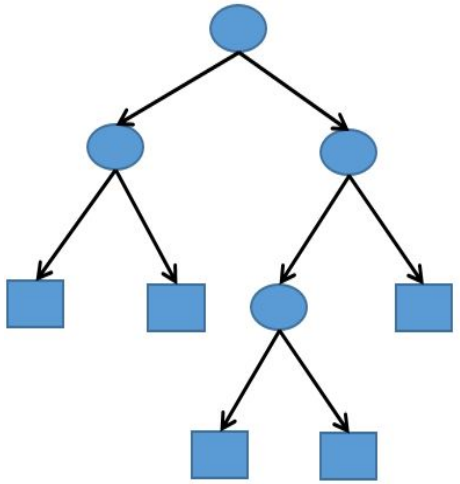
Introduction to Machine Learning

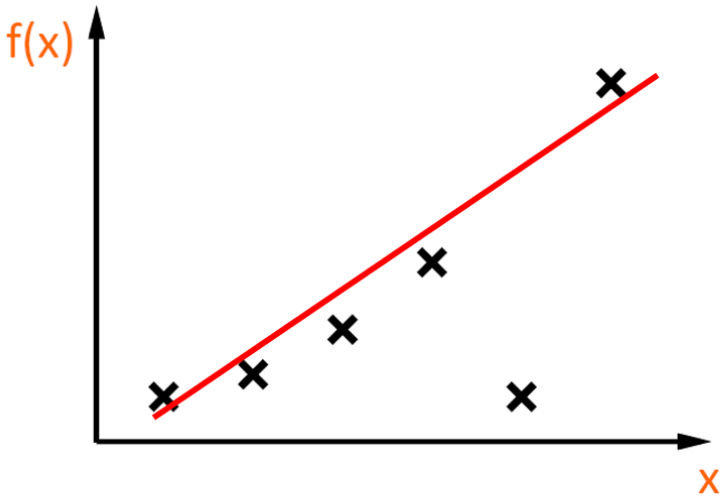
Part 4: Linear Models and SVM

Zengchang Qin (Ph.D.)

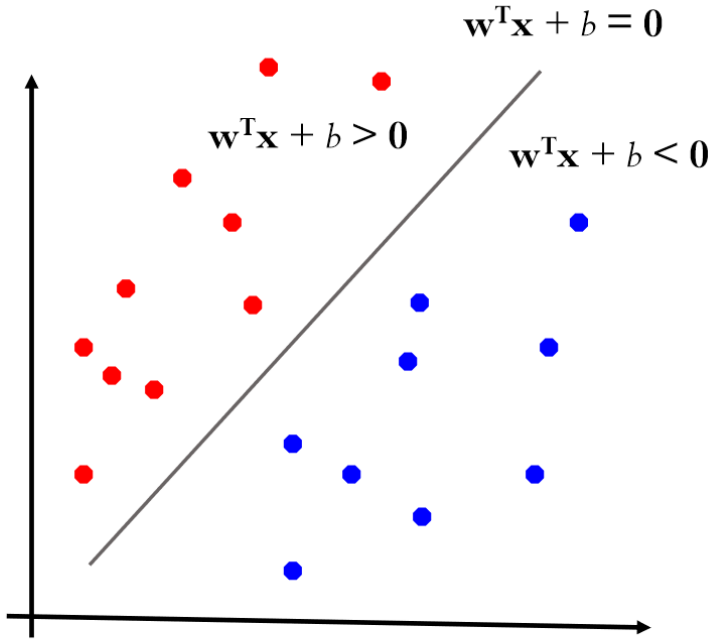


Linear Model

Linear Model

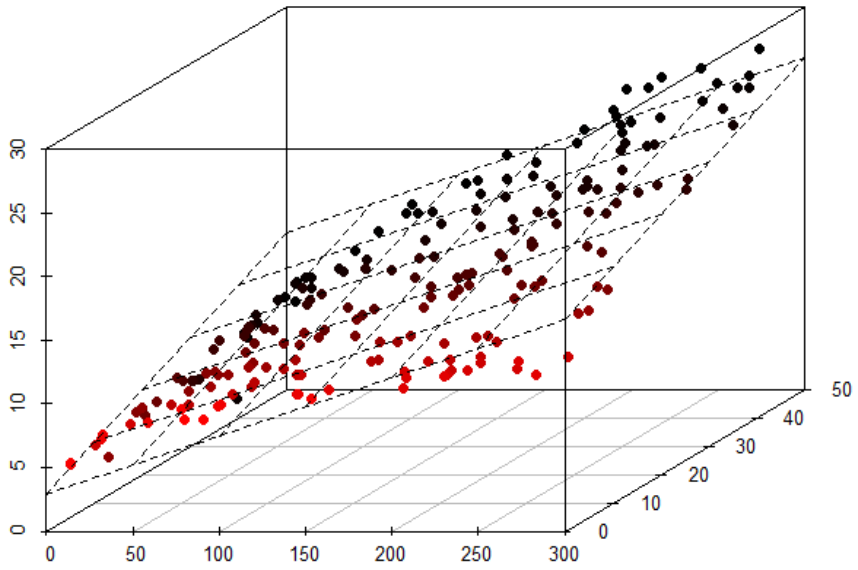


Linear fit



Linear decision boundary

Linear Regression



$$h_\theta(x) = \theta_0 + \theta_1 x$$

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

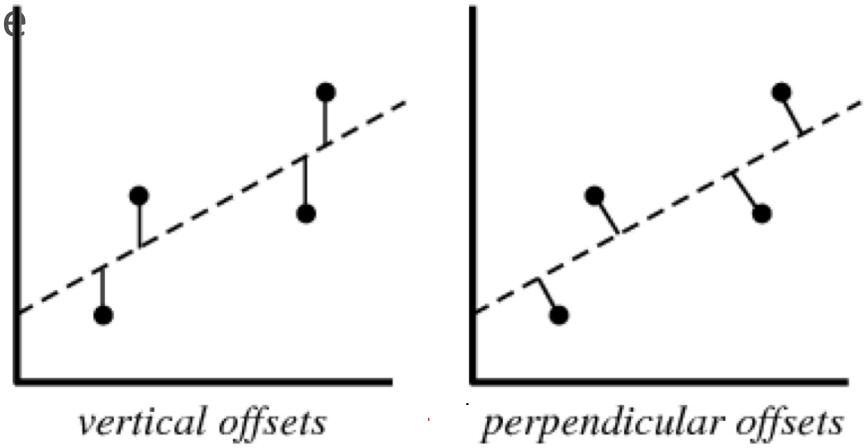
$$h(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x$$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

In statistics, the most common occurrence is in connection with regression models and the term is often taken as synonymous with linear regression model.

