ML&MEA (2024)
Lecture 7 - Support Vector Machine (SVM)

Quanying Liu

BME, SUSTech

2024.3.19

Content

1. Recap
2. The core ideas of SVM
3. Solving SVM with KKT
4. *Lagrangian and KKT conditions
5. Summary

1. Recap
2. The core ideas of SVM
3. Solving SVM with KKT
4. *Lagrangian and KKT conditions
5. Summary

Cross-entropy loss & its gradient

Given a single training data $(\mathbf{x}, y)$, the cross-entropy loss is:

$$\mathcal{L}(\beta) = H(y, h_\beta(x)) = -[y \log h_\beta(x) + (1-y)(1 - \log h_\beta(x))]$$

The gradient of cross-entropy loss is:

$$\begin{aligned}
\nabla \mathcal{L}(\beta) &= -\left[y \frac{1}{\sigma\left(\beta^T \mathbf{x}\right)} - (1-y) \frac{1}{1-\sigma\left(\beta^T \mathbf{x}\right)}\right] \frac{\partial}{\partial \beta} \sigma\left(\beta^T \mathbf{x}\right) \\
&= -[\cdots] \sigma\left(\beta^T \mathbf{x}\right)\left(1-\sigma\left(\beta^T \mathbf{x}\right)\right) \frac{\partial(\beta^T \mathbf{x})}{\partial \beta} \\
&= -[\cdots] \sigma\left(\beta^T \mathbf{x}\right)\left(1-\sigma\left(\beta^T \mathbf{x}\right)\right) \mathbf{x} \\
&= -\left(y\left(1-\sigma\left(\beta^T \mathbf{x}\right)\right)-(1-y) \sigma\left(\beta^T \mathbf{x}\right)\right) \mathbf{x} \\
&= -\left(y-h_\beta(\mathbf{x})\right) \mathbf{x}
\end{aligned}$$

GD for logistic regression

Model: $y = h_\beta(\mathbf{x}) = \sigma\left(\beta^T \mathbf{x}\right)$
Gradient of cross-entropy loss: $\nabla\mathcal{L}(\beta) = -\left(y - h_\beta(\mathbf{x})\right)\mathbf{x}$

The entire training data: $(\mathbf{X}, \mathbf{y})$

1. Initiate β
2. Calculate $h_\beta(\mathbf{x}^{(i)}) = \sigma\left(\beta^T \mathbf{x}^{(i)}\right)$ for for each $\mathbf{x}^{(i)}$
3. Calculate $\nabla\mathcal{L}(\beta)|_{(\mathbf{x}^{(i)}, y^{(i)})} = -\left(y^{(i)} - h_\beta(\mathbf{x}^{(i)})\right)\mathbf{x}^{(i)}$ for each data pair $(\mathbf{x}^{(i)}, y^{(i)})$
4. Calculate $\nabla\mathcal{L} = \sum_{i=1}^{i=n} \nabla\mathcal{L}(\beta)|_{(\mathbf{x}^{(i)}; y^{(i)})}$
5. Update β: $\quad \beta \leftarrow \beta + \eta\nabla\mathcal{L}$

For logistic regression, MLE = minimizing cross-entropy loss

1 Recap

2 The core ideas of SVM

3 Solving SVM with KKT

4 *Lagrangian and KKT conditions

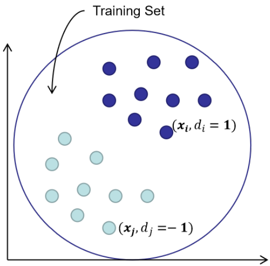
5 Summary

SVM Classification: Problem Definition

- Given a set $\mathcal{S}$ of n training samples, i.e.,

$$\mathcal{S} = \{(\mathbf{x}_i, d_i)\}_{i=1}^{n}$$

where $\mathbf{x}_i$ is the p-dimension feature vector, $d_i \in \{1, -1\}$ is the label of the ith sample.



- The main goal of SVM classification is to separate data with minimum classification error.

Sign function (符号函数)

- SVM is a two-class classification task.

