Support Vector Machine

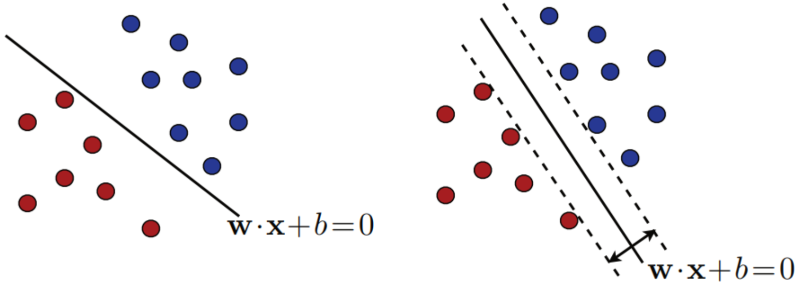
Wu Kaixiang

College of Electronic Science and Enginering,Jilin University

1. Hard-Margin SVM
2. Soft-margin SVM
3. Soft-SVM: Dual Problem
4. Soft-SVM: Dual Problem
5. Hinge Loss Function

Which is the Best Classifier?

Which classifier is the best classifier, ie. will yield the lowest test error?



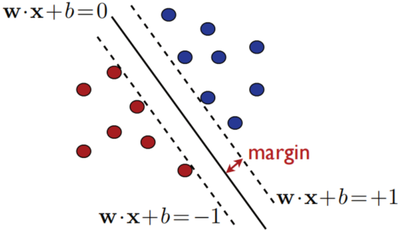
- This classifier has the **largest margin** to training data.
- This classifier is **the most robust classifier** to the noisy data.

Support Vector Machine: Margin

- Margin: Twice of the distance to the closest points of either class.
- Problem: How to find the linear classifier with the largest margin?
- Requirements:
    - The margin is the largest.
    - Classify all data points correctly.
- Constrained optimization problem:

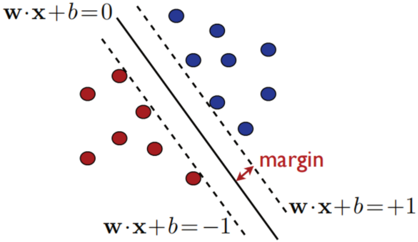
$$\max_{w,b} \operatorname{margin}(\boldsymbol{w}, b)$$

$$\text{s.t. } y_i \left(\boldsymbol{w} \cdot \boldsymbol{x}_i + b\right) \geq 1, 1 \leq i \leq n$$



Support Vector Machine: Margin

- How to quantify the margin?



$$r = \frac{|\boldsymbol{w} \cdot \boldsymbol{x} + b|}{\|\boldsymbol{w}\|_2}$$

$$\gamma = \frac{1}{\|\boldsymbol{w}\|_2} + \frac{|-1|}{\|\boldsymbol{w}\|_2} = \frac{2}{\|\boldsymbol{w}\|_2}$$

Hard-margin Support Vector Machine

- **Hard-margin** Support Vector Machine:

$$\max_{\boldsymbol{w},b} \frac{2}{\|\boldsymbol{w}\|_2}$$
$$\text{s.t. } y_i\left(\boldsymbol{w}\cdot\boldsymbol{x}_i+b\right)\geq 1, 1\leq i\leq n$$

- Non-convex problem. But it is equivalent to:

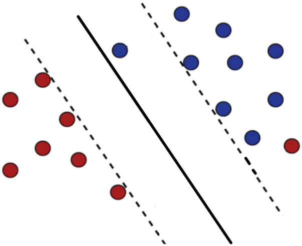
$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|_2^2$$
$$\text{s.t. } y_i\left(\boldsymbol{w}\cdot\boldsymbol{x}_i+b\right)\geq 1, 1\leq i\leq n$$

- This is only for the linearly separable case. Hardly used in practice!