Least Square

A mathematical procedure for finding the best-fitting curve to a given set of points by minimizing the sum of the squares of the offsets.



$$S = \sum_{i=1}^{n} {r_i}^2 \qquad r_i = y_i - f(x_i, \boldsymbol{\beta}).$$

Solving the least squares problem:

$$\frac{\partial S}{\partial \beta_j} = 2 \sum_i r_i \frac{\partial r_i}{\partial \beta_j} = 0, \ j = 1, \ldots, m,$$

$$= -2 \sum_i r_i \frac{\partial f(x_i, \boldsymbol{\beta})}{\partial \beta_j} = 0, \ j = 1, \ldots, m.$$

since $r_i = y_i - f(x_i, \boldsymbol{\beta})$

Residuals are the vertical distances between the data points and the corresponding predicted values.

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}.$$

Matrix

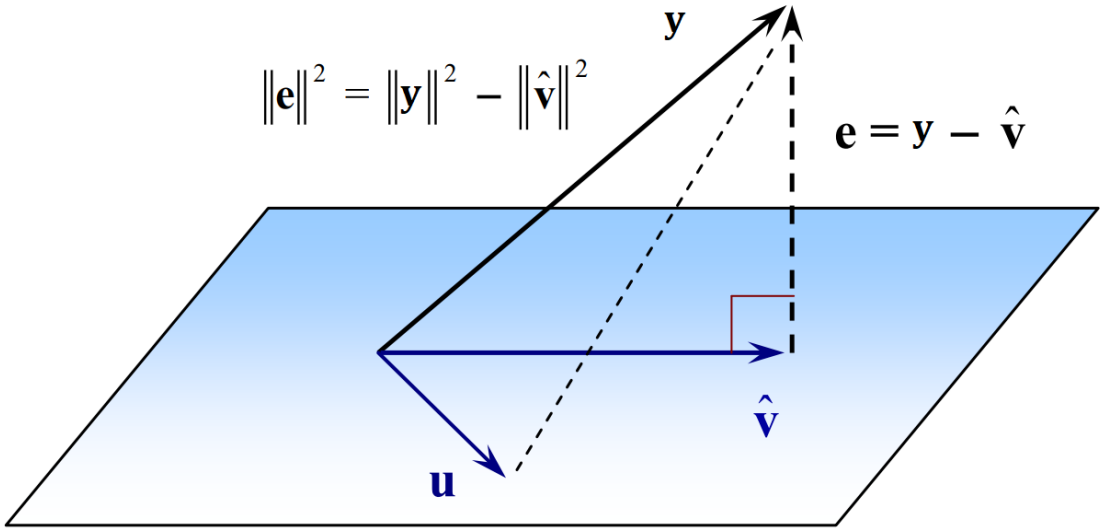
$$\sum_{j=1}^{n} X_{ij}\beta_j = y_i, \ (i = 1, 2, \dots, m), \quad \mathbf{X}\boldsymbol{\beta} = \mathbf{y},$$

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}), \quad S(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left| y_i - \sum_{j=1}^{n} X_{ij}\beta_j \right|^2 = \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2$$

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

Geometrical Interpretation



$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

Probabilistic Interpretation

- Let us assume that the target variables and the inputs are related via the equation:

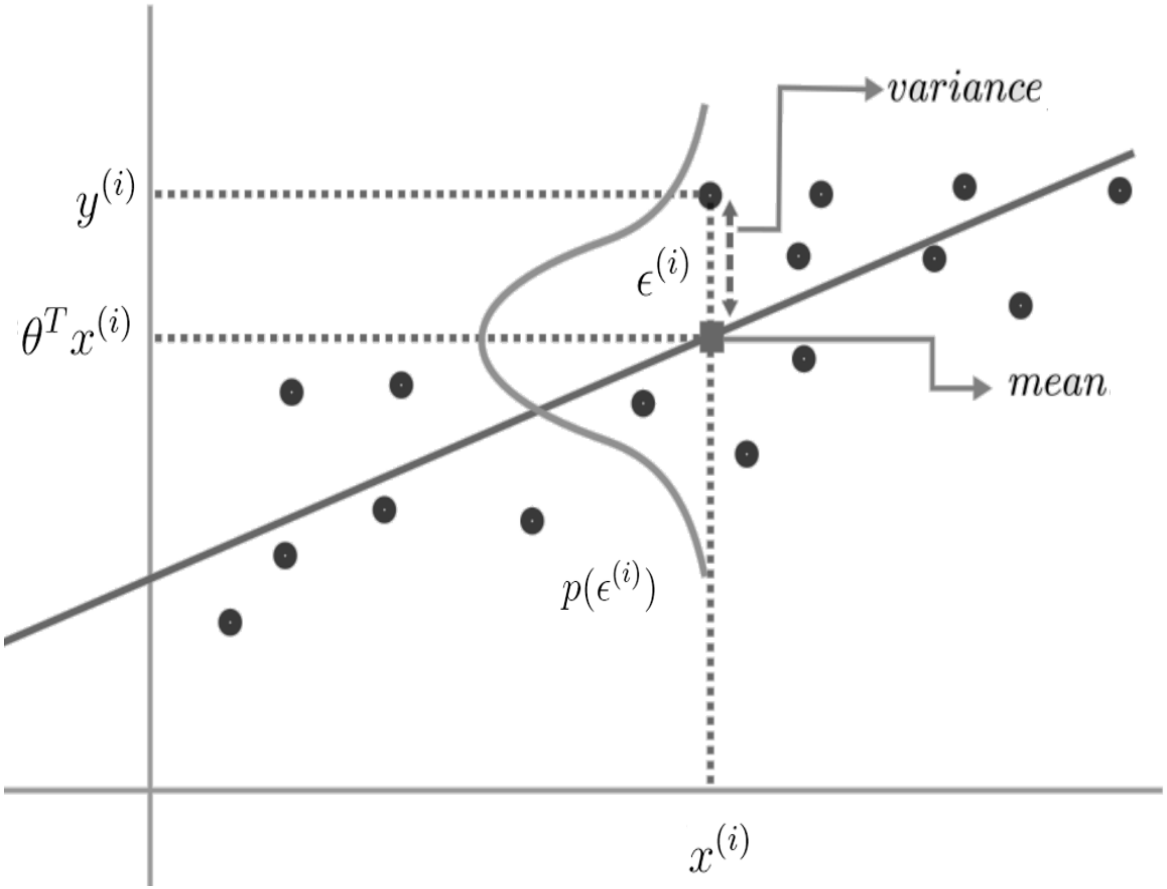
$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

where $\varepsilon^{(i)}$ is an error term.

