

Logistic Regression

Linear Regression

Linear Regression

How to create an approximated function?

Machine learning is the function approximation process

Hypothesis

$$h: \hat{f}(x; \theta) = \theta_0 + \sum_{i=1}^n \theta_i x_i = \sum_{i=0}^n \theta_i x_i$$

n is the number of the feature values

Two aspects: the linearly weight sum(Linear model), the parameter θ

How to find the better θ ?

Finding θ in Linear Regression

$$h: \hat{f}(x; \theta) = \sum_{i=0}^n \theta_i x_i \rightarrow \hat{f}(x; \theta) = X\theta$$

$$X = \begin{pmatrix} 1 & \cdots & x_n^1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_n^D \end{pmatrix}, \theta = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_n \end{pmatrix}$$

But, the reality would be the noisy

$$f(x; \theta) = \sum_{i=0}^n \theta_i x_i + e = y \rightarrow f(x; \theta) = X\theta + e = Y$$

$$\begin{aligned}
\hat{\theta} &= \operatorname{argmin}_{\theta} (f - \hat{f})^2 = \operatorname{argmin}_{\theta} (Y - X\theta)^2 \\
&= \operatorname{argmin}_{\theta} (Y - X\theta)^T (Y - X\theta) \\
&= \operatorname{argmin}_{\theta} (Y^T - \theta^T X^T) (Y - X\theta) \\
&= \operatorname{argmin}_{\theta} (Y^T Y - Y^T X\theta - \theta^T X^T Y + \theta^T X^T X\theta) \\
&= \operatorname{argmin}_{\theta} (Y^T Y - 2\theta^T X^T Y + \theta^T X^T X\theta) \\
&= \operatorname{argmin}_{\theta} (\theta^T X^T X\theta - 2\theta^T X^T Y)
\end{aligned}$$

Now, we need to optimize θ

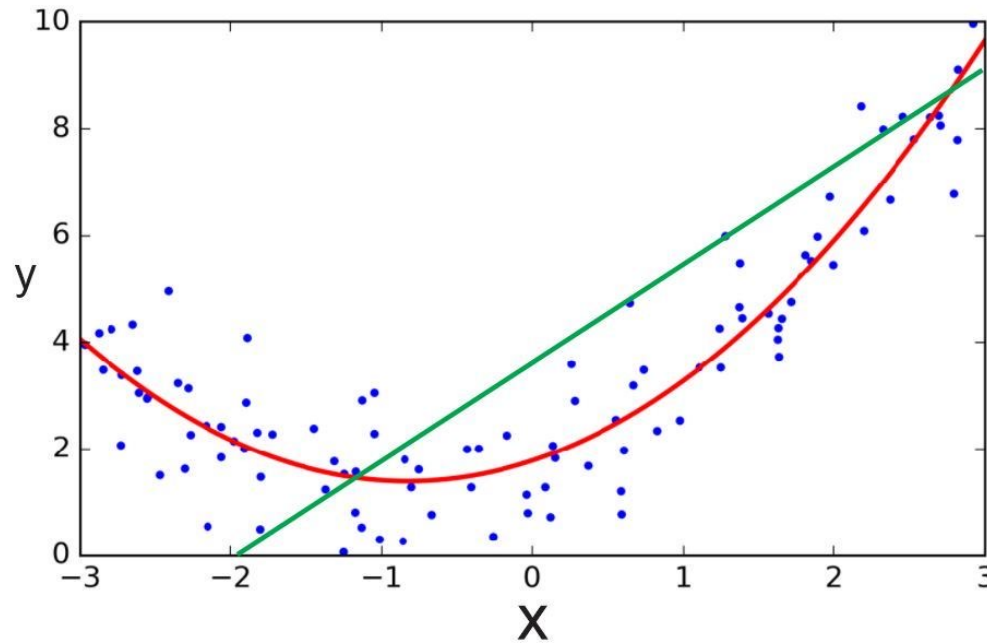
Optimized θ

$$\hat{\theta} = \operatorname{argmin}_{\theta} (\theta^T X^T X \theta - 2\theta^T X^T Y)$$