1 Hard-Margin SVM
2 Soft-margin SVM
3 Soft-SVM: Dual Problem
4 Soft-SVM: Dual Problem
5 Hinge Loss Function

Soft-margin Support Vector Machine

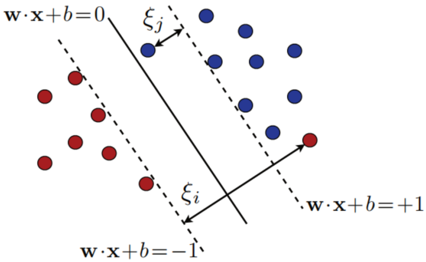
- In the linearly non-separable case, we **cannot find a solution** to the hard-margin support vector machine (left).



- Instead of constraining all data points to be correctly classified:
  - Allow some points on the wrong side of the margin.
  - Their number should be small.

Soft-margin Support Vector Machine

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}\|\boldsymbol{w}\|_2^2$$
$$\text{s.t. } y_i\left(\boldsymbol{w}\cdot\boldsymbol{x}_i + b\right) \ge 1 - \xi_i$$
$$\xi_i \ge 0, \sum_{i=1}^{n} \xi_i \le n'$$
$$1 \le i \le n$$



- Slack variables: $\xi_i, i \in [n]$
- Computationally, we re-express in the (Lagrangian) equivalent form:

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{n} \xi_i$$
$$\text{s.t. } y_i\left(\boldsymbol{w}\cdot\boldsymbol{x}_i + b\right) \ge 1 - \xi_i$$
$$\xi_i \ge 0, 1 \le i \le n$$

- C: penalty parameter

1. Hard-Margin SVM
2. Soft-margin SVM
3. Soft-SVM: Dual Problem
4. Soft-SVM: Dual Problem
5. Hinge Loss Function

Soft-SVM: Dual Problem

Soft-margin SVM

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{s.t. } y_i\left(\boldsymbol{w}\cdot\boldsymbol{x}_i + b\right) \geq 1 - \xi_i$$
$$\xi_i \geq 0, 1 \leq i \leq n$$

Lagrangian function (with 2n inequality constraints):

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu})$$
$$= \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - \xi_i - y_i\left(\boldsymbol{w}\cdot\boldsymbol{x}_i + b\right)\right) - \sum_{i=1}^{n}\mu_i\xi_i$$
$$\alpha_i \geq 0, \mu_i \geq 0, i = 1, \ldots, n$$

1. Hard-Margin SVM
2. Soft-margin SVM
3. Soft-SVM: Dual Problem
4. Soft-SVM: Dual Problem
5. Hinge Loss Function

Soft-SVM: Dual Problem

- Take the partial derivatives of Lagrangian w.rt $\boldsymbol{w}, b, \xi_i$ and set to zero:

$$
\begin{aligned}
\nabla_{\boldsymbol{w}} L = \mathbf{0} &\Rightarrow \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \\
\nabla_b L = 0 &\Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \\
\nabla_{\boldsymbol{\xi_i}} L = 0 &\Rightarrow C = \alpha_i + \mu_i, i = 1, \ldots, n
\end{aligned}
$$

Soft-SVM: Dual Problem

- Dual Problem of Soft-SVM:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \left(\boldsymbol{x}_i \cdot \boldsymbol{x}_j \right)$$

$$\text{s.t.} \ \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$0 \le \alpha_i \le C, 1 \le i \le n$$

After solving for $\boldsymbol{\alpha}$, we can solve for

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \boldsymbol{x}_i, b^* = y_j - \sum_{i=1}^{n} \alpha_i^* y_i (\boldsymbol{x}_i \cdot \boldsymbol{x}_j)$$

- Solved by Quadratic Program with linear constraints: slow!
- Solved by Sequential Minimal Optimization (SMO): fast!

1 Hard-Margin SVM

2 Soft-margin SVM

3 Soft-SVM: Dual Problem

4 Soft-SVM: Dual Problem

5 Hinge Loss Function

Hinge Loss Function

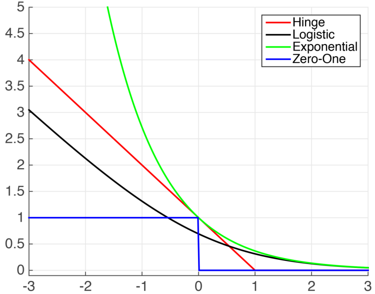
$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{s.t. } y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0, 1 \leq i \leq n$$

$$\Updownarrow$$

$$\min_{\boldsymbol{w},b} \lambda\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{n}[1 - y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b)]_+$$