- We can write this assumption as $\varepsilon^{(i)}$ follows a Normal distribution:

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right)$$

Likelihood and Cost Function



Likelihood

$X_1, X_2, X_3, \ldots X_n$ have joint density denoted

$$f_\theta(x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta)$$

Given observed values $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, the likelihood of θ is the function

If $X_1, X_2, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$ random variables their density is written:

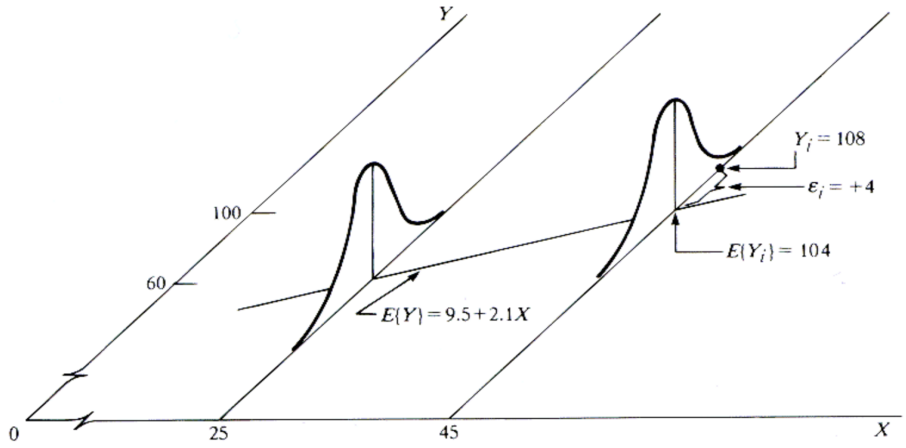
$$f(x_1, \ldots, x_n|\mu, \sigma) = \prod_i^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}[\frac{x_i - \mu}{\sigma}]^2\right)$$

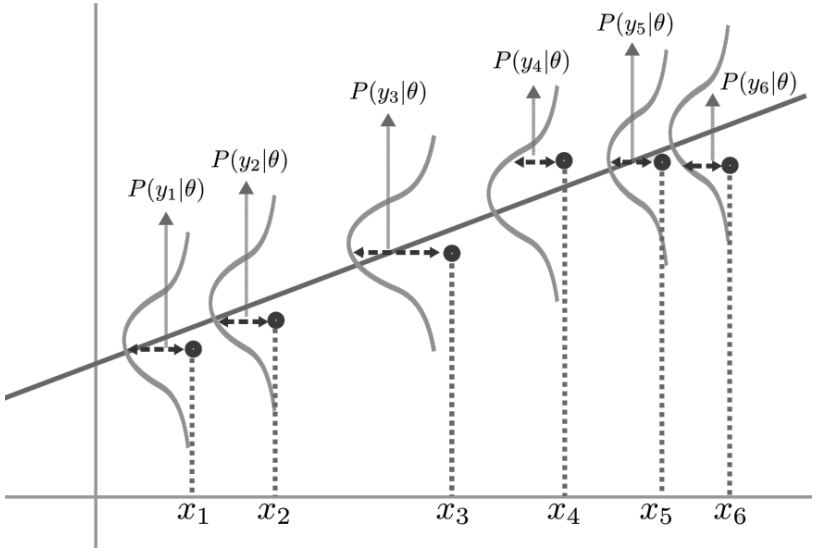
$$f(x_1, x_2, \ldots, x_n \mid \theta) = f(x_1 \mid \theta) \times f(x_2 \mid \theta) \times \cdots \times f(x_n \mid \theta).$$

$$\mathcal{L}(\theta\,;\, x_1, \ldots, x_n) = f(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

$$\ln \mathcal{L}(\theta\,;\, x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i \mid \theta),$$

ML Estimation



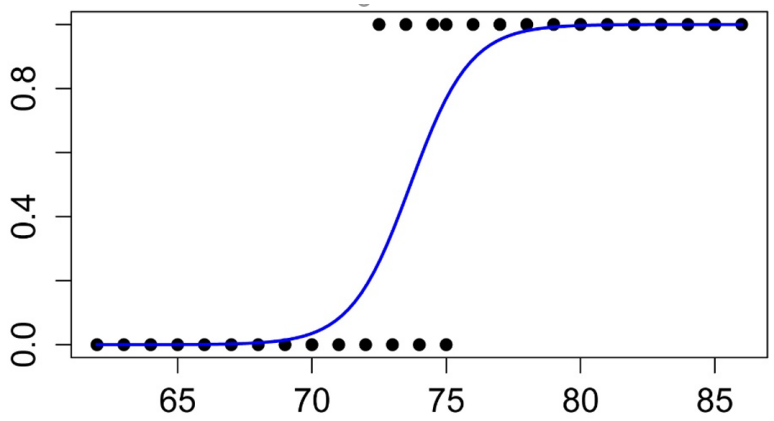


It implies that: $$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right)$$

■ When we wish to explicitly view this as a function of θ, we will instead call it the likelihood function:

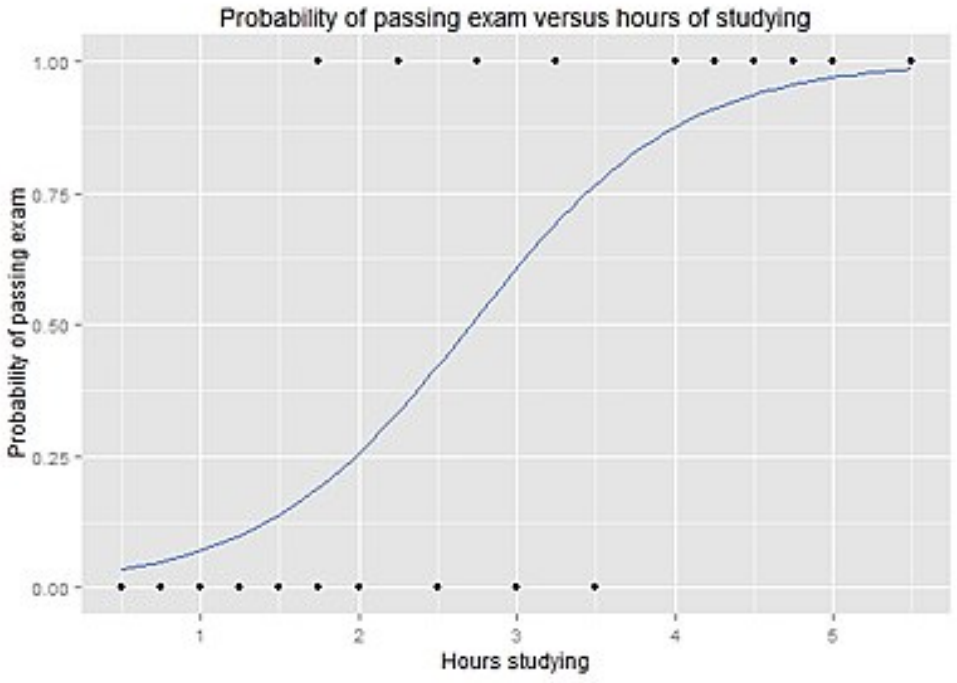
$$L(\theta) = L(\theta; X, \vec{y}) = p(\vec{y}|X;\theta) = \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right)$$