$$\nabla_{\theta} (\theta^T X^T X \theta - 2\theta^T X^T Y) = 0$$

$$2X^T X \theta - 2X^T Y = 0$$

$$\theta = (X^T X)^{-1} X^T Y$$



$$h: \hat{f}(x; \theta) = \sum_{i=0}^n \sum_{j=0}^m \theta_{i,j} \phi_j(x_i)$$

Simple
Linear
Regression

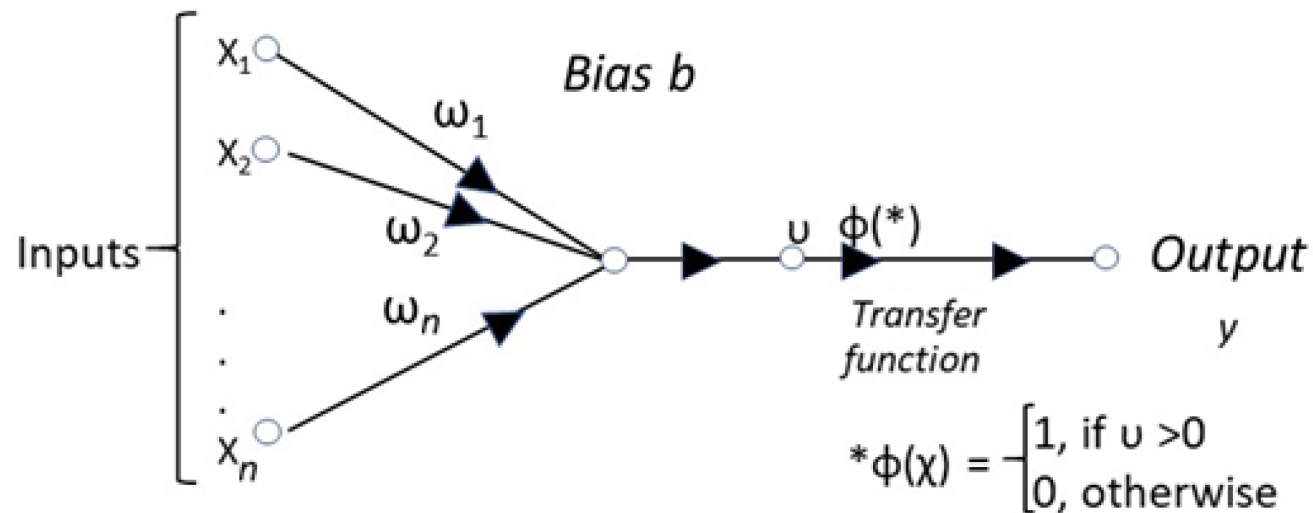
$$y = b_0 + b_1 x_1$$

Multiple
Linear
Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$