Sign function (符号函数)

- SVM is a two-class classification task.
- 注意：**SVM** 中 y_i 只能是 1 或 -1。

$$d_i \in \{1, -1\} \text{ for } i = \{1, 2, \ldots, n\}$$

Sign function (符号函数)

- SVM is a two-class classification task.
- 注意：**SVM** 中 y_i 只能是 1 或 -1。

$$d_i \in \{1, -1\} \text{ for } i = \{1, 2, \dots, n\}$$

- Q: Which function can describe the desired output $\{1, -1\}$?

Sign function (符号函数)

- SVM is a two-class classification task.
- 注意：**SVM** 中 y_i 只能是 1 或 -1。

$$d_i \in \{1, -1\} \text{ for } i = \{1, 2, \dots, n\}$$

- Q: Which function can describe the desired output $\{1, -1\}$?
- Sign function (符号函数):

$$\text{sign}(z) \triangleq \begin{cases} +1, & \text{if} \quad z > 0 \\ -1, & \text{if} \quad z < 0 \end{cases},$$

where $\text{sign}(\cdot)$ is the sign function.

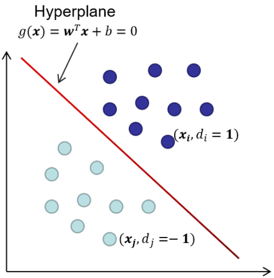
Hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$

- A hyperplane is defined by parameters $(\mathbf{w}, b)$, expressed as

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0 \qquad (1)$$

- A hyperplane classifies a given $\mathbf{x}_i$ with

$$\text{sign}(g(\mathbf{x}_i)) = \begin{cases} +1, & \text{if } g(\mathbf{x}_i) > 0 \\ -1, & \text{if } g(\mathbf{x}_i) < 0 \end{cases}$$



Two classes of data are **linearly separable** if and only if there exists a hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ that separates two classes.

Q: How to find such a hyperplane?

The optimal hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$

- Many potential hyperplanes can separate the two classes.

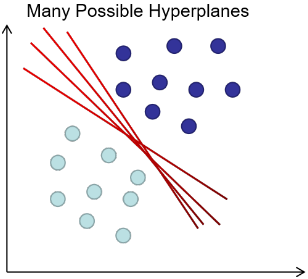


图 1: Hyperplanes

The optimal hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$

- Many potential hyperplanes can separate the two classes.

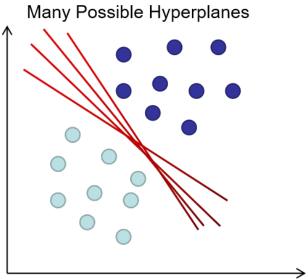


图 1: Hyperplanes

- Q: Which one is the optimal hyperplane?

The optimal hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$

- Many potential hyperplanes can separate the two classes.

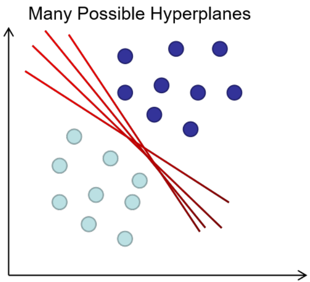


图 1: Hyperplanes

- Q: Which one is the optimal hyperplane?
- Now we introduce a concept called 'margin'.

What is margin? Intuitively, ...

- **Margin** is the core concept in SVM.

What is margin? Intuitively, ...

- **Margin** is the core concept in SVM.

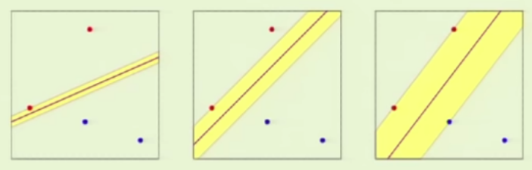


图 2: Margin (*e.g.*, the yellow shadowed area)

What is margin? Intuitively, ...

- **Margin** is the core concept in SVM.

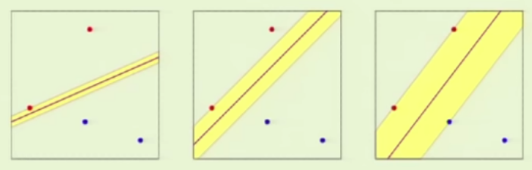


图 2: Margin (*e.g.*, the yellow shadowed area)

- Q: Which one is the optimal hyperplane?