Logistic Regression

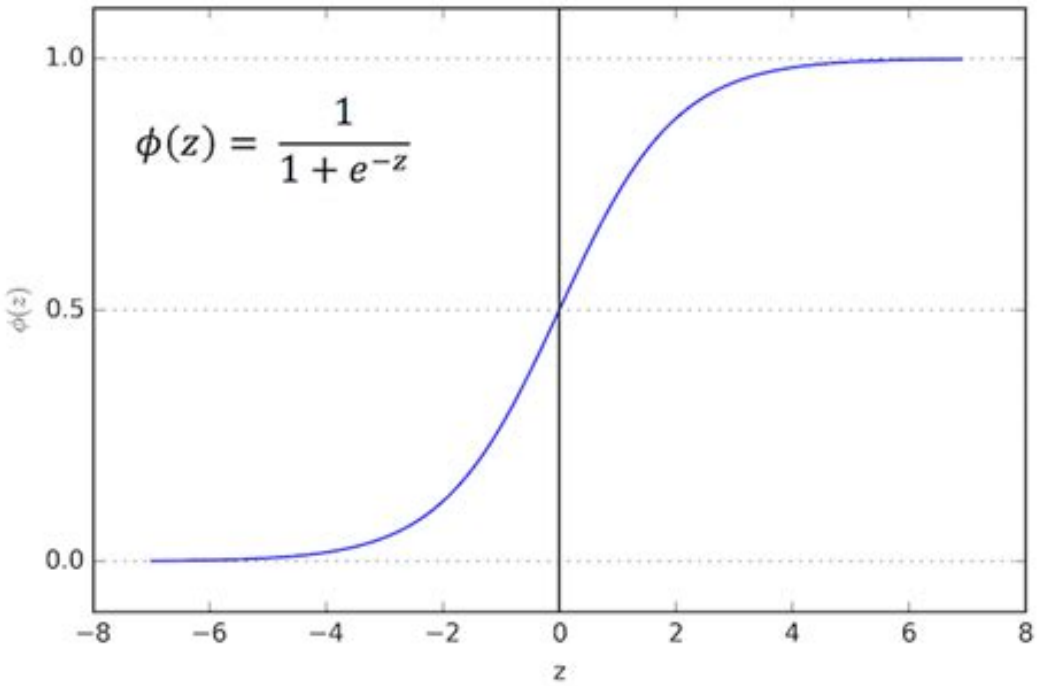


A Simple Classification Problem

Hours	0.5	0.75	1	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1



Sigmoid Function

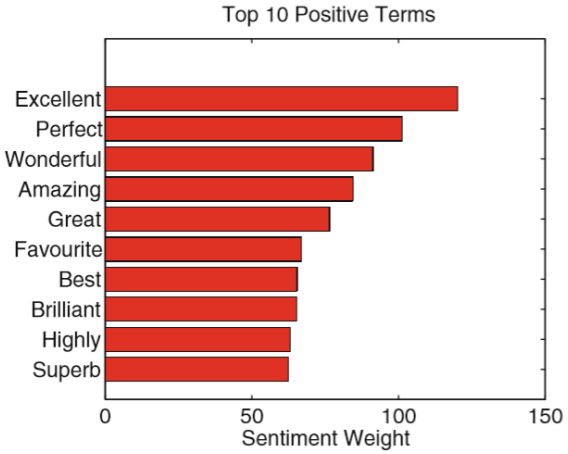


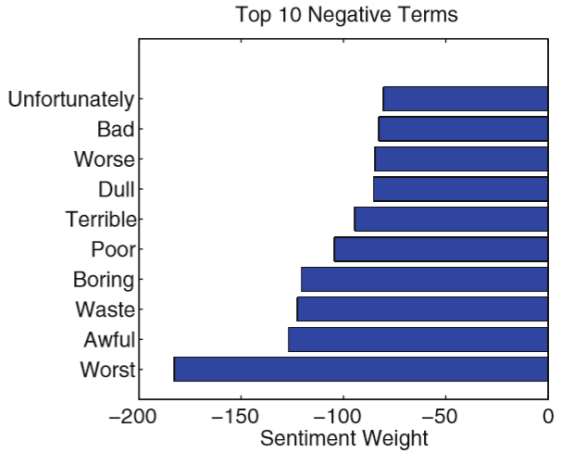
$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x})}$$

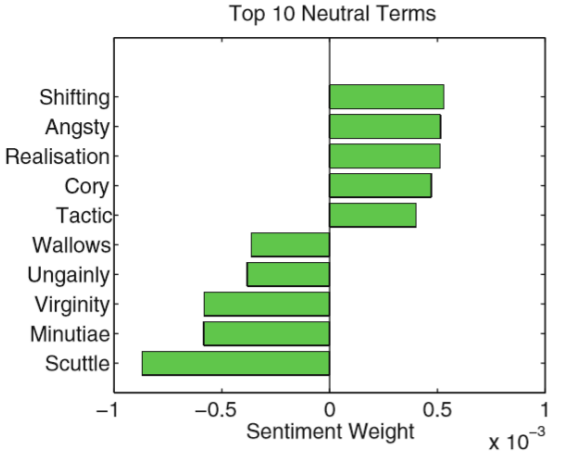
- How to learn the parameters?
- Can we do MLE?

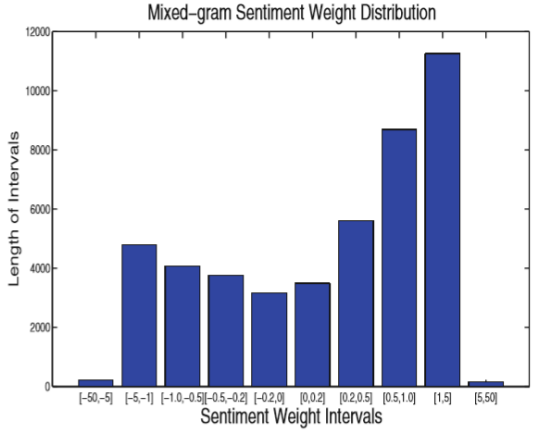
Application

$$h = f\left(\sum_{i=1}^{N} w_i x_i\right) = f(\mathbf{w}^T\mathbf{x}) \qquad h_j = \frac{1}{1+\exp\left(-\mathbf{w}^T\mathbf{x}_j\right)} \quad ; \quad 1 \le j \le M$$