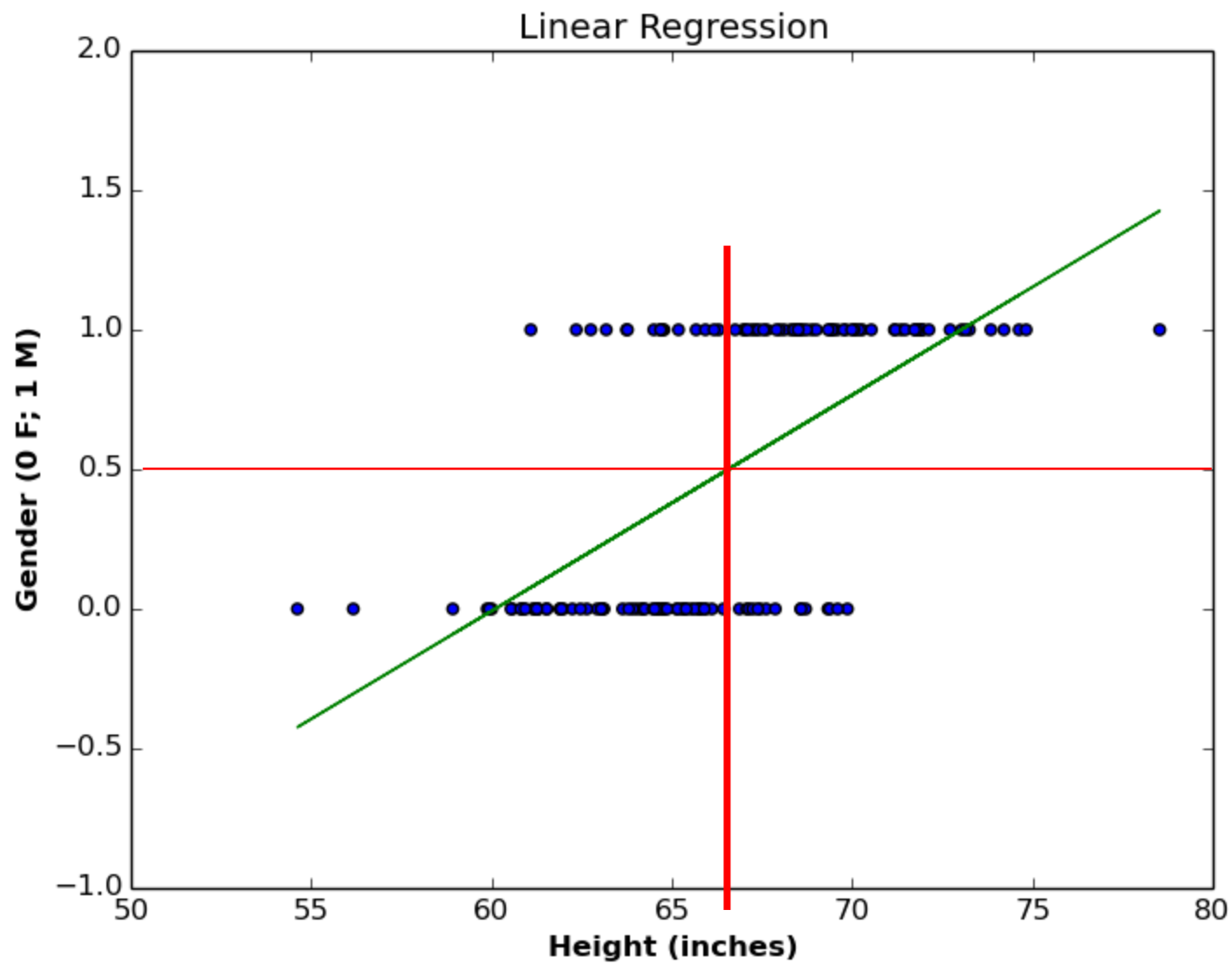
Polynomial
Linear
Regression

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \dots + b_n x_1^n$$



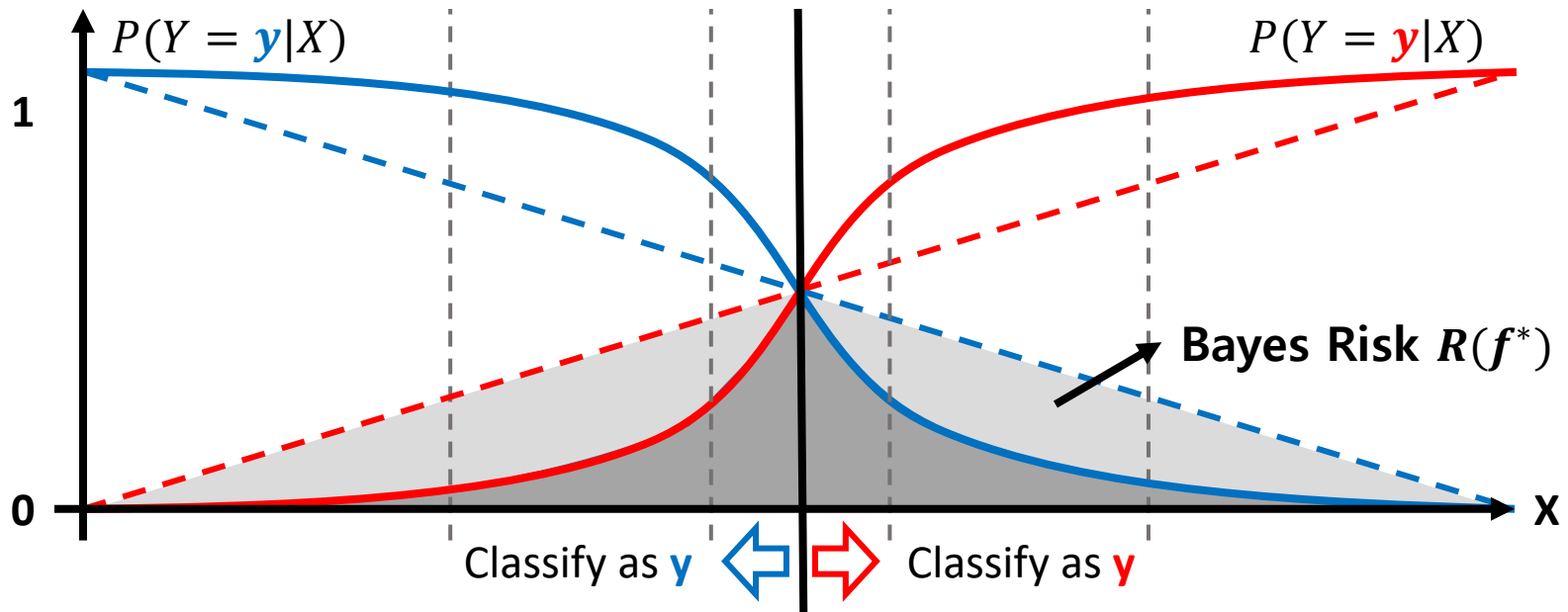
$$X = \begin{bmatrix} x_0 \\ \vdots \\ x_n \end{bmatrix} \quad W = \begin{bmatrix} \omega_0 \\ \vdots \\ \omega_n \end{bmatrix} \quad v = \sum_{j=1}^n x_j \omega_j + b = \omega^T x$$

