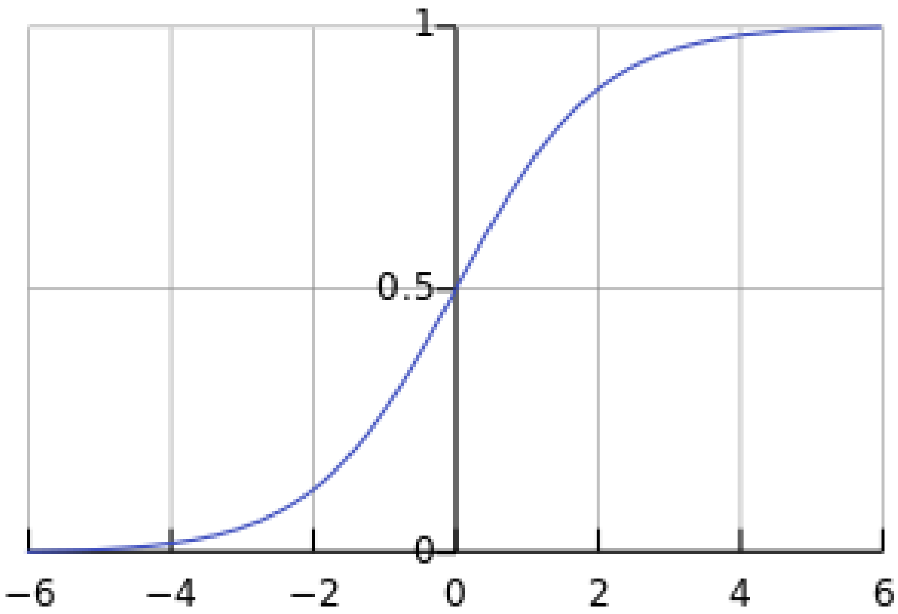
Logistic Regression

Idea

- Similar to perceptron, but sigmoid function applied to linearity.



$$S(\mathbf{w}^T\mathbf{x}) = S(z) = \frac{1}{1+e^{-z}}$$

Probability Interpretation

- Assume instances (**x**, *y*) are generated from some noisy data-source according to some distribution.

$$p(y \mid \mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{for } y = 1 \\ 1 - f(\mathbf{x}) & \text{for } y = -1 \end{cases}$$

- Will try to learn $f(\mathbf{x})$ by approximating it with S(z),

  i.e. S($\mathbf{w}^T\mathbf{x}$)

$$\mathbf{w} = [w_0 = b, w_1, .., w_m] \qquad \mathbf{x} = [x_0 = 1, x_1, .., x_m]$$

- Makes sense as S(z) is a function from 0 to 1.

$$p(y=\pm 1 \mid \mathbf{x})$$

- What is the probability of an instance to have *y*=1?

  (according to our approximation)

$$p(y=1 \mid \mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$$

- What is the probability of an instance to have *y*=-1?

  (according to our approximation)

$$p(y=-1 \mid \mathbf{x}) = 1 - \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$$

$$= \frac{1+e^{-\mathbf{w}^T\mathbf{x}}-1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$$

$$= \frac{e^{-\mathbf{w}^T\mathbf{x}}}{1+e^{-\mathbf{w}^T\mathbf{x}}}$$

$$= \frac{e^{-\mathbf{w}^T\mathbf{x}}/e^{-\mathbf{w}^T\mathbf{x}}}{1/e^{-\mathbf{w}^T\mathbf{x}}+e^{-\mathbf{w}^T\mathbf{x}}/e^{-\mathbf{w}^T\mathbf{x}}}$$

$$= \frac{1}{1+e^{\mathbf{w}^T\mathbf{x}}}$$

- Combining the two eq. on the right, we get:

$$p(y=\pm 1 \mid \mathbf{x}) = \frac{1}{1+e^{-y\mathbf{w}^T\mathbf{x}}}$$

Maximum likelihood

- Find **w** that gives the greatest likelihood (probability) of producing the given training data.
  – i.e. maximize **likelihood function**:

$$L(\mathbf{w}) = \prod_{k=1}^{n} p(y^k \mid \mathbf{x}^k)$$

$$= \prod_{k=1}^{n} \frac{1}{1+e^{-y^k \mathbf{w}^T \mathbf{x}^k}}$$

Probability that the data-source will generate y^k given $\mathbf{x}^k$.
Plain lang: Probability the training instances have the class they have. (Assuming the training instances are independent)

Same **w** for all the training instances.

- Maximizing $L(\mathbf{w})$ is the same as maximizing:

$$\ln L(\mathbf{w}) = \sum_{k=1}^{n} \ln\left(\frac{1}{1+e^{-y^k \mathbf{w}^T \mathbf{x}^k}}\right)$$

We are not saying that ln L(**w**) is the same as E(**w**).