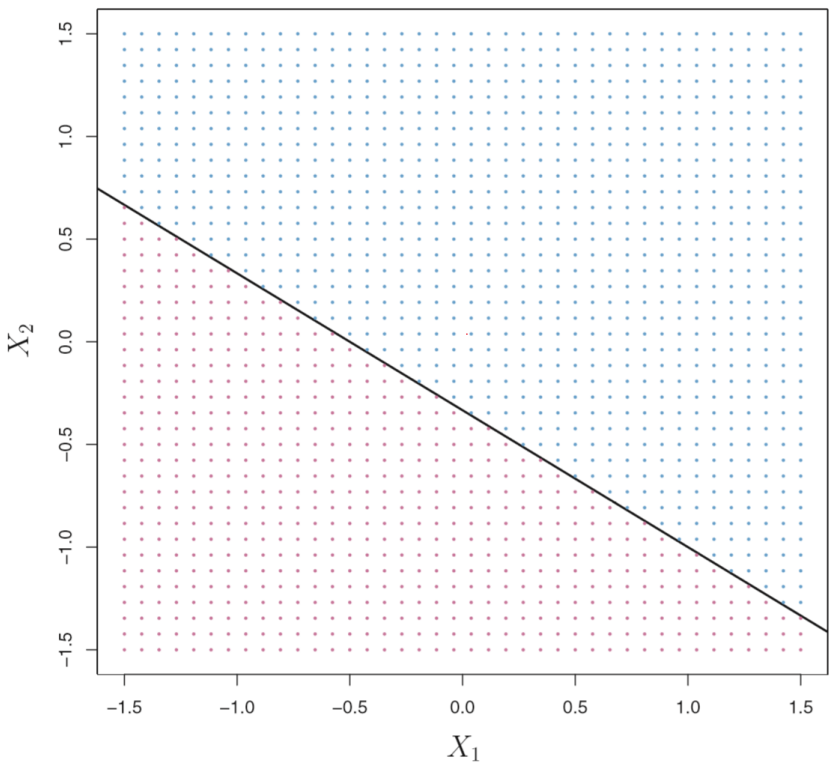


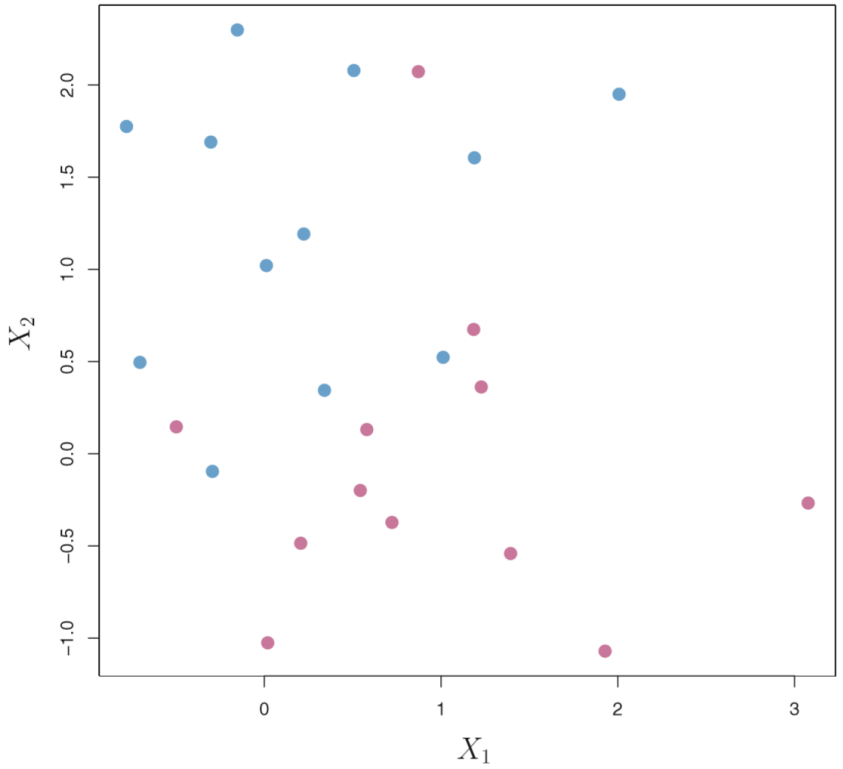




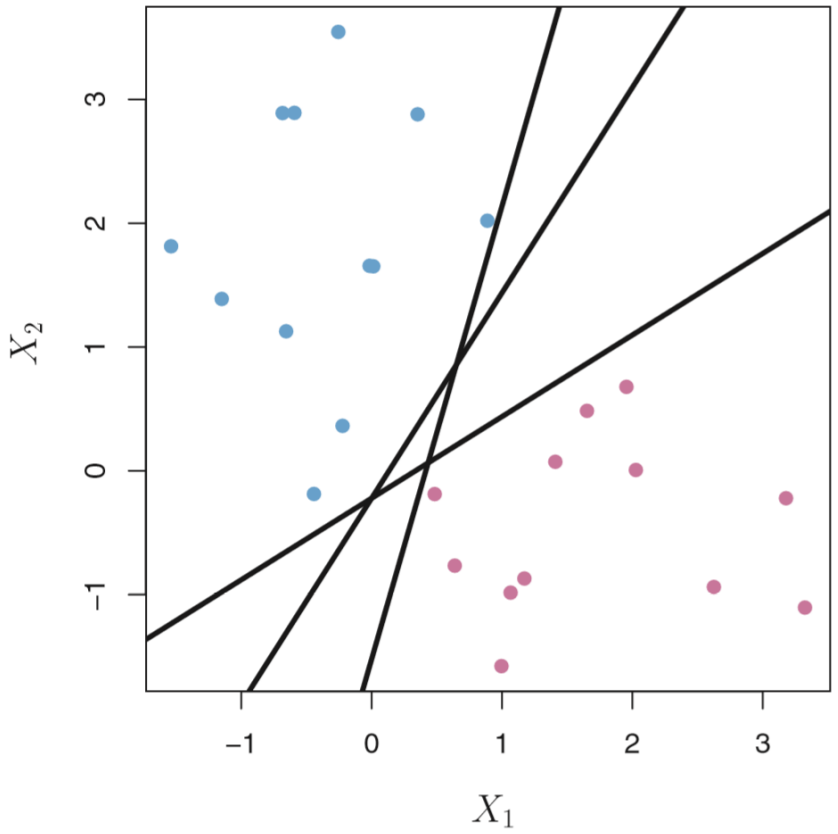
Hyperplane

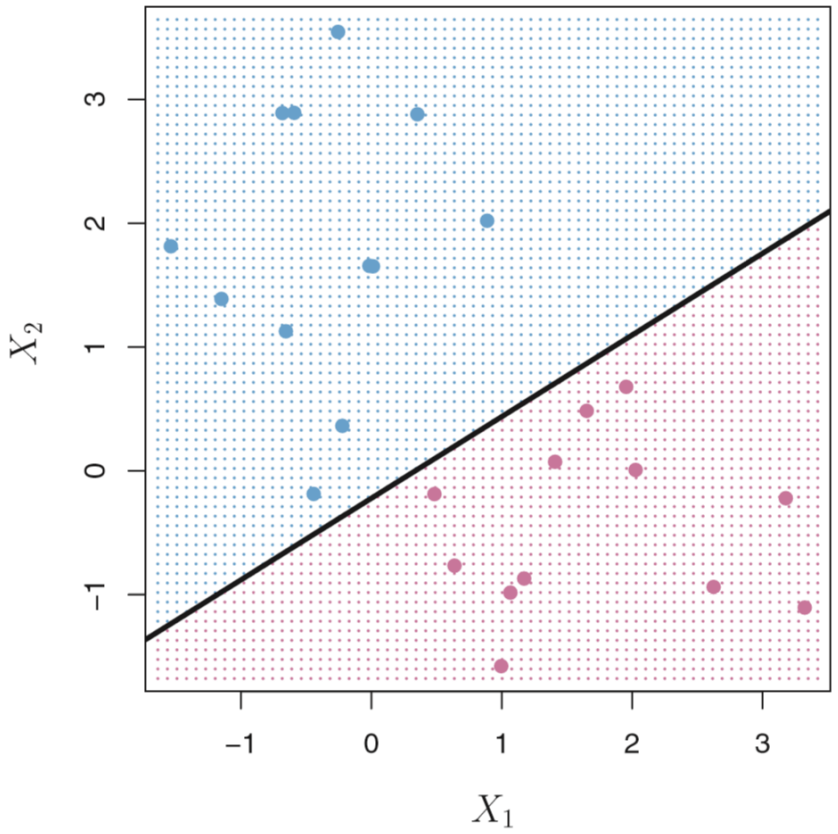
The hyperplane 1+2X1+3X2 =0 is shown. The blue region is the set of points for which 1+2X1 +3X2 > 0, and the purple region is the set of points for which 1+2X1 +3X2 < 0.





Which One is Better?





The Problem

Given a data set,

$$D = \{ (\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_m}, y_m)\}, y_i \in \{-1,1\},$$

How can we find a hyperplane to classify them?
Is it the best one?

A hyperplane can be described as the following function

