Logistic Regression

- Consider a data set $\{(x_n, t_n)\}$, $t_n \in \{0, 1\}$, $n = 1, \ldots, N$.

  Example: Test (GRE) scores and admission to a graduate school

x	272	331	295	287	315	266	303	294	317	309
t	0	1	1	0	1	0	0	0	1	1

$t = 1$ accepted; $t = 0$ rejected

If $x = 299$, what is the **probability** to be accepted?

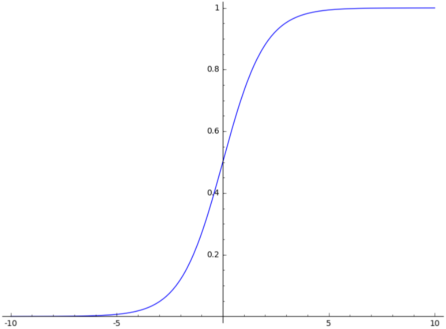
- This problem is called logistic regression.

- Idea: Transform x_n into probability y_n of admission so that $t_n = 1$ with probability y_n.
  In other words, y_n is the probability of success for score x_n.

- We need a function from $(-\infty, \infty)$ to $(0, 1)$.
  Use the logistic sigmoid function

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

Graph of sigmoid



- First, set $\boldsymbol{x}_n := [x_n, 1]$ and

$$a_n := w_1 x_n + w_2 = \boldsymbol{x}_n \boldsymbol{w}$$

  for some (unknown) $\boldsymbol{w} := [w_1, w_2]^\top$.

  This is exactly a linear regression.

- Second,

$$y_n := p(t_n = 1 | x_n) = \sigma(a_n).$$

- Third, each result is determined by a Bernoulli trial.

$$x_n \rightsquigarrow y_n \rightsquigarrow t_n = 0, 1$$

Let T be a Bernoulli random variable.

$$\Pr(T = 1) = y \quad \text{and} \quad \Pr(T = 0) = 1 - y$$

We consider y as the probability of success.

- The probability mass function $\mathrm{Ber}(\cdot|y)$ is given by

$$\mathrm{Ber}(1|y) = y \quad \text{and} \quad \mathrm{Ber}(0|y) = 1 - y.$$

- We write

$$T \sim \mathrm{Ber}(t|y) = y^t (1-y)^{1-t}, \quad t = 0, 1.$$

- $\mathrm{E}(T) = y \quad \text{and} \quad \mathrm{Var}(T) = y(1-y)$

- data set: $\{(x_n, t_n)\}$, $t_n \in \{0, 1\}$, $n = 1, \ldots, N$

$$\begin{array}{ccccc} \text{(score)} & \rightsquigarrow & \text{(probability of success)} & \rightsquigarrow & \text{(success or failure)} \\ x_n & \rightsquigarrow & y_n = \sigma(w_1 x_n + w_0) & \rightsquigarrow & t_n \end{array}$$

- Assume that the test scores are independent.

  Each score brings about a Bernoulli random variable.

  The likelihood function can be written

$$p(\mathbf{t}|\boldsymbol{w}) = \prod_{n=1}^{N} \mathrm{Ber}(t_n|y_n) = \prod_{n=1}^{N} y_n^{t_n}(1 - y_n)^{1-t_n},$$

  where $\mathbf{t} = (t_1, \ldots, t_N)$.

- Task: Determine $\boldsymbol{w} = [w_1, w_2]^\top$ to obtain maximum likelihood.

- We want to maximize

$$\ln p(\mathbf{t}|\boldsymbol{w}) = \sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\},$$

or equivalently, we want to minimize

$$\boxed{E(\boldsymbol{w}) = -\ln p(\mathbf{t}|\boldsymbol{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}}.$$

It is called the cross-entropy error function.

- How can we minimize this error function?

We can use Gradient Descent or Newton's Method.

k features

- Data set: $\{(x_{n,1}, x_{n,2}, \ldots, x_{n,k}; t_n)\}$, $t_n = 0, 1$, $n = 1, 2, \ldots, N$

  Example

  $x_{n,1}$ = GRE score, $x_{n,2}$ = GPA, ...
- Set $\boldsymbol{x}_n := [x_{n,1}, x_{n,2}, \ldots, x_{n,k}, 1]$ and

$$a_n := w_1 x_{n,1} + w_2 x_{n,2} + \cdots + w_k x_{n,k} + x_{k+1} = \boldsymbol{x}_n \boldsymbol{w}$$

  for some (unknown) $\boldsymbol{w} := [w_1, w_2, \ldots, w_{k+1}]^\top$.
- Set $y_n = \sigma(a_n)$ and

$$E(\boldsymbol{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}.$$

- We have

$$\nabla E(\boldsymbol{w}) = \left[\sum_{n=1}^{N} (y_n - t_n) x_{nj} \right] = X^\top (\boldsymbol{y} - \mathbf{t}),$$

where

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1k} & 1 \\ x_{21} & \cdots & x_{2k} & 1 \\ \vdots & & \vdots & \vdots \\ x_{N1} & \cdots & x_{Nk} & 1 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \text{and } \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}.$$

- Gradient Descent

$$\boxed{\boldsymbol{w}_{i+1} = \boldsymbol{w}_i - \eta X^\top (\boldsymbol{y} - \mathbf{t})}.$$

- $$\mathbf{H}E = \left[\sum_{n=1}^{N} y_n(1-y_n)x_{ni}x_{nj}\right] = X^\top R X,$$
  where $R = \mathrm{diag}(y_n(1-y_n))$.
- Newton's Method
  $$\boxed{\boldsymbol{w}_{i+1} = \boldsymbol{w}_i - \eta(X^\top R X)^{-1}X^\top(\boldsymbol{y}-\mathbf{t})},$$
  where R and $\boldsymbol{y}$ are determined by $\boldsymbol{w}_i$ in each step.

Example (continued): GRE scores and admission

- Choose $k = 1$. Write $a_n = w_1 x_n + w_2$.

x	272	331	295	287	315	266	303	294	317	309
t	0	1	1	0	1	0	0	0	1	1

- Choose $\boldsymbol{w}_0 = [0, 0]^\top$. Then $\boldsymbol{w}_i$ converges to

$$[0.1910, -57.2937]^\top.$$

- If $x = 299$, then the probability of admission is

$$y = \sigma(0.1910 * 299 - 57.2937) \approx 0.454.$$