What is margin? Intuitively, ...

- **Margin** is the core concept in SVM.

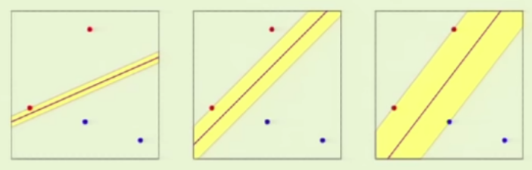


图 2: Margin (*e.g.*, the yellow shadowed area)

- Q: Which one is the optimal hyperplane?
- A: The hyperplane with the biggest margin.

Two questions:

What is margin? Intuitively, ...

- **Margin** is the core concept in SVM.

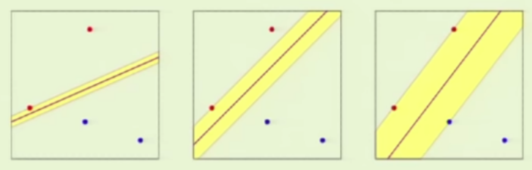


图 2: Margin (*e.g.*, the yellow shadowed area)

- Q: Which one is the optimal hyperplane?
- A: The hyperplane with the biggest margin.

Two questions:

1. Why is bigger margin better? $\longrightarrow$ robustness, generalibility

What is margin? Intuitively, ...

- **Margin** is the core concept in SVM.

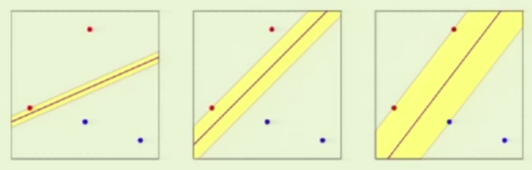


图 2: Margin (*e.g.*, the yellow shadowed area)

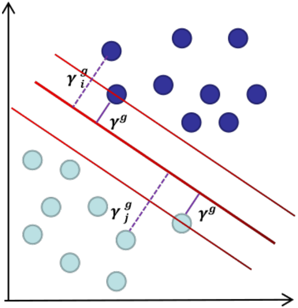
- Q: Which one is the optimal hyperplane?
- A: The hyperplane with the biggest margin.

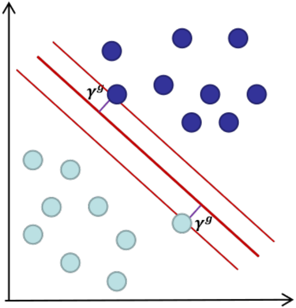
Two questions:

1. Why is bigger margin better? $\longrightarrow$ robustness, generalibility
2. Which parameters $(\boldsymbol{w}, b)$ maximize the margin?

Maximizing the margin

In SVM, the goal is to find the **optimal hyperplane** for a given dataset S where has the **maximal margin** over all possible hyperplanes $(\mathbf{w}, b)$.





We have to define margin mathematically.

What is margin? Mathematically, ...

- **Margin = Distance** between the data to the hyperplane.

What is margin? Mathematically, ...

- **Margin = Distance** between the data to the hyperplane.
- We notice that the hyperplane with $(\mathbf{w}, b)$ with $\mathbf{w}^T\mathbf{x} + b = 0$ is the same hyperplane with $(c\mathbf{w}, cb)$ with $c\mathbf{w}^T\mathbf{x} + cb = 0$, where c is any constant.

What is margin? Mathematically, ...

- **Margin = Distance** between the data to the hyperplane.
- We notice that the hyperplane with $(\mathbf{w}, b)$ with $\mathbf{w}^T\mathbf{x} + b = 0$ is the same hyperplane with $(c\mathbf{w}, cb)$ with $c\mathbf{w}^T\mathbf{x} + cb = 0$, where c is any constant.

Normalization technique:

$$|\mathbf{w}^T\mathbf{x}_i + b| = 1 \tag{2}$$

What is margin? Mathematically, ...

- **Margin = Distance** between the data to the hyperplane.
- We notice that the hyperplane with $(\mathbf{w}, b)$ with $\mathbf{w}^T\mathbf{x} + b = 0$ is the same hyperplane with $(c\mathbf{w}, cb)$ with $c\mathbf{w}^T\mathbf{x} + cb = 0$, where c is any constant.