- Hinge loss $\ell(f(\boldsymbol{x}), y) = \max\{0, 1 - yf(\boldsymbol{x})\}$

1 Hard-Margin SVM

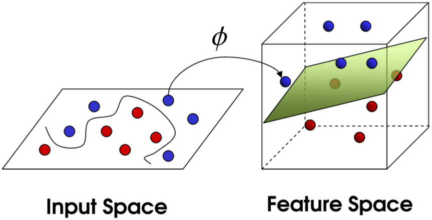
2 Soft-margin SVM

3 Soft-SVM: Dual Problem

4 Soft-SVM: Dual Problem

5 Hinge Loss Function
- Kernel Trick

Polynomial Kernel



$\mathcal{X}$ is the input space and $\mathcal{H}$ is the feature space. If there exists a mapping from $\mathcal{X}$ to $\mathcal{H}$

$$\phi(x) : \mathcal{X} \to \mathcal{H}$$

such that for all $x, z \in \mathcal{X}$, the function $K(x, z)$ satisfies the condition

$$K(x, z) = \phi(x) \cdot \phi(z)$$

then $K(x, z)$ is a kernel function and $\phi(x)$ is a mapping function.

Polynomial Kernel

- Suppose the input space is $\mathbf{R}^2$ and the kernel function is

$$K(x,z) = (x \cdot z).$$

Try to find the feature space $\mathcal{H}$ and the mapping function $\phi(x)$.

▶ Take the feature space $\mathcal{H} = \mathbf{R}^3$, $x = (x^{(1)}, x^{(2)})^T$,$z = (z^{(1)}, z^{(2)})^T$

$$\begin{aligned} (x \cdot z)^2 &= \left(x^{(1)} z^{(1)} + x^{(2)} z^{(2)}\right)^2 \\ &= \left(x^{(1)} z^{(1)}\right)^2 + 2x^{(1)} z^{(1)} x^{(2)} z^{(2)} + \left(x^{(2)} z^{(2)}\right)^2 \end{aligned} \tag{1}$$

So you can take the mapping function

$$\phi(x) = \left(\left(x^{(1)}\right)^2, \sqrt{2}x^{(1)}x^{(2)}, \left(x^{(2)}\right)^2\right)^{\mathrm{T}}$$

Easy to verify $\phi(x) \cdot \phi(z) = (x \cdot z)^2 = K(x,z)$

Kernel Trick in SVM

- Dual Problem of Soft-SVM:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \left(x_i \cdot x_j \right)$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C, 1 \leq i \leq n$$

- Replace $(x_i \cdot x_j)$ with $K(x_i, x_j)$

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K \left(\boldsymbol{x}_i, \boldsymbol{x}_j \right)$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C, 1 \leq i \leq n$$

Kernel Matrix

- How to verify that a function can be used as a kernel function? Find its basis function ϕ? It is too hard for most kernel functions.
- Theorem(Mercer): If $k(\cdot,\cdot)$ is a symmetric function on space $\mathcal{X}\times\mathcal{X}$ then k is a kernel function
  For any input set

$$\left(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\right), \boldsymbol{K}=\left(\begin{array}{ccc} k\left(\boldsymbol{x}_1, \boldsymbol{x}_1\right) & \cdots & k\left(\boldsymbol{x}_1, \boldsymbol{x}_m\right) \\ \vdots & \ddots & \vdots \\ k\left(\boldsymbol{x}_1, \boldsymbol{x}_m\right) & \cdots & k\left(\boldsymbol{x}_m, \boldsymbol{x}_m\right) \end{array}\right)$$

  The kernel matrix is **semi-definite**.
- For kernel functions $k_1, k_2, ...k_s$ and $\gamma_1, \gamma_2, ..., \gamma_s > 0$
  $\sum_{i=1}^{S}\gamma_i k_i$ is also (multi-)kernel function, because $\sum_{i=1}^{S}\gamma_i \boldsymbol{K}_i \geq 0$

Guassian Kernel function

$$K(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right)$$

吉林大学
JILIN UNIVERSITY

THANKS

— Wu Kaixiang —