$$\boldsymbol{\omega}^T \boldsymbol{x} + b = 0,$$

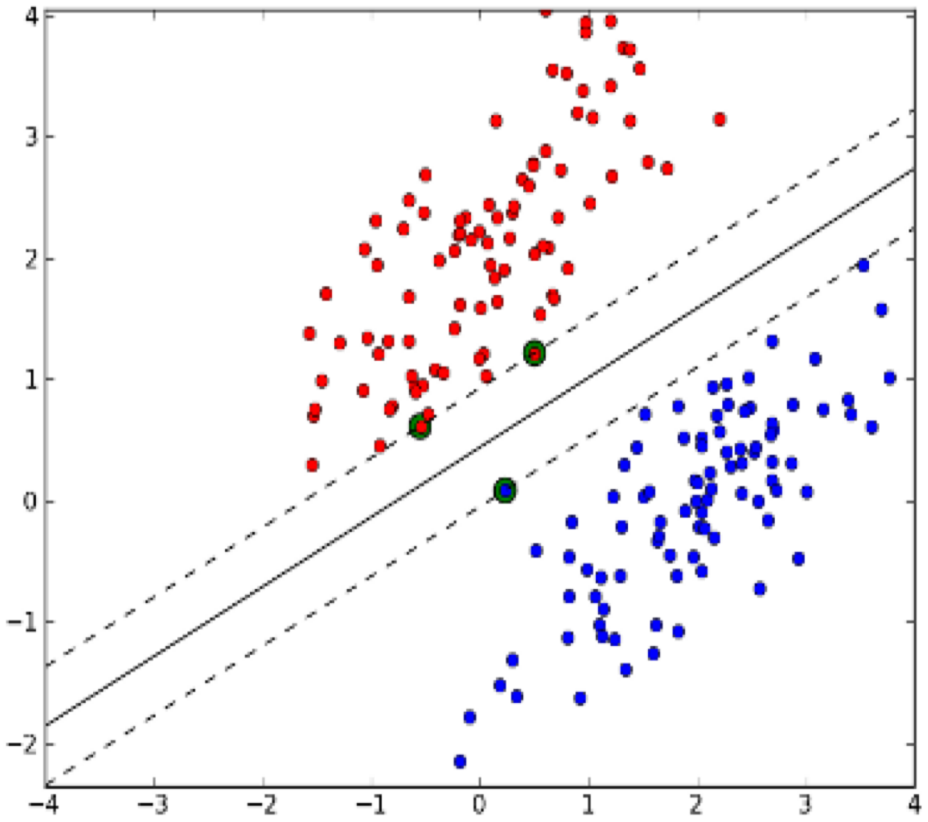
where $\boldsymbol{\omega} = \{\omega_1; \omega_2; \ldots; \omega_d\}$ is the normal vector of the hyperplane

$$\boldsymbol{\omega}^T \boldsymbol{x_i} + b \geq +\sigma,$$

$$\boldsymbol{\omega}^T \boldsymbol{x_i} + b \leq -\sigma,$$

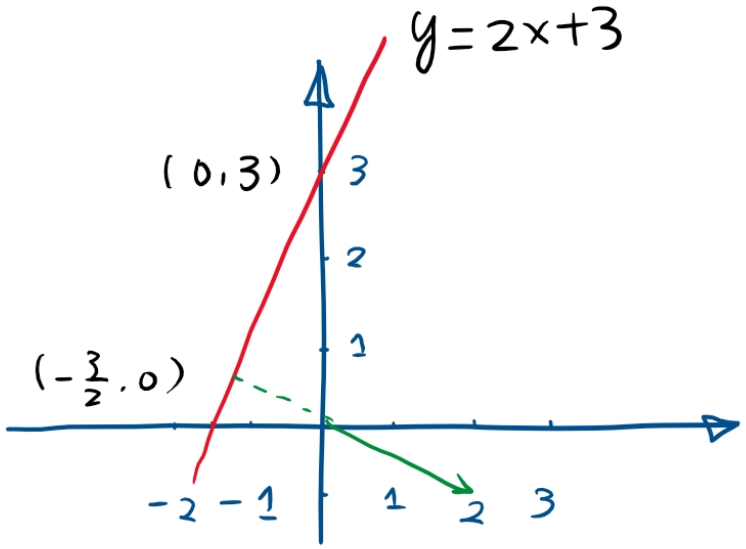
Normalize to $y_i(\boldsymbol{\omega}^T \boldsymbol{x_i} + b) >= 1, \quad i=1,2, \ldots, r$

Support Vectors



Linear Function

1) Linear Function



Given a linear function $y = 2x+3$, we have $(0,3)$ and $(-\frac{3}{2}, 0)$ on the hyperplane.