Normalization technique:

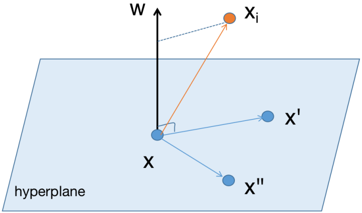
$$|\mathbf{w}^T\mathbf{x}_i + b| = 1 \tag{2}$$

- Recall some geometry knowledge.
  To get the **distance** between a point $\mathbf{x}_i \in \mathbb{R}^p$ and the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$, we just need project $\mathbf{x}_i$ to a vector orthogonal to the hyperplane.

What is margin? Mathematically, ...

Hyperplane: $\mathbf{w}^T\mathbf{x} + b = 0$
Normalization: $|\mathbf{w}^T\mathbf{x}_i + b| = 1$

Any two points in the hyperplane:
$\mathbf{w}^T\mathbf{x} + b = 0, \quad \mathbf{w}^T\mathbf{x}' + b = 0$
$\Longrightarrow \mathbf{w}^T(\mathbf{x} - \mathbf{x}') = 0$
$\Longrightarrow \mathbf{w}$ is $\perp$ to the hyperplane



Unit vector $\hat{\mathbf{w}} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$

$$\gamma_i = \text{distance} = |\hat{\mathbf{w}}^T(\mathbf{x}_i - \mathbf{x})| = \frac{1}{\|\mathbf{w}\|}|\mathbf{w}^T(\mathbf{x}_i - \mathbf{x})|$$
$$= \frac{1}{\|\mathbf{w}\|}|(\mathbf{w}^T\mathbf{x}_i + b) - (\mathbf{w}^T\mathbf{x} + b)| = \frac{1}{\|\mathbf{w}\|}$$

SVM: the constrained optimization problem

The constrained optimization problem:

$$\begin{aligned} \mathbf{w}^*, b^* &= \arg\max_{\mathbf{w}, b} \frac{1}{\|\mathbf{w}\|} \\ \text{s.t.,} \quad & \min_{i=1,2,\ldots,n} |\mathbf{w}^T \mathbf{x}_i + b| = 1 \end{aligned} \tag{3}$$

SVM: the constrained optimization problem

The constrained optimization problem:

$$\mathbf{w}^*, b^* = \arg\max_{\mathbf{w}, b} \frac{1}{\|\mathbf{w}\|}$$
$$\text{s.t.,} \quad \min_{i=1,2,\ldots,n} |\mathbf{w}^T \mathbf{x}_i + b| = 1 \tag{3}$$

Notice: $|\mathbf{w}^T \mathbf{x}_i + b| = d_i(\mathbf{w}^T \mathbf{x}_i + b)$

Maximize $\frac{1}{\|\mathbf{w}\|} \Longrightarrow$ Minimize $\mathbf{w}^T \mathbf{w}$

$$\mathbf{w}^*, b^* = \arg\min_{\mathbf{w}, b} \frac{1}{2} \mathbf{w}^T \mathbf{w}$$
$$\text{s.t.,} \; d_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 \quad \text{for } i = 1, 2, \ldots, n \tag{4}$$

1. Recap
2. The core ideas of SVM
3. Solving SVM with KKT
4. *Lagrangian and KKT conditions
5. Summary

Solving SVM with KKT conditions

The optimization problem for SVM in Eq. (4):

$$\mathbf{w}^*, b^* = \arg\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{s.t. } d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \geq 0$$

The **Lagrangian multipliers**: $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_n]^T$
The **generalized Lagrangian function** is:

$$\begin{aligned}\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{n}\alpha_i\left(d_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1\right)\\ &= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{n}\alpha_i d_i\mathbf{w}^T\mathbf{x}_i - b\sum_{i=1}^{n}\alpha_i d_i + \sum_{i=1}^{n}\alpha_i\end{aligned}$$

KKT conditions

The **generalized Lagrangian function** is:

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{n} \alpha_i d_i \mathbf{w}^T\mathbf{x}_i - b\sum_{i=1}^{n} \alpha_i d_i + \sum_{i=1}^{n} \alpha_i \quad (5)$$