$b = \omega x_0, \quad x_0 = 1$



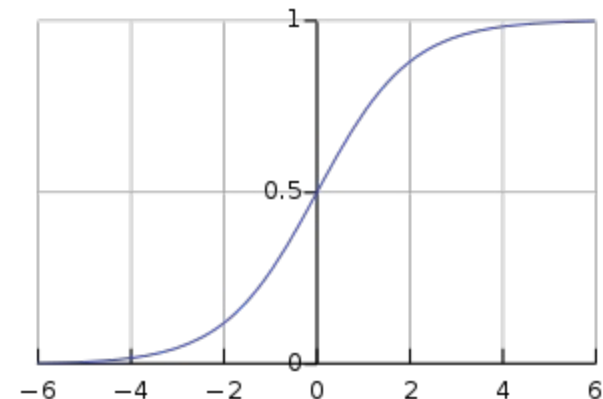
Logistic Regression

Optimal Classification and Bayes Risk

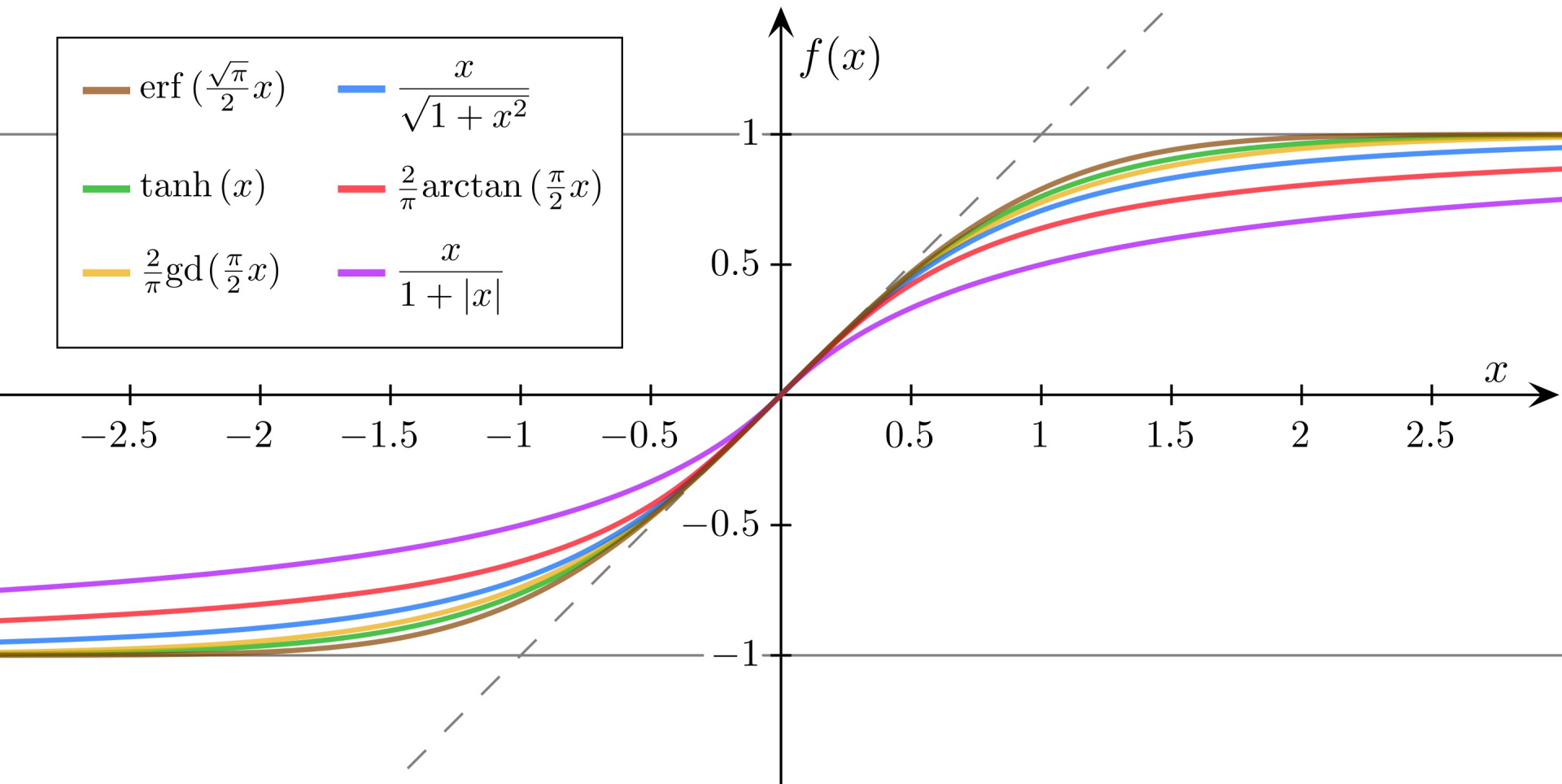


Which is better?

