Logistic regression

- Classification setup
- Logistic regression is the classification counterpart of linear regression.
  - we focus on binary classification, i.e. we assume
  $$x \in \mathbb{R}^d, \quad \mathcal{Y} = \{yes, no\}$$

For example,
- the two classes can be associated with whether an individual will default on their credit card payments.
  - The attributes may be the person's annual income & monthly credit card balance
    - ($d = 2$ in this case).

Probability prediction

- The default labels fall into one of the two categories, i.e. $Y \in \{yes, no\}$
- Idea: rather than modelling Y directly, logistic regression models
  $$p(x) = \text{Prob}(Y = yes \mid X).$$
  i.e. the probability that Y is yes (given x)
- Let $\tilde{x} = [1, x]^T$ with $x \in \mathbb{R}^d$, then

$$F(x, \lambda) = p_\lambda(x) = \sigma(\tilde{x}^T \lambda) = \sigma\left(\lambda_0 + \sum_{i=1}^{d} \lambda_i x_i\right)$$

Where; $\sigma : \mathbb{R} \to [0, 1]$ is the logistic function.

The sigmoid function

- The logistic function

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}}$$

- is nonnegative and bounded.
- is monotone
  i.e. its derivative $\sigma'(x) = \sigma(x)(1-\sigma(x))$
  - is positive everywhere
- maps from the real line to the interval $[0,1]$ & it can be naturally interpreted as a probability

- $\lim_{x \to \infty} \sigma(x) = 1$ and $\lim_{x \to -\infty} \sigma(x) = 0$

Odds

- The odds of the conditional probability
  $Prob(Y = yes | X) = p_\lambda(X)$
  are defined as
  $$\frac{p_\lambda(X)}{1 - p_\lambda(X)} = e^{\vec{x}^T \lambda}$$
  and their logarithms,
  i.e. the log-odds of $p_\lambda(X)$, are linear in the attribute
  i.e.
  $$\log \frac{P(X)}{1 - P(X)} = \vec{X}^T \lambda$$
- Note that: $1 - p_\lambda(X) = Prob(Y = no | X)$

Maximum likelihood estimation

- Let $D = \{(x_n, y_n) \in \mathbb{R}^d \times \{yes, no\}\}_{n=1}^{N}$ be a training data set.
- The interpretation of $F(X, \lambda) = \sigma(\vec{x}^T \lambda) \in [0, 1]$
  - as the conditional probability of observing $Y = yes$ given X
  - allows you to estimate λ by
  - maximizing the likelihood of D.

The ML estimate on D is the parameter that maximises the probability of observing D.

Conditional likelihood maximisation

- As the attributes are assumed to be always known,
  - we focus on maximizing the probability of observing the labels $y_1, \ldots, y_N \in D$

⇒ More precisely, we let
$$\hat{\lambda} = \arg\max_\lambda L(D, \lambda)$$
where;
$$L(D, \lambda) = Prob(y_1, \ldots, y_N | x_1, \ldots, x_N, \lambda).$$
$$= \prod_{i=1}^{N} 1[y_i = yes] F(x_i, \lambda) + 1[y_i = no](1 - F(x_i, \lambda))$$

Log-likelihood minimization

- An numerically easier but equivalent problem
  - is to minimize the negative of the logarithm of $\mathcal{L}(\mathcal{D}, \lambda)$,

  i.e. to let

$$\hat{\lambda} = \arg\min_{\lambda \in \mathbb{R}^{d+1}} \ell(\mathcal{D}, \lambda) = -\log\left(\mathcal{L}(\mathcal{D}, \lambda)\right)$$

$$\hat{\lambda} = -\log\left(\prod_{i=1}^{N} \mathbb{1}[y_i = yes] F(x_i, \lambda) + \mathbb{1}[y_i = no](1 - F(x_i, \lambda))\right)$$

- In particular, using standard property of the logarithm we have

$$\ell(\mathcal{D}, \lambda) = \sum_{i=1}^{N} \log\left(y_i F(x_i, \lambda) + (1 - y_i)(1 - F(x_i, \lambda))\right).$$

where;
- the representation numerical of the nominal variables, i.e. $\{yes, no\} \rightarrow \{1, 0\}$
- allows us to rewrite the i^{th} factor in $\mathcal{L}(\mathcal{D}, \lambda)$ as

$$y_i F(x_i, \lambda) + (1 - y_i)(1 - F(x_i, \lambda))$$

- which is differentiable on λ.

Example :