tag{17}$$

$$\psi(\mu) = \sigma^{-1}(p) = \ln \frac{p}{1 - p} = \text{logit } p \tag{18}$$

3 GLM: Cross-entropy and log-loss

Model: Logistic regression represents a special case of GLM where the binary response variable Y follows a Bernoulli distribution:

$$y_i \sim \mathcal{B}(p), \quad p := \mathbb{P}[y_i = 1] \tag{19}$$

Here, p represents the success probability in a single trial. The canonical form of the Bernoulli distribution is:

$$y_i \sim f(y|\theta) = \sigma(-\theta) \cdot e^{\theta \cdot y}, \quad \sigma(\theta) = \frac{1}{1+e^{-\theta}} \tag{20}$$

Starting from the general GLM form:

$$Y \sim f(y|\boldsymbol{\theta}) = \exp[\boldsymbol{\theta} \cdot T(y) - A(\boldsymbol{\theta}) + C(y)] \tag{21}$$

We can derive both cross-entropy and log-loss directly, assuming only the Bernoulli distribution of Y.

Link Function: The link function ψ connects the response variable's mean $\mu = \mathbb{E}[Y]$ to the distribution's canonical parameters $\boldsymbol{\theta}$:

$$\boldsymbol{\mu} = \psi(\boldsymbol{\theta}) \tag{22}$$

In GLM, we assume the canonical parameters are linear:

$$\theta_i = \boldsymbol{x}_i^\top \beta, \quad \boldsymbol{\theta} = X\boldsymbol{\beta} \tag{23}$$

where $\boldsymbol{\beta}$ represents the linear coefficients corresponding to features in $\boldsymbol{x}$.

For the Bernoulli distribution, the link function takes the form:

$$\psi(\mu) = \log \frac{\mu}{1-\mu} = \text{logit } \mu \tag{24}$$

Cross-entropy Loss: We begin with the log-likelihood function $l(\theta)$ for the Bernoulli-distributed response variable Y, assuming $\theta = \boldsymbol{x}^\top \beta$:

$$\begin{aligned}
l(\theta) &= \log \prod_i f(y_i|\theta) \\
&= \log \prod_i \sigma(-\theta) \cdot e^{\theta \cdot y_i} \\
&= \sum_i \{\theta \cdot y_i + \log \sigma(-\theta)\} \\
&= \sum_i \left\{\theta \cdot y_i + \log \frac{1}{1+e^{-(-\theta)}}\right\} \\
&= \sum_i \left\{y_i \log \frac{\mu}{1-\mu} + \log \frac{1}{1+\frac{\mu}{1-\mu}}\right\} \\
&= \sum_i \left\{y_i \log \frac{\mu}{1-\mu} + \log \frac{1-\mu}{1-\mu+\mu}\right\} \\
&= \sum_i \{y_i \log \mu - y_i \log(1-\mu) + \log(1-\mu)\} \\
&= \sum_i \{y_i \log \mu + (1-y_i) \log(1-\mu)\} \\
&= \sum_i \{y_i \log p + (1-y_i) \log(1-p)\} \\
&= l(p(\boldsymbol{\beta})) \to \max_{\boldsymbol{\beta}}
\end{aligned} \tag{25}$$

Log-loss: The log-loss function $\ell(M)$ can be derived by taking the negative log-likelihood:

$$\begin{aligned}
-l(\theta) &= -\sum_i \left\{\theta \cdot y_i + \log \frac{e^{-\theta}}{1+e^{-\theta}}\right\} \\
&= \sum_i \begin{cases} -\log e^{\theta} + \log \frac{e^{-\theta}}{1+e^{-\theta}}, \text{if } y=1 \\ -\log \frac{e^{-\theta}}{1+e^{-\theta}}, \text{if } y=0 \end{cases} \\
&= \sum_i \begin{cases} \log(1+e^{-\theta}), \text{if } y=1 \\ \log(1+e^{\theta}), \text{if } y=0 \end{cases} \\
&= \sum_i \log(1+e^{\theta \cdot \operatorname{sgn} y_i}) \\
&= \sum_i \log(1+e^{\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle \cdot \operatorname{sgn} y_i}) \\
&= \sum_i \log(1+e^{-M_i}) \\
&= \ell(M(\boldsymbol{\beta})) \to \min_{\boldsymbol{\beta}}
\end{aligned} \tag{26}$$

Making Predictions: To make a prediction:

$$\hat{p}(\boldsymbol{x}) = \mu(\boldsymbol{x}) = \psi(\theta = \boldsymbol{x} \cdot \boldsymbol{\beta}) = \frac{1}{1 + e^{\boldsymbol{x} \cdot \boldsymbol{\beta}}} \tag{27}$$