$$y = 2x+3 \Rightarrow$$

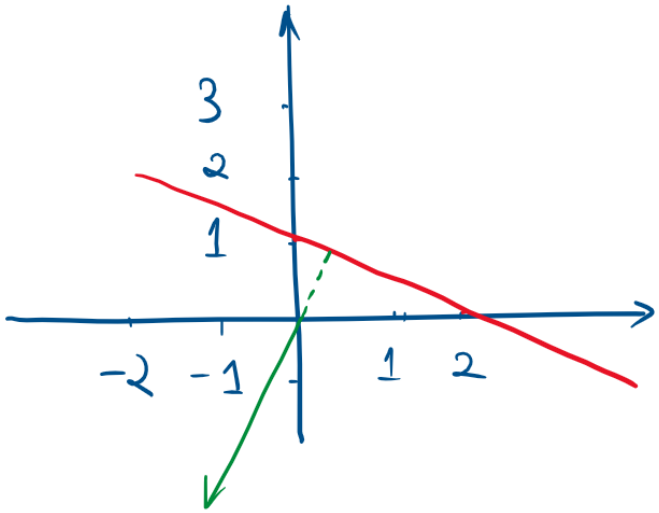
$$2x - y + 3 = 0$$

$$(2, -1)\begin{pmatrix} x \\ y \end{pmatrix} + 3 = 0$$

Probabilistic Interpretation

For a linear function $w^T x + b = 0$, the direction of w^T is perpendicular to the original linear function.

E.g.



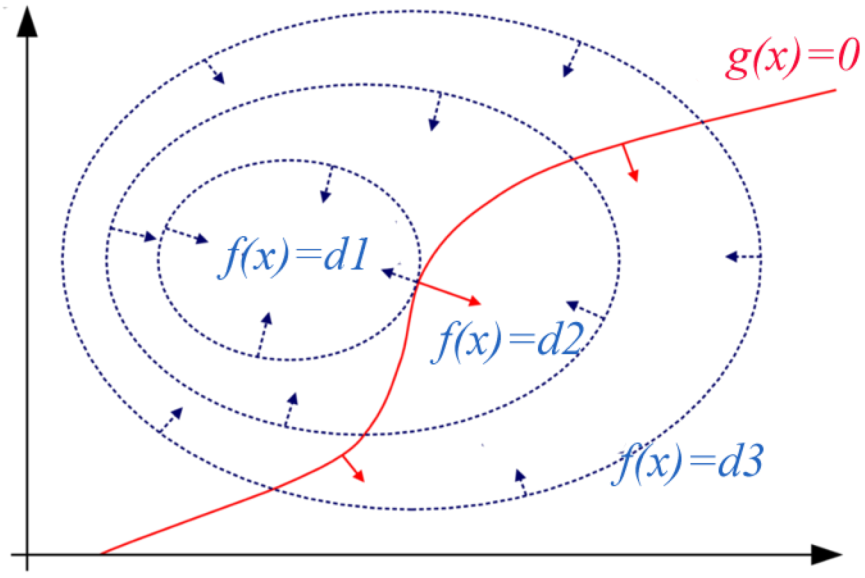
$$\left(-\frac{1}{2}, -1\right)\begin{pmatrix} x \\ y \end{pmatrix} + 1 = 0$$

$$w^T = \left(-\frac{1}{2}, -1\right)$$

$$y = -\frac{1}{2}x + 1$$

Lagrange Multiplier

Lagrange multipliers:



Case1: equality constraint

$$\min f(\boldsymbol{x})$$

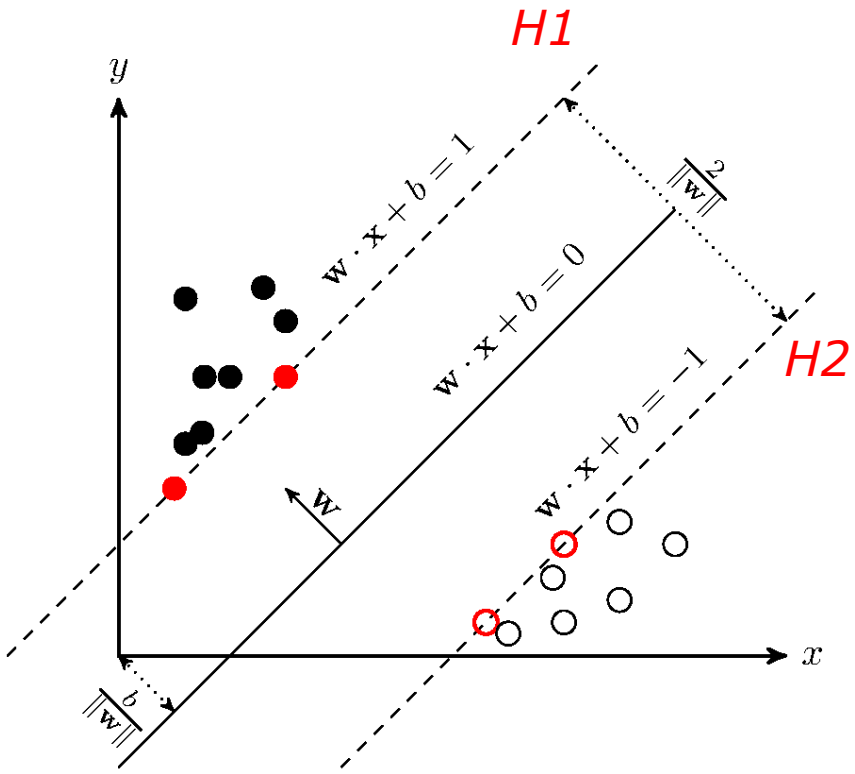
$$s.t.: \quad g(\boldsymbol{x}) = 0$$

$$\nabla f(\boldsymbol{x}^*) + \lambda \nabla g(\boldsymbol{x}^*) = 0, \lambda \neq 0$$

We can combine the constraints with objective function together

$$L(x, \lambda) = f(x) + \lambda g(x)$$

Margin



$$H1: \quad \boldsymbol{\omega}^T \boldsymbol{x_i} + b = +1,$$

$$H2: \quad \boldsymbol{\omega}^T \boldsymbol{x_i} + b = -1,$$

Recall that in 2-D, the distance from a point (x_0, y_0) to a line $Ax + By + C = 0$ is

$$\frac{|Ax_0 + By_0 + C|}{\sqrt{A^2 + B^2}}$$

So the distance from hyperplane *H1* to H2 can be computed as

$$\frac{2}{||\boldsymbol{\omega}||}$$

Maximization

In conclusion, the objective function is

$$\max_{\boldsymbol{\omega},b} \frac{2}{||\boldsymbol{\omega}||}$$

$$\text{s. t.} \quad y_i(\boldsymbol{\omega}^T \boldsymbol{x}_i + b) \geq 1, \qquad i=1,2,\ldots,\text{m.}$$

