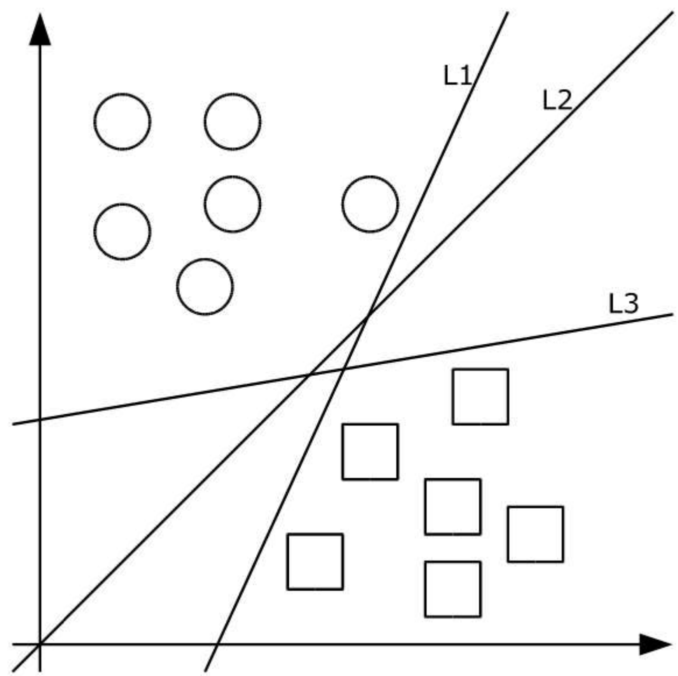
Preliminaries

- Data set is pairs $\{(x^{(1)}, y^{(1)}), \cdots, (x^{(m)}, y^{(m)})\}$, where each $y^{(i)} \in \{-1, 1\}$
- We can build different hyperplanes

$$w \cdot x_i > b \rightarrow y_i = 1$$

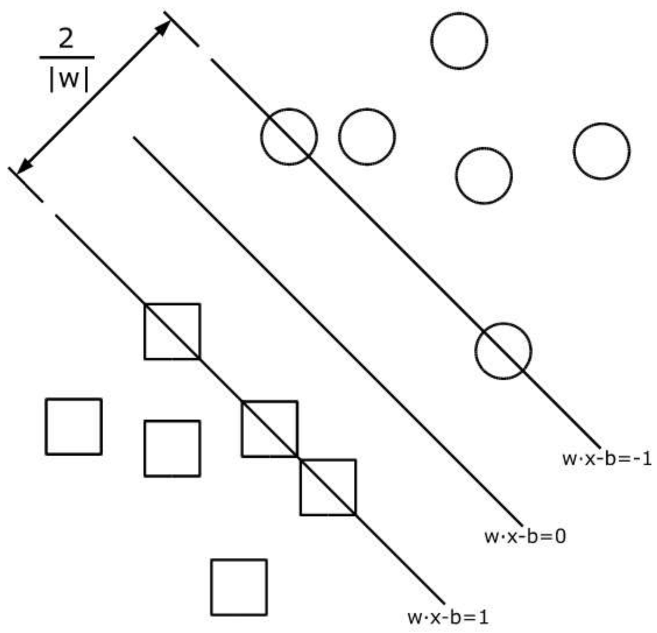
$$w \cdot x_i < b \rightarrow y_i = -1$$



Margin

$$w \cdot x_i > b + \varepsilon \rightarrow y_i = 1$$

$$w \cdot x_i < b - \varepsilon \rightarrow y_i = -1$$



Margin

- Separating hyperplane dosen't change if we multiply its equation by some constant
- Let's multiply by $\frac{1}{\varepsilon}$

$$w \cdot x_i - b > +1 \rightarrow y_i = 1$$

$$w \cdot x_i - b < -1 \rightarrow y_i = -1$$

- The best hyperplane is one that gives the widest margin
- The width of the margin is $\frac{2}{||w||}$

Optimization Task, Linearly Separable Case

- We would like to minimize $||w|| = w \cdot w^T$
- with respect to constraints

$$y_i(w \cdot x_i - b) \geq 1$$

Lagrangian for Linearly Separable Case

We want to find w, b such that

- $||w|| = w \cdot w^T \to \min$
- $y_i(w \cdot x_i - b) - 1 \geq 0, \; i = 1, \cdots, m$

By Karush-Kuhn-Tucker theorem this is equivalent to optimization of the Lagrangian

$$L(w, b; \lambda) = \frac{1}{2} w \cdot w^T - \sum_{i=1}^{m} \lambda_i (y_i (w \cdot x_i - b) - 1) \to \min_{w,b} \max_{\lambda}$$

$$\lambda_i \geq 0, \quad i = 1, \cdots, m$$

Lagrangian

$$\frac{\partial L}{\partial w} = w - \sum_{i=0}^{m} \lambda_i y_i x_i = 0 \rightarrow w = \sum_{i=0}^{m} \lambda_i y_i x_i \ (1)$$

$$\frac{\partial L}{\partial b} = -\sum_{i=0}^{m} \lambda_i y_i = 0 \rightarrow \sum_{i=0}^{m} \lambda_i y_i = 0 \ (2)$$

- w is linear combination of training set vectors those $\lambda_i \neq 0$ (support vectors)
- Using (1) and (2) we can get

$$\hat{L}(\lambda) = \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle \rightarrow \max_{\lambda}$$

$$\lambda_i \geq 0, i = 1, \cdots, m$$

$$\sum_{i=0}^{m} \lambda_i y_i = 0$$

Obtaining Parameters

- Solving with respect to each λ_i and using (1) obtain w
- To find b we can calculate average $w \cdot x_i$ over all support vectors
- Note that

$$w^T x + b = \left(\sum_{i=0}^{m} \lambda_i y_i x_i \right)^T x + b = \sum_{i=0}^{m} \lambda_i y_i \langle x_i, x \rangle + b$$

Kernels

- The algorithm can be written in terms of the inner products $\langle x, z \rangle$
- We could replace all those inner products with $\langle \phi(x), \phi(z) \rangle$
- Where $\phi(x)$ some feature mapping e.g.

$$\phi(x) = \begin{pmatrix} x \\ x^2 \\ x^3 \end{pmatrix}$$

- Specifically, given a feature mapping ϕ, we define corresponding Kernel to be

$$K(x, z) = \phi(x)^T \phi(z)$$

Optimization Task, Soft Margin

- In case of linearly inseparable case examples are permitted to have margin less then 1

$$y_i(w \cdot x_i - b) \geq 1 - \xi_i$$

- If example has margin $1 - \xi_i$ with $\xi > 0$ we would pay a cost of objective function to being increased by $C\xi_i$

$$||w|| = w \cdot w^T + C \sum_i \xi_i$$