Linear function vs. Non-linear function of $P(Y|X)$



Sigmoid functions



Logistic Function

Sigmoid function is

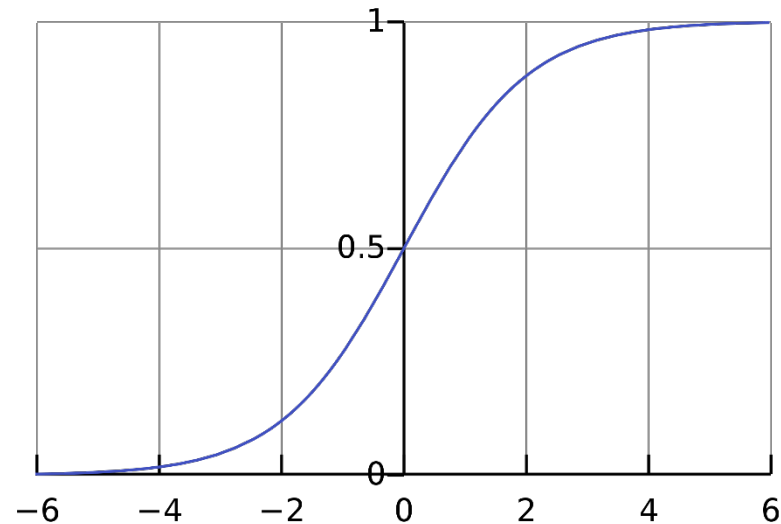
Bounded

Differentiable

Real function

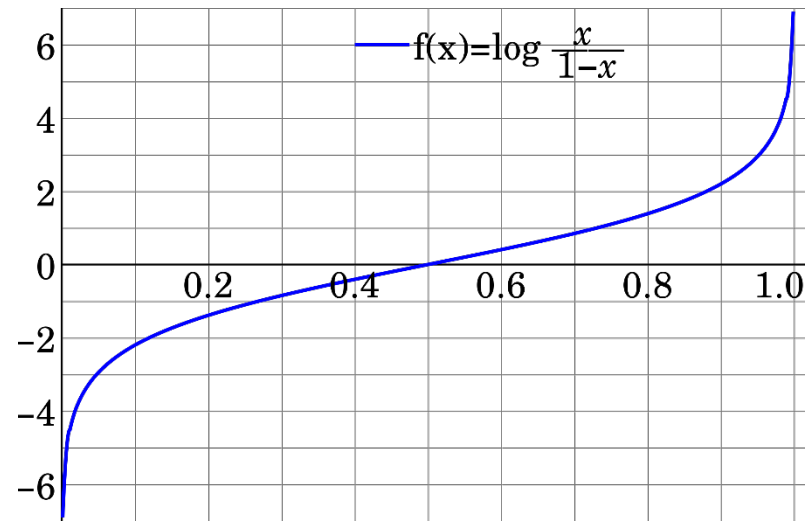
Defined for all real inputs

With positive derivative



Logistic function is

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



Logistic Function Fitting

Logic to Logistic

Inverse of X and Y

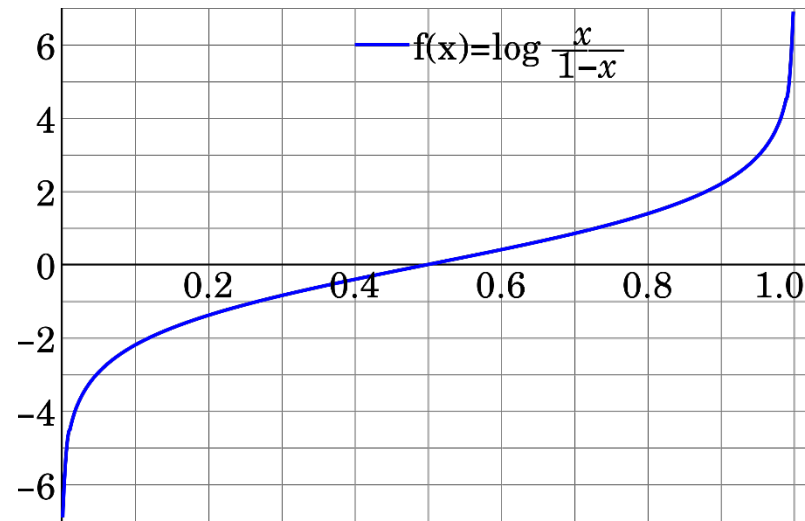
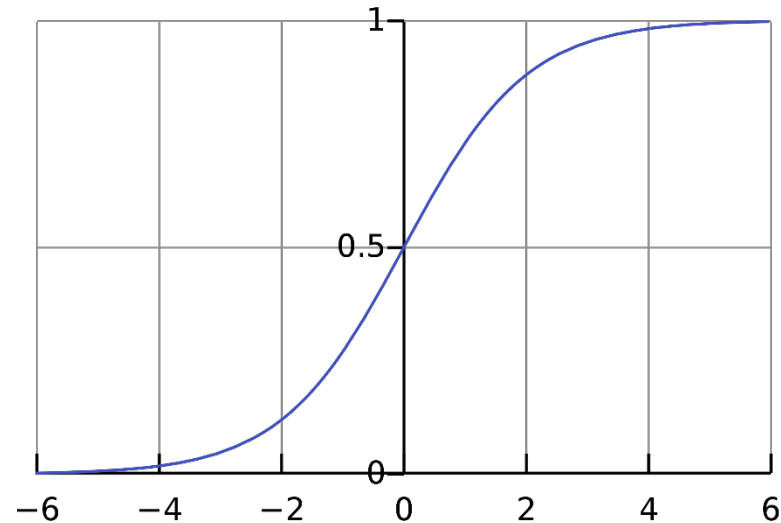
$$f(x) = \log \frac{x}{1-x} \rightarrow x = \log \frac{p}{1-p}$$

Linear function fitting

$$ax + b = \log \frac{p}{1-p} \rightarrow X\theta = \log \frac{p}{1-p}$$

Linear to Logistic

$$X\theta = P(Y|X) \rightarrow X\theta = \log \frac{P(Y|X)}{1-P(Y|X)}$$



Logistic Regression

Probabilistic classifier to predict the binomial or the multinomial outcome
by fitting the conditional probability to the logistic function

Bernoulli experiments

$$P(y|x) = \mu(x)^y (1 - \mu(x))^{1-y}$$

$$\mu(x) = \frac{1}{1 + e^{-\theta^T x}} = P(y = 1|x)$$

Here, $\mu(x)$ is the logistic function

$$X\theta = \log \frac{P(Y|X)}{1 - P(Y|X)} \rightarrow P(Y|X) = \frac{e^{X\theta}}{1 + e^{X\theta}}$$

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

Finding the Parameter, θ

Maximum Likelihood Estimation(MLE) of θ