Alternatively, we can minimize the denominator

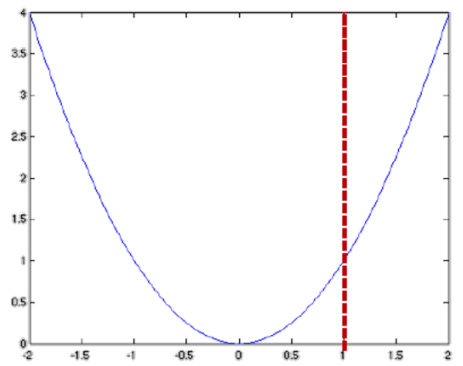
$$\min_{\boldsymbol{\omega},\boldsymbol{b}} \frac{1}{2}||\boldsymbol{\omega}||^2$$

$$\text{s. t.} \quad y_i(\boldsymbol{\omega}^T \boldsymbol{x}_i + b) \geq 1, \qquad i=1,2,\ldots,\text{m.}$$

ML Estimation

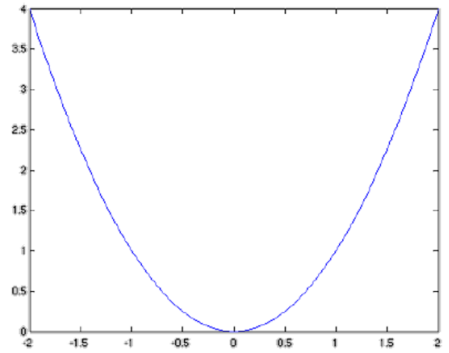
$\min_x \; x^2$
s.t. $\quad x \geq b$

$\min_x \; x^2$
s.t. $\quad x \geq 1$



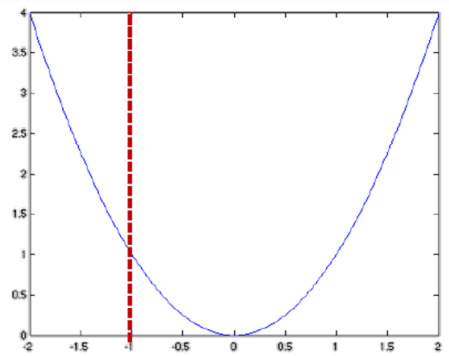
$x^* = 1$

$\min_x \; x^2$



$x^* = 0$

$\min_x \; x^2$
s.t. $\quad x \geq -1$



$x^* = 0$

Langrangian Muliplier

Move the constraint to objective function – **Lagrangian**

$$L(x, \alpha) = x^2 - \alpha(x - b), \quad \text{s.t.: } \alpha \geq 0$$

$$\min_x \max_\alpha \quad L(x, \alpha)$$
$$\text{s.t.:} \quad \alpha \geq 0$$

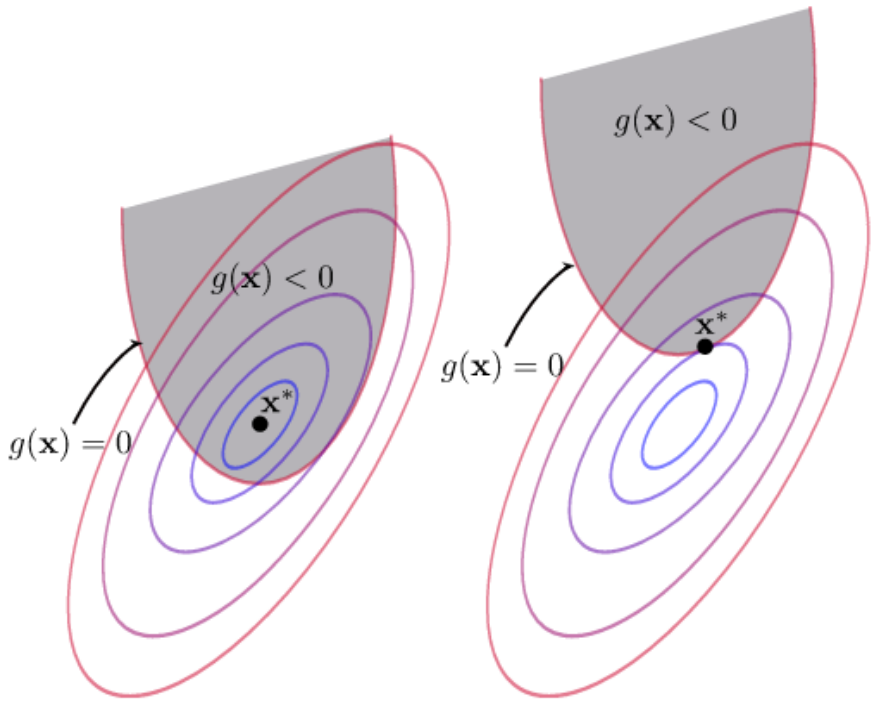
$$\min_x \max_\alpha \quad L(x, \alpha) = x^2 - \alpha(x - b)$$
$$\text{s.t.:} \quad \alpha \geq 0$$

To solve the min max problem

$$\frac{\partial L}{\partial x} = 0 \Rightarrow x^* = \frac{\alpha}{2}$$

$$\frac{\partial L}{\partial \alpha} = 0 \Rightarrow \alpha^* = \max(2b, 0)$$

Inequality Constraint



Case 2: inequality constraint

$$\min f(\boldsymbol{x})$$

$$s.\ b. \quad g(\boldsymbol{x}) \leq 0$$

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$$

$$g(\boldsymbol{x}) < 0, \quad \lambda = 0$$

$$g(\boldsymbol{x}) = 0, \quad \nabla f(\boldsymbol{x}^*) + \lambda \nabla g(\boldsymbol{x}^*) = 0, \ \lambda > 0$$

$$\rightarrow \quad \lambda\, g(\boldsymbol{x}) = 0$$

Dural Form

1. Primal problem

$$\min_{\boldsymbol{\omega},\boldsymbol{b}} \ \frac{1}{2}||\boldsymbol{\omega}||^2$$

$$\text{s. t.} \quad y_i(\boldsymbol{\omega}^T \boldsymbol{x}_i + b) \geq 1, \qquad i=1,2,\ldots,\text{m.}$$

2. Lagrange function

$$L(\boldsymbol{\omega},\ b,\ \boldsymbol{\alpha}) = \frac{1}{2}||\boldsymbol{\omega}||^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\boldsymbol{\omega}^T \boldsymbol{x}_i + b))$$

$$\frac{\partial L(\boldsymbol{\omega},\ b,\ \boldsymbol{\alpha})}{\partial \boldsymbol{\omega}} = 0 \qquad \boldsymbol{\omega} = \sum_{i=1}^{m} \alpha_i y_i x_i$$

$$\frac{\partial L(\boldsymbol{\omega},\ b,\ \boldsymbol{\alpha})}{\partial \mathrm{b}} = 0 \qquad 0 = \sum_{i=1}^{m} \alpha_i y_i$$

Dual Form

Move the constraint to objective function – **Lagrangian**

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i}^T \boldsymbol{x_j}$$

s.t. $\sum_{i=1}^{m} \alpha_i y_i = 0,$

$\alpha_i \geq 0, \qquad i = 1, \ldots, m$

More details are available in the written lecture notes!