- Which is the same as minimizing:

$$E(\mathbf{w}) = \frac{1}{n}\sum_{k=1}^{n} \ln\left(1+e^{-y^k \mathbf{w}^T \mathbf{x}^k}\right)$$

Gradient Descent Algorithm

Initialize **w**=**0**

For t=0,1,2,…do

Compute the gradient $$\nabla_E(\mathbf{w}) = -\frac{1}{n}\sum_{k=1}^{n}\frac{y^k\mathbf{x}^k}{1+e^{y^k\mathbf{w}^T\mathbf{x}^k}}$$

Update the weights $$\mathbf{w} \leftarrow \mathbf{w} - \kappa\nabla_E(\mathbf{w})$$

Iterate with the next step until **w** doesn't change too much

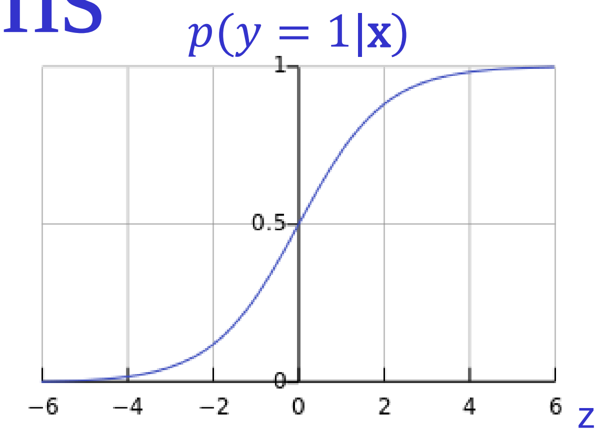
(or for a fixed number of iterations)

Return final **w**.

Making predictions

- A new tuple comes: (**x**,?)

$$p(y=1|\mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$$



- Fix a threshold in [0,1] to make predictions.

$p(y=1|\mathbf{x}) > \text{threshold}$ Predict *y*=1

$p(y=1|\mathbf{x}) \le \text{threshold}$ Predict *y*=-1

Threshold typically is 0.5

Logistic Regression gives a linear separator, namely the hyperplane defined by **w**.
(in 2D this is just a line)

If p(y=1|**x**) = 0.5,
then z $= \mathbf{w}^T\mathbf{x} = 0$
i.e. **x** lies on the hyperplane.

If p(y=1|**x**) > 0.5,
then z $= \mathbf{w}^T\mathbf{x} > 0$
i.e. **x** is above the hyperplane.

If p(y=1|**x**) <= 0.5,
then z $= \mathbf{w}^T\mathbf{x} \le 0$
i.e. **x** is below the hyperplane.

Example

GPA, GRE, and success.

100, 800, 1
90, 800, 1
90, 700, 1
70, 600, -1
60, 700, -1
60, 700, 1
50, 600, -1
50, 650, -1
50, 800, 1
50, 700, -1
50, 700, 1

Dummy	GPA	GRE	y
1	1.0	1.0	1
1	0.9	1.0	1
1	0.9	0.875	1
1	0.7	0.75	-1
1	0.6	0.875	-1
1	0.6	0.875	1
1	0.5	0.75	-1
1	0.5	0.8125	-1
1	0.5	1.0	1
1	0.5	0.875	-1
1	0.5	0.875	1

Normalized data

Example

After many iterations:

w=[7.94, 83.12, – 77.04]

$$p(y=1 \mid gpa, gre) = \frac{1}{1+e^{-(7.93*gpa+83.12*gre-77.04)}}$$

Fix a threshold in [0,1], **e.g. 0.5**, to make predictions.

$p(y=1 \mid gpa, gre) > \text{threshold}$ Predict y=1

$p(y=1 \mid gpa, gre) \leq \text{threshold}$ Predict y=-1

Odds

Definition: $$odds(y \text{ vs. -}y \text{ given } \mathbf{x}) = \frac{p(y \mid \mathbf{x})}{1 - p(y \mid \mathbf{x})}$$

Formula: $$odds(y = 1 \text{ vs. } y = -1 \text{ given } \mathbf{x}) = \frac{\dfrac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}}{1 - \dfrac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}}$$

$$= \frac{1}{e^{-\mathbf{w}^T \mathbf{x}}}$$

$$= e^{\mathbf{w}^T \mathbf{x}}$$

Interpretation

$$odds(\text{successful vs. unsuccessful given } gpa \text{ and } gre) = e^{7.93*\text{gpa}+83.12*\text{gre}-77.04}$$

If GPA increases by .1 (10%)then the odds of success will increase $e^{0.793} \approx 2$ times

If GRE increases by .1 then the odds of success will increase $e^{8.312} \approx 4000$ times

In general the increase of odds is by a factor of:

$$\frac{e^{w_1x_1+...+w_j(x_j+change)+...+w_mx_m}}{e^{w_1x_1+...+w_jx_j+...+w_mx_m}} = e^{change*w_j}$$