From SVM to SMO and Random Feature Kernel Approximation

106033233 資工21 周聖諺

Abstract

Lagrange Multiplier

Karush, Kuhn, Tucker(KKT) Condition

Hard-Margin SVM

Soft-Margin SVM

Kernel Trick

Sequential Minimal Optimization(SMO)

Based on the paper **Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines**.

We've known he dual problem of soft-SVM is

$$\sup_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$
$$\text{subject to } 0 \leq \alpha_i \leq C, \sum_{i=1}^{N} \alpha_i y_i = 0$$

We also define the kernel.

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

where ϕ is an embedding function projecting the data points to a high dimensional space.

However, it's very hard to solve because we need to optimize N variables.

Notation

We denote the target function as $\mathcal{L}_d(\alpha, C)$

$$\mathcal{L}_d(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$

We also denote the kernel of x_1, x_2 as $K_{1,2} = k(x_1, x_2)$.

Step 1. Update 2 Variable

First, we need to pick 2 variables to update in sequence, so we split the variables α_1, α_2 from the summation.

$$\begin{aligned}
\mathcal{L}_d(\alpha) = \alpha_1 + \alpha_2 - \frac{1}{2}\alpha_1^2 y_1^2 K_{1,1} - \frac{1}{2}\alpha_2^2 y_2^2 K_{2,2} \\
- \frac{1}{2}\alpha_1 \alpha_2 y_1 y_2 K_{1,2} - \frac{1}{2}\alpha_2 \alpha_1 y_2 y_1 K_{2,1} \\
- \frac{1}{2}\alpha_1 y_1 \sum_{i=3}^{N} \alpha_i y_i K_{i,1} - \frac{1}{2}\alpha_1 y_1 \sum_{i=3}^{N} \alpha_i y_i K_{1,i} \\
- \frac{1}{2}\alpha_2 y_2 \sum_{i=3}^{N} \alpha_i y_i K_{i,2} - \frac{1}{2}\alpha_2 y_2 \sum_{i=3}^{N} \alpha_i y_i K_{2,i} \\
+ \sum_{i=3}^{N} \alpha_i - \frac{1}{2} \sum_{i=3}^{N} \sum_{j=3}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j)
\end{aligned}$$

$$\begin{aligned}
= \alpha_1 + \alpha_2 - \frac{1}{2}\alpha_1^2 y_1^2 K_{1,1} - \frac{1}{2}\alpha_2^2 y_2^2 K_{2,2} - \alpha_1 \alpha_2 y_1 y_2 K_{1,2} \\
- \alpha_1 y_1 \sum_{i=3}^{N} \alpha_i y_i K_{i,1} - \alpha_2 y_2 \sum_{i=3}^{N} \alpha_i y_i K_{i,2} + \mathcal{C}onst
\end{aligned}$$

$$
\begin{aligned}
&= \alpha_1 + \alpha_2 - \frac{1}{2}\alpha_1^2 K_{1,1} - \frac{1}{2}\alpha_2^2 K_{2,2} - \alpha_1\alpha_2 y_1 y_2 K_{1,2} \\
&- \alpha_1 y_1 \sum_{i=3}^{N} \alpha_i y_i K_{i,1} - \alpha_2 y_2 \sum_{i=3}^{N} \alpha_i y_i K_{i,2} + \mathcal{C}onst
\end{aligned}
$$

where $\mathcal{C}onst = \sum_{i=3}^{N} \alpha_i - \frac{1}{2}\sum_{i=3}^{N}\sum_{j=3}^{N} \alpha_i\alpha_j y_i y_j k(x_i, x_j)$. We see it as a constant because it is regardless to α_1, α_2.

The Relation Between The Updated Values and The Hyperplane

We've derive the partial derivative of the dual problem.

$$
\frac{\partial L(w, b, \xi, \alpha, \mu)}{\partial w} = w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0
$$

We can get

$$
w = \sum_{i=1}^{N} \alpha_i y_i x_i
$$

Thus, we can rewrite the hyperplane $f_\phi(x)$ with kernel.

$$
f_\phi(x) = w^\top \phi(x) + b = b + \sum_{i=1}^{N} \alpha_i y_i k(x_i, x)
$$

We also denote v_1, v_2 as

$$
\begin{aligned}
v_1 = \sum_{i=3}^{N} \alpha_i y_i K_{i,1} &= \sum_{i=1}^{N} \alpha_i y_i k(x_i, x_1) - \alpha_1^{old} y_1 k(x_1, x_1) - \alpha_2^{old} y_2 k(x_2, x_1) \\
&= f_\phi(x_1) - b - \alpha_1^{old} y_1 K_{1,1} - \alpha_2^{old} y_2 K_{2,1}
\end{aligned}
$$

and v_2 is similar.

$$
\begin{aligned}
v_2 = \sum_{i=3}^{N} \alpha_i y_i K_{i,2} &= \sum_{i=1}^{N} \alpha_i y_i k(x_i, x_2) - \alpha_1^{old} y_1 k(x_1, x_2) - \alpha_2^{old} y_2 k(x_2, x_2) \\
&= f_\phi(x_2) - b - \alpha_1^{old} y_1 K_{1,2} - \alpha_2^{old} y_2 K_{2,2}
\end{aligned}
$$

where α_1^{old} and α_2^{old} are α_1 and α_2 of the previous iteration. Since we see $\alpha_i, i \geq 3$ as constant, α_i shouldn't depends on update variables α_1, α_2 .

