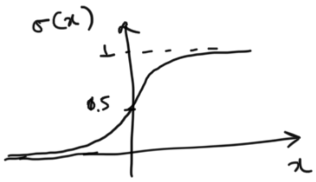
Logistic regression (classification)

Example: $p(s|x)$, $x = (x_1, x_2, x_3)$, x_1: age, x_2: M/F , x_3: chole

The simplest approach would assume some linear model:

(linear regression) $\quad w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 = w^T x$, $x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}$, $w^T = [w_0, w_1, w_2$

... but this is not a probability! We can fix this by "warping" $w^T x$ with a sigmoid function:



$$\left.\begin{array}{l} \sigma(x) = \dfrac{1}{1 + e^{-x}} \\ \text{logistic sigmoid} \end{array}\right\} \quad p(y|x) = \sigma(w^T x)$$

Formal definition:

Setup: Given $\mathcal{D} := \{(x_1, y_1), \ldots, (x_n, y_n)\}$, $x_i \in \mathbb{R}^d$, $y_i \in \{0, 1\}$

Model: $y_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(\underbrace{\sigma(w^T x_i)}_{a_i})$, $\theta := \{w_0, w_1, \ldots, w_d\}$

Pros:
- Interpretable (= model parameters are meaningful)
- Reveal which variables are more influential.
- Small number of trainable parameters $(d+1)$
  ↳ simple model that is "statistically easy to train
- Computationally efficient ways to estimate w.
- Extension to multi-class is straightforward.
- Forms the foundation for more complex models. (GLM, NNs

Cons :
- Being a simple model, its performance is inferior to more complex models.

Maximum Likelihood Estimation :

$$w_{MLE} = \arg\max_{w} \; p(y|X,w)$$

$$a_i := \sigma(w^T x_i)$$

$$p(y_1,\dots,y_n \mid x_1,\dots,x_n, w_0, w_1,\dots,w_d) \overset{i.i.d}{=} \prod_{i=1}^{n} p(y_i \mid x_i, w)$$

$$= \prod_{i=1}^{n} a_i^{y_i}(1-a_i)^{1-y_i}$$

$$\mathcal{L}(w) := -\log p(y|X,w) = -\sum_{i=1}^{n} y_i \log a_i + (1-y_i)\log(1-a_i) \Big\} \text{ Binary cross-entropy loss}$$

- Recall that $a := \sigma(w^T x)$

$$\rightarrow \quad \log a = -\log(1+e^{-w^T x})$$

$$\rightarrow \quad \log(1-a) = -w^T x - \log(1+e^{-w^T x})$$

$$\rightarrow \quad \frac{\partial}{\partial w_i}\log a = x(1-a)$$

$$\rightarrow \quad \frac{\partial}{\partial w_i}\log(1-a) = -ax$$

$$\Longrightarrow \quad \underset{1\times 1}{\nabla_{w_j}\mathcal{L}(w)} = -\sum_{i=1}^{n} y_i x_{ij}(1-a_i) - (1-y_i)x_{ij}a_i$$

$$= \sum_{i=1}^{n}(a_i - y_i)x_{ij}$$

$$\boxed{\underset{(d+1)\times 1}{\nabla_w \mathcal{L}(w)} = \underbrace{X^T(a-y)}}, \quad \underset{n\times(d+1)}{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & \vdots & & \\ 1 & x_{n1} & \cdots & x_{nd} \end{bmatrix}, \quad a = \begin{bmatrix} \sigma(w^T x_1) \\ \vdots \\ \sigma(w^T x_n) \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \end{bmatrix}$$

$$\boxed{\nabla_w^2 \mathcal{L}(w) = X^T A X}, \quad \underset{n\times n}{A} := \begin{bmatrix} a_1(1-a_1) & & 0 \\ & \ddots & \\ 0 & & a_n(1-a_n) \end{bmatrix}$$

$\hookrightarrow$ symmetric PSD

<u>Newton's methods</u>: (<u>Iterative</u> <u>re-weighted</u> <u>least squares</u>)

$$w_{t+1} = w_t - \eta H_t^{-1} g_t, \quad H_t := X^T A_t X, \quad g_t := X^T (a_t - y)$$

$$\Rightarrow w_{t+1} = w_t - (X^T A_t X)^{-1} X^T (a_t - y)$$

Rewrite as:

$$w_{t+1} = (X^T A_t X)^{-1} X^T A_t \overbrace{\left[X w_t - A_t^{-1}(a_t - y) \right]}^{z_t}$$

$$\Rightarrow w_{t+1} = \boxed{(X^T A_t X)^{-1} X^T A_t z_t}$$ (Newton for logistic)

Recall Linear regression: $\underline{W_{MLE}} = (\underbrace{X^T X}) ^{-1} X^T y = \overbrace{(X^T A X)^{-1} X^T A y}$,

where $\underline{A := I}$.

<u>Multi-class logistic regression</u>:

$\sum [0 \ 0.8 \ 0 \ 0.2] = 1$

<u>Model</u>: $\underbrace{p(y = c \mid X, w)}_{c = 1, \dots, q \ \hookrightarrow \# \text{ of classes}} = \dfrac{\exp(w_c^T x)}{\sum_{c'=1}^{q} \exp(w_{c'}^T x)}$ } soft-max is a generalizati of the logistic sigmoid for multiple classes.

where w_c is the c-th column of $\underset{(d\times 1)\times q}{W}$,

and y is a "one-hot" encoding vector:

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

$n \times c$

$$\underline{\underline{y_{ic}}} = \mathbb{1}_{\{y_i = c\}} = \begin{cases} 1, & \text{if } y_i = c \\ 0, & \text{if } y_i \neq c \end{cases} \qquad \begin{array}{c} \left[\qquad\qquad \right] \\ c \end{array}$$

$$p(y \mid X, w) = \prod_{i=1}^{n} \prod_{c=1}^{C} p(y_i = c \mid x_i, w)$$

$$\Rightarrow \quad -\log p(y \mid X, w) = \underbrace{\sum_{i=1}^{n} \left[\left(\sum_{c=1}^{C} y_{ic} w_c^T x_i \right) - \log \left(\sum_{c'=1}^{C} \exp(w_{c'}^T x_i \right. \right.}_{\text{multi-class cross entropy.}}$$