Choose θ that maximizes the probability of observed data

$$\hat{\theta} = \operatorname{argmax}_{\theta} P(D|\theta)$$

Maximum Conditional Likelihood Estimation(MCLE)

$$\hat{\theta} = \operatorname{argmax}_{\theta} P(D|\theta)$$

$$= \operatorname{argmax}_{\theta} \prod_{1 \leq i \leq N} P(Y_i|X_i; \theta)$$

$$= \operatorname{argmax}_{\theta} \log \left(\prod_{1 \leq i \leq N} P(Y_i|X_i; \theta) \right)$$

$$= \operatorname{argmax}_{\theta} \sum_{1 \leq i \leq N} \log(P(Y_i|X_i; \theta))$$

$$P(Y_i|X_i; \theta) = \mu(X_i)^{Y_i}(1 - \mu(X_i))^{1-Y_i}$$

$$\log(P(Y_i|X_i; \theta)) = Y_i \log(\mu(X_i)) + (1 - Y_i) \log(1 - \mu(X_i))$$

$$= Y_i(\log(\mu(X_i)) - \log(1 - \mu(X_i))) + \log(1 - \mu(X_i))$$

$$= Y_i \log\left(\frac{\mu(X_i)}{1 - \mu(X_i)}\right) + \log(1 - \mu(X_i))$$

$$= Y_i X_i \theta + \log(1 - \mu(X_i)) \quad \longleftarrow \quad X\theta = \log \frac{P(Y|X)}{1 - P(Y|X)}$$

$$= Y_i X_i \theta - \log(1 + e^{X_i \theta})$$

$$\longleftarrow P(y = 1|x) = \mu(x)$$

$$= \frac{1}{1 + e^{-\theta^T x}} = \frac{e^{X\theta}}{1 + e^{X\theta}}$$

$$\begin{aligned}\hat{\theta} &= \operatorname{argmax}_{\theta} \sum_{1 \leq i \leq N} \log(P(Y_i|X_i; \theta)) \\ &= \operatorname{argmax}_{\theta} \sum_{1 \leq i \leq N} Y_i X_i \theta - \log(1 + e^{X_i \theta})\end{aligned}$$