KKT conditions:

$$\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} = \mathbf{0} \quad (6)$$

$$\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} = 0 \quad (7)$$

$$d_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 \quad (8)$$

$$\alpha_i\left(d_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1\right) = 0 \quad (9)$$

$$\alpha_i \geq 0 \quad (10)$$

The first term in KKT conditions

Consider the first KKT condition, Eq(6):

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} = \mathbf{0}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{n} \alpha_i d_i \mathbf{w}^T \mathbf{x}_i - b \sum_{i=1}^{n} \alpha_i d_i + \sum_{i=1}^{n} \alpha_i \right) \\
&= \frac{1}{2} \frac{\partial \left(\mathbf{w}^T\mathbf{w}\right)}{\partial \mathbf{w}} - \sum_{i=1}^{n} \alpha_i d_i \left(\frac{\partial \left(\mathbf{w}^T \mathbf{x}_i\right)}{\partial \mathbf{w}} \right) \\
&= \frac{2\mathbf{w}}{2} - \sum_{i=1}^{n} \alpha_i d_i \left(\frac{\partial \left(\mathbf{x}_i^T \mathbf{w}\right)}{\partial \mathbf{w}} \right) \\
&= \mathbf{w} - \sum_{i=1}^{n} \alpha_i d_i \mathbf{x}_i = \mathbf{0}
\end{aligned}$$

The second term in KKT conditions

Consider the second KKT condition, Eq(7):

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} = 0$$

$$\begin{aligned}\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} &= \frac{\partial}{\partial b}\left(\frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{n}\alpha_i d_i \mathbf{w}^T\mathbf{x}_i - b\sum_{i=1}^{n}\alpha_i d_i + \sum_{i=1}^{n}\alpha_i\right) \\ &= \sum_{i=1}^{n}\alpha_i d_i = 0\end{aligned}$$

Solving SVM with KKT conditions

So far, we have: $\mathbf{w} = \sum_{i=1}^{n} \alpha_i d_i \mathbf{x}_i$ and $\sum_{i=1}^{n} \alpha_i d_i = 0$

Recall the generalized Lagrangian function in Eq. (5)

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{n} \alpha_i d_i \mathbf{w}^T\mathbf{x}_i - b\sum_{i=1}^{n} \alpha_i d_i + \sum_{i=1}^{n} \alpha_i$$

Let's derive the first two terms of $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})$

$$\frac{1}{2}\mathbf{w}^T\mathbf{w} = \frac{1}{2}\left[\sum_{i=1}^{n} \alpha_i d_i \mathbf{x}_i^T\right]\left[\sum_{j=1}^{n} \alpha_j d_j \mathbf{x}_j\right] = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j d_i d_j \mathbf{x}_i^T\mathbf{x}_j$$

$$\sum_{i=1}^{n} \alpha_i d_i \mathbf{w}^T\mathbf{x}_i = \sum_{i=1}^{n} \alpha_i d_i \left[\sum_{j=1}^{n} \alpha_j d_j \mathbf{x}_j^T\right]\mathbf{x}_i = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j d_i d_j \mathbf{x}_i^T\mathbf{x}_j$$

Solving SVM with KKT conditions

Therefore, it becomes a dual problem of $Q(\boldsymbol{\alpha})$

$$
\begin{aligned}
\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{n} \alpha_i \left(d_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) - 1 \right) \\
&= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{n} \alpha_i d_i \mathbf{w}^T \mathbf{x}_i - b \sum_{i=1}^{n} \alpha_i d_i + \sum_{i=1}^{n} \alpha_i \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j \\
&\equiv Q(\boldsymbol{\alpha})
\end{aligned}
$$

We **maximize** $Q(\boldsymbol{\alpha})$ to get the optimal Lagrange multipliers $\boldsymbol{\alpha}^*$

The dual problem

The optimal Lagrange multipliers $\boldsymbol{\alpha}^*$:

$$\begin{aligned}\boldsymbol{\alpha}^* &= \arg\max_{\boldsymbol{\alpha}} Q(\boldsymbol{\alpha}) \\ &= \arg\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j\end{aligned}$$

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

$$\alpha_i \geq 0$$

It can be solved by quadratic programming using **Linear Quadratic Solver**.

