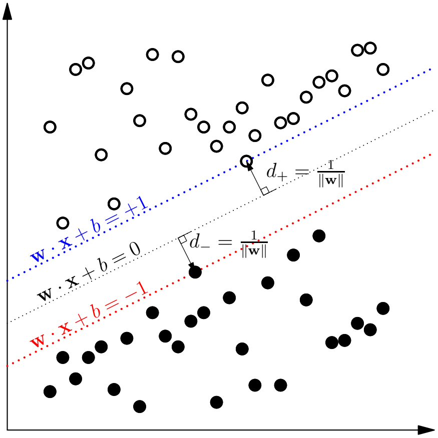
Operations Research
09. Support Vector Machine (SVM)

Binary Classification

Given the data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, $y_i \in \{-1, +1\}$, $\mathbf{x}_i \in \mathbb{R}^n$, find the hyperplane with maximum "margin" — the gap between parallel hyperplanes seperating two classes where the vectors of neither class can lie



SVM: Linearly Separable (Hard-Margin)

Let $\mathbf{w}\cdot\mathbf{x}+b=0$ be the seperating hyperplane and d_+, d_- be the shortest distance to the closest objects from the class $+1$, -1, respectively.

Suppose that

$$\begin{aligned} \mathbf{w}\cdot\mathbf{x}_i+b &\geqslant +1 \quad \text{for } \ y_i = +1 \\ \mathbf{w}\cdot\mathbf{x}_i+b &\leqslant -1 \quad \text{for } \ y_i = -1 \end{aligned}$$

which can be combined as

$$1 - y_i\,(\mathbf{w}\cdot\mathbf{x}_i + b) \leqslant 0, \quad \forall\, i = 1,\, 2,\, \ldots,\, m$$

Theorem. The distance between planes $\mathbf{w}\cdot\mathbf{x}=b_1$ and $\mathbf{w}\cdot\mathbf{x}=b_2$ is $\dfrac{|b_1-b_2|}{\|\mathbf{w}\|}$.

Proof. For $\mathbf{x}_1, \mathbf{x}_2$ s.t. $\mathbf{w}\cdot\mathbf{x}_1 = b_1$, $\mathbf{w}\cdot\mathbf{x}_2 = b_2$ and $\overline{\mathbf{x}_1\mathbf{x}_2}$ be the shortest path, $\exists\, t \in \mathbb{R}$ such that $\mathbf{x}_1 - \mathbf{x}_2 = t\,\mathbf{w} \implies b_1 - b_2 = \mathbf{w}\cdot(\mathbf{x}_1 - \mathbf{x}_2) = t\,\mathbf{w}\cdot\mathbf{w} = t\,\|\mathbf{w}\|^2 \implies t = \dfrac{b_1 - b_2}{\|\mathbf{w}\|^2}$. So the distance is $\|t\,\mathbf{w}\| = \dfrac{|b_1 - b_2|}{\|\mathbf{w}\|}$.

The margin between $\mathbf{w}\cdot\mathbf{x} = 1 - b$ and $\mathbf{w}\cdot\mathbf{x} = -1 - b$ is simply $\dfrac{2}{\|\mathbf{w}\|}$.

Determine the hyperplane with maximum margin

$$\begin{aligned} &\text{maximize } \frac{1}{\|\mathbf{w}\|} \iff \text{minimize } \frac{1}{2}\,\|\mathbf{w}\|^2 \\ &\text{subject to} \quad 1 - y_i\,(\mathbf{w}\cdot\mathbf{x}_i + b) \leqslant 0 \quad \forall\, i = 1,\, 2,\, \ldots,\, m. \end{aligned}$$

Set the Lagrangian $\mathcal{L}$

$$\mathcal{L} = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{m} \lambda_i \left\{1 - y_i\,(\mathbf{w}\cdot\mathbf{x}_i + b)\right\} \tag{1}$$

The KKT conditions are

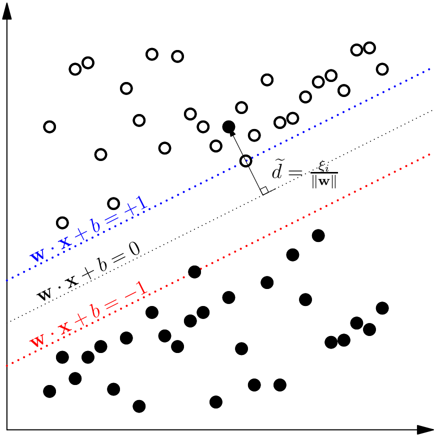
$$
\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 \implies \mathbf{w} = \sum_{i=1}^{m} \lambda_i \, y_i \, \mathbf{x}_i \\
&\frac{\partial \mathcal{L}}{\partial b} = 0 \implies \sum_{i=1}^{m} \lambda_i \, y_i = 0 \\
&1 - y_i \left(\mathbf{w} \cdot \mathbf{x}_i + b \right) \leqslant 0, \quad i = 1,\, 2,\, \dots,\, m \\
&\lambda_i \geqslant 0, \quad i = 1,\, 2,\, \dots,\, m \\
&\lambda_i \left\{ 1 - y_i \left(\mathbf{w} \cdot \mathbf{x}_i + b \right) \right\} = 0, \quad i = 1,\, 2,\, \dots,\, m
\end{aligned} \tag{2}
$$

Substitute the KKT conditions (2) into $\mathcal{L}$ (1), the dual Lagrangian

$$
\mathcal{L}_D = \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j \, y_i \, y_j \, \mathbf{x}_i^\top \mathbf{x}_j
$$

Now the problem becomes maximizing $\mathcal{L}_D$ subject to $\displaystyle\sum_{i=1}^{m} \lambda_i y_i = 0$.