Now, we need to optimize θ

$$\begin{aligned}& \frac{\partial}{\partial \theta_j} \sum_{1 \leq i \leq N} Y_i X_i \theta - \log(1 + e^{X_i \theta}) \\ &= \sum_{1 \leq i \leq N} Y_i X_{i,j} + \sum_{1 \leq i \leq N} -\frac{1}{1 + e^{X_i \theta}} \times e^{X_i \theta} \times X_{i,j} \\ &= \sum_{1 \leq i \leq N} X_{i,j} \left(Y_i - \frac{e^{X_i \theta}}{1 + e^{X_i \theta}} \right) \\ &= \sum_{1 \leq i \leq N} X_{i,j} (Y_i - P(Y_i = 1|X_i; \theta)) = 0 \quad \longleftarrow \text{Open form solution!}\end{aligned}$$

$$\theta = (X^T X)^{-1} X^T Y \quad \longleftarrow \text{Closed form solution!}$$

Gradient Descent

Taylor Expansion

Taylor series is a representation of a function

A infinite sum of terms

Calculated from the values of the function's derivatives at a fixed point

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \dots$$
$$= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x-a)^n$$

Here, a is a constant value

Taylor series is possible when

Infinitely differentiable at a real or complex number of a

Gradient Descent/Ascent

Gradient descent/ascent method

Given a differentiable function of $f(x)$ and a initial parameter of x_1

Iteratively moving the parameter to the lower/higher value of $f(x)$

By taking the direction of the negative/positive gradient of $f(x)$

$$f(x) = f(a) + \frac{f'(a)}{1!} (x - a) + O(|x - a|^2) \quad (a = x_1, x = x_1 + hu)$$

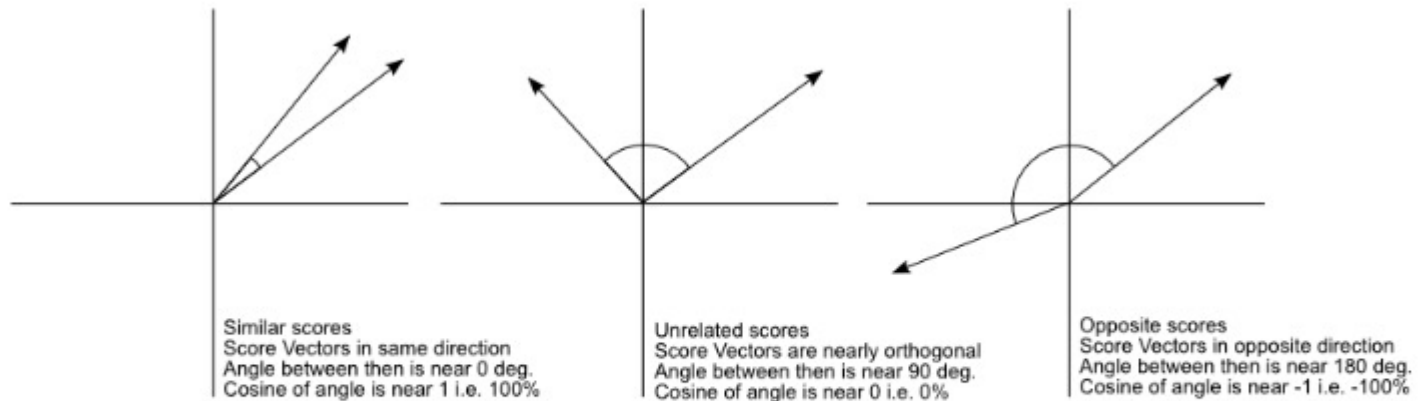
$$f(x_1 + hu) = f(x_1) + f'(x_1)(x_1 + hu - x_1) + O(|x_1 + hu - x_1|^2)$$

$$f(x_1 + hu) = f(x_1) + hf'(x_1)u + h^2O(1)$$

$$f(x_1 + hu) - f(x_1) \approx hf'(x_1)u$$

$$\begin{aligned}
 u^* &= \operatorname{argmin}_u f(x_1 + hu) - f(x_1) \\
 &= \operatorname{argmin}_u h f'(x_1) \mathbf{u} \\
 &= - \frac{f'(x_1)}{|f'(x_1)|} \quad (\text{Gradient Descent})
 \end{aligned}$$

$$x_{t+1} \leftarrow x_t + hu^* = x_t - h \frac{f'(x_1)}{|f'(x_1)|}$$



The Cosine Similarity values for different documents, 1 (same direction), 0 (90 deg.), -1 (opposite directions).