Quadratic programming

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j - \sum_{i=1}^{n} \alpha_i$$

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

$$\arg\min_{\boldsymbol{\alpha}} \frac{1}{2} \boldsymbol{\alpha}^T \begin{bmatrix} d_1 d_1 \mathbf{x}_1^T \mathbf{x}_1 & d_1 d_2 \mathbf{x}_1^T \mathbf{x}_2 & \cdots & d_1 d_n \mathbf{x}_1^T \mathbf{x}_n \\ d_2 d_1 \mathbf{x}_2^T \mathbf{x}_1 & d_2 d_2 \mathbf{x}_2^T \mathbf{x}_2 & \cdots & d_2 d_n \mathbf{x}_2^T \mathbf{x}_n \\ \vdots & \vdots & \ddots & \vdots \\ d_n d_1 \mathbf{x}_n^T \mathbf{x}_1 & d_n d_2 \mathbf{x}_n^T \mathbf{x}_2 & \cdots & d_n d_n \mathbf{x}_n^T \mathbf{x}_n \end{bmatrix} \boldsymbol{\alpha} + (-\mathbf{1}^T)\boldsymbol{\alpha}$$

$$\text{s.t.} \quad \mathbf{d}^T \boldsymbol{\alpha} = 0$$

$$\mathbf{0} \leq \boldsymbol{\alpha} \leq \infty$$

1 Recap

2 The core ideas of SVM

3 Solving SVM with KKT

4 *Lagrangian and KKT conditions

5 Summary

Primal Problem

The **Primal Problem**:

$$\begin{aligned} \text{Minimize:} \quad & f(\mathbf{w}) \\ \text{Subject to:} \quad & q_i(\mathbf{w}) \leq 0, \quad i = 1, \ldots, k \\ & h_j(\mathbf{w}) = 0, \quad j = 1, \ldots, m \end{aligned} \tag{11}$$

We formulate the Generalized Lagrangian function:

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{w}) + \sum_{i=1}^{k} \alpha_i q_i(\mathbf{w}) + \sum_{j=1}^{m} \beta_j h_j(\mathbf{w}) \tag{12}$$

- α_i and β_i are the **Lagrange multipliers**.
- Goal: the constrained problem Eq. (11) $\longrightarrow$ a unconstrained problem Eq. (12).

Dual Problem (对偶问题)

$$\mathbf{w} = [w_1, w_2, \ldots, w_p]^T$$
$$\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_k]^T$$
$$\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_m]^T$$

Let's denote p^* as the solution for problem Eq. (11);
d^* as the dual problem Eq. (11).

Theorem (' 宁做凤尾不做鸡头' 定理):

$$d^* = \max_{\boldsymbol{\alpha},\boldsymbol{\beta},\alpha_i \geq 0} \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq \min_{\mathbf{w}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta},\alpha_i \geq 0} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = p^* \tag{13}$$

KKT conditions $\Longleftrightarrow d^* = p^*$

KKT conditions

$$d^* = \max_{\boldsymbol{\alpha},\boldsymbol{\beta},\alpha_i \geq 0} \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq \min_{\mathbf{w}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta},\alpha_i \geq 0} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = p^*$$

(充分必要条件) **KKT conditions** $\Longleftrightarrow$ $d^* = p^*$

Kuhn-Tucker Theorem: The solution meets KKT

$$\frac{\partial \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \mathbf{w}} = 0, \quad \frac{\partial \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{\alpha}} = 0, \quad \frac{\partial \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$$

$$\alpha_i q_i(\mathbf{w}) = 0, \quad i = 1, \ldots, k$$

$$\alpha_i \geq 0, \quad i = 1, \ldots, k$$

$$q_i(\mathbf{w}) \leq 0, \quad i = 1, \ldots, k$$

1. Recap
2. The core ideas of SVM
3. Solving SVM with KKT
4. *Lagrangian and KKT conditions
5. Summary

Solving with KKT conditions

The dual problem of $Q(\boldsymbol{\alpha})$

$$\text{Maximize: } Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{Subject to: } \sum_{i=1}^{N} \alpha_i d_i = 0 \text{ and } \alpha_i \geq 0$$

Please watch the following youtube video

`https://www.youtube.com/watch?v=eHsErlPJWUU&list=PLD63A284B7615313A&index=14`