SVM: Linearly Non-Separable (Soft-Margin)



Introduce *positive slack variables* $\{\xi_i\}_{i=1}^m$ into the constraints

$$\begin{aligned} \mathbf{w}\cdot\mathbf{x}_i + b &\geqslant +1 - \xi_i && \text{for } \ y_i = +1 \\ \mathbf{w}\cdot\mathbf{x}_i + b &\leqslant -1 + \xi_i && \text{for } \ y_i = -1 \\ \xi_i &\geqslant 0 && i = 1,\, 2,\, \ldots,\, m. \end{aligned}$$

The first two can be combined into

$$1 - \xi_i - y_i\,(\mathbf{w}\cdot\mathbf{x}_i + b) \leqslant 0, \qquad i = 1,\, 2,\, \ldots,\, m$$

If error occurs, $\xi_i > 1$; the objective function is changed to

$$\text{minimize} \quad \frac{1}{2}\,\|\mathbf{w}\|^2 + c\sum_{i=1}^{m}\xi_i$$

where $c > 0$ controls the tolerance to errors on the training set. The Lagrangian with $2m$ multipliers $\lambda_i \geqslant 0$ and KKT conditions are

$$\mathcal{L} = \frac{1}{2}\,\|\mathbf{w}\|^2 + c\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\lambda_i\,\{1 - \xi_i - y_i\,(\mathbf{w}\cdot\mathbf{x}_i + b)\} - \sum_{i=1}^{m}\lambda_{m+i}\,\xi_i \tag{3}$$

$$
\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 \implies \mathbf{w} = \sum_{i=1}^{m} \lambda_i \, y_i \, \mathbf{x}_i \\
&\frac{\partial \mathcal{L}}{\partial b} = 0 \implies \sum_{i=1}^{m} \lambda_i y_i = 0 \\
&\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \implies c - \lambda_i - \lambda_{m+i} = 0, \qquad i = 1,\, 2,\, \dots,\, m \\
&\lambda_i \left\{ 1 - \xi_i - y_i \left(\mathbf{w} \cdot \mathbf{x}_i + b \right) \right\} = 0, \qquad i = 1,\, 2,\, \dots,\, m \\
&1 - \xi_i - y_i \left(\mathbf{w} \cdot \mathbf{x}_i + b \right) \leqslant 0, \qquad i = 1,\, 2,\, \dots,\, m \\
&\lambda_{m+i} \, \xi_i = 0, \quad \xi_i \geqslant 0, \qquad i = 1,\, 2,\, \dots,\, m \\
&\lambda_i \geqslant 0 \qquad i = 1,\, 2,\, \dots,\, 2m
\end{aligned}
\tag{4}
$$

Substitute the KKT conditions (4) into $\mathcal{L}$ (3), the dual Lagrangian

$$
\mathcal{L}_D = \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \, \lambda_j \, y_i \, y_j \, \mathbf{x}_i^{\top} \mathbf{x}_j
$$

Now the problem becomes maximizing $\mathcal{L}_D$ subject to

$$
0 \leqslant \lambda_i \leqslant c \quad \text{and} \quad \sum_{i=1}^{m} \lambda_i y_i = 0.
$$

Nonlinear SVM: Kernel Trick

- When the seperating boundary is not linear, map the data into another space $\mathcal{H}$ and perform classification there
- Say the mapping function be $\Psi : \mathbb{R}^d \rightarrow \mathcal{H}$, the training algorithm now depends on $\Psi(\mathbf{x}_i) \cdot \Psi(\mathbf{x}_j)$
- If there were a "kernel function" K such that $K(\mathbf{x}_i, \mathbf{x}_j) = \Psi(\mathbf{x}_i) \cdot \Psi(\mathbf{x}_j)$, we don't need to know the exact form of Ψ
- **Mercer's condition**: $K(\mathbf{x}_i, \mathbf{x}_j) = \Psi(\mathbf{x}_i) \cdot \Psi(\mathbf{x}_j) \iff$
$$\int K(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \geqslant 0 \quad \text{for square integrable functions } g$$
- kernel examples:

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^\top \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$$
$$K(\mathbf{x}_i, \mathbf{x}_j) = \left(\mathbf{x}_i^\top \mathbf{x}_j + 1\right)^p$$
$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh\left(k\, \mathbf{x}_i^\top \mathbf{x}_j + \delta\right)$$