Rosenbrock Function

$$f(x_1, x_2) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2$$

$$\frac{\partial}{\partial x_1} f(x_1, x_2) = -2(1 - x_1) - 400x_1(x_2 - x_1^2)^2$$

$$\frac{\partial}{\partial x_2} f(x_1, x_2) = 200(x_2 - x_1^2)$$

Assume the initial point

$$x^0 = (x_1^0, x_2^0) = (-1.3, 0.9)$$

Partial derivative vector at the point

$$f'(x^0) = \left(\frac{\partial}{\partial x_1} f(x_1, x_2), \frac{\partial}{\partial x_2} f(x_1, x_2) \right) = (-415.4, -158.0)$$

$$x^1 \leftarrow x^0 - h \frac{f'(x^0)}{|f'(x^0)|} \quad (\text{Repeat updating until convergence})$$

Logistic Regression

$$\hat{\theta} = \operatorname{argmax}_{\theta} \sum_{1 \leq i \leq N} \log(P(Y_i|X_i; \theta)) \quad (\text{Gradient Ascent})$$

$$f(\theta) = \sum_{1 \leq i \leq N} \log(P(Y_i|X_i; \theta)) \quad (a = \theta_1, \theta = \theta_1 + hu)$$

$$\theta_{t+1} \leftarrow \theta_t + hu^* = \theta_t + h \frac{f'(\theta_1)}{|f'(\theta_1)|}$$

Setup an initial parameter of θ_1

Iteratively moving θ_t to the higher value of $f(\theta_t)$

By taking the direction of the **positive** gradient of $f(\theta_t)$

$$f(\theta) = \sum_{1 \leq i \leq N} \log(P(Y_i|X_i; \theta))$$

$$\begin{aligned} \frac{\partial}{\partial \theta_j} f(\theta) &= \frac{\partial}{\partial \theta_j} \sum_{1 \leq i \leq N} \log(P(Y_i|X_i; \theta)) \\ &= \frac{\partial}{\partial \theta_j} \sum_{1 \leq i \leq N} X_{i,j} (Y_i - P(Y_i = 1|X_i; \theta)) \end{aligned}$$

To utilize the gradient method

$$\begin{aligned} \theta_j^{t+1} &= \theta_j^t + h \frac{\partial f(\theta^t)}{\partial \theta_j^t} \\ &= \theta_j^t + h \left(\sum_{1 \leq i \leq N} X_{i,j} (Y_i - P(Y_i = 1|X_i; \theta^t)) \right) \\ &= \theta_j^t + \frac{h}{C} \left(\sum_{1 \leq i \leq N} X_{i,j} \left(Y_i - \frac{e^{X_i \theta^t}}{1 + e^{X_i \theta^t}} \right) \right) \end{aligned}$$

Linear Regression

$$\hat{\theta} = \operatorname{argmin}_{\theta} (f - \hat{f})^2 = \operatorname{argmin}_{\theta} (Y - X\theta)^2$$

$$\theta = (X^T X)^{-1} X^T Y$$

To utilize the gradient method

$$\hat{\theta} = \operatorname{argmin}_{\theta} (f - \hat{f})^2 = \operatorname{argmin}_{\theta} (Y - X\theta)^2$$

$$= \sum_{1 \leq i \leq N} (Y_i - \sum_{1 \leq j \leq d} X_j^i \theta_j)^2$$

$$\frac{\partial}{\partial \theta_k} \sum_{1 \leq i \leq N} (Y_i - \sum_{1 \leq j \leq d} X_j^i \theta_j)^2 = - \sum_{1 \leq i \leq N} 2 \left(Y_i - \sum_{1 \leq j \leq d} X_j^i \theta_j \right) X_k^i$$

$$\theta_k^{t+1} \leftarrow \theta_k^t - h \frac{\partial f(\theta^t)}{\partial \theta_k^t} = \theta_k^t + h \sum_{1 \leq i \leq N} 2 \left(Y_i - \sum_{1 \leq j \leq d} X_j^i \theta_j \right) X_k^i$$