Rewrite The Complementary Slackness

The constraint can be represented as

$$
\sum_{i=1}^{N} \alpha_i y_i = \alpha_1 y_1 + \alpha_2 y_2 + \sum_{i=3}^{N} \alpha_i y_i = 0
$$

$$
\alpha_1 y_1 + \alpha_2 y_2 = -\sum_{i=3}^{N} \alpha_i y_i = \zeta
$$

$$
\alpha_1 = \frac{\zeta - \alpha_2 y_2}{y_1}
$$

Since y_1 is either 1 or -1, thus

$$
\alpha_1 = \zeta y_1 - \alpha_2 y_1 y_2
$$

The old ones are the same.

$$
\alpha_1^{old} = \zeta y_1 - \alpha_2^{old} y_1 y_2
$$

Replace the symbol α_1, v_1, v_2

$$
\begin{aligned}
\mathcal{L}_d(\alpha) &= (\zeta y_1 - \alpha_2 y_1 y_2) + \alpha_2 \\
&- \frac{1}{2}(\zeta y_1 - \alpha_2 y_1 y_2)^2 K_{1,1} - \frac{1}{2}\alpha_2^2 K_{2,2} - (\zeta y_1 - \alpha_2 y_1 y_2)\alpha_2 y_1 y_2 K_{1,2} \\
&- (\zeta y_1 - \alpha_2 y_1 y_2) y_1 v_1 - \alpha_2 y_2 v_2
\end{aligned}
$$

$$
\begin{aligned}
&= (\zeta y_1 - \alpha_2 y_1 y_2) + \alpha_2 \\
&- \frac{1}{2}(\zeta^2 + \alpha_2^2 - 2\zeta\alpha_2 y_2) K_{1,1} - \frac{1}{2}\alpha_2^2 K_{2,2} - (\zeta\alpha_2 y_2 - \alpha_2^2) K_{1,2} \\
&- (\zeta - \alpha_2 y_2) v_1 - \alpha_2 y_2 v_2
\end{aligned}
$$

Combine the v_1, v_2 and ζ

$$v_1 - v_2 = [\ f_\phi(x_1) - b - \alpha_1^{old} y_1 K_{1,1} - \alpha_2^{old} y_2 K_{2,1}\ ] - [\ f_\phi(x_2) - b - \alpha_1^{old} y_1 K_{1,2} - \alpha_2^{old} y_2 K_{2,2}\ ]$$
$$= [\ f_\phi(x_1) - b - (\zeta y_1 - \alpha_2^{old} y_1 y_2) y_1 K_{1,1} - \alpha_2^{old} y_2 K_{2,1}\ ] - [\ f_\phi(x_2) - b - (\zeta y_1 - \alpha_2^{old} y_1 y_2) y_1 K_{1,2} - \alpha_2^{old} y_2 K_{2,2}\ ]$$
$$= [\ f_\phi(x_1) - f_\phi(x_2)\ ] + [\ -(\zeta - \alpha_2^{old} y_2) K_{1,1} - \alpha_2^{old} y_2 K_{2,1}\ ] - [\ -(\zeta - \alpha_2^{old} y_2) K_{1,2} - \alpha_2^{old} y_2 K_{2,2}\ ]$$
$$= [\ f_\phi(x_1) - f_\phi(x_2)\ ] + [\ -\zeta K_{1,1} + \alpha_2^{old} y_2 K_{1,1} - \alpha_2^{old} y_2 K_{2,1}\ ] - [\ -\zeta K_{1,2} + \alpha_2^{old} y_2 K_{1,2} - \alpha_2^{old} y_2 K_{2,2}\ ]$$
$$= f_\phi(x_1) - f_\phi(x_2) - \zeta K_{1,1} + \zeta K_{1,2} + (K_{1,1} + K_{2,2} - 2K_{1,2}) \alpha_2^{old} y_2$$

Derive Gradient of α_2

$$\frac{\partial \mathcal{L}_d(\alpha)}{\partial \alpha_2} = -y_1 y_2 + 1 - \frac{1}{2}(2\alpha_2 - 2\zeta y_2) K_{1,1} - \alpha_2 K_{2,2} - (\zeta y_2 - 2\alpha_2) K_{1,2} - (-y_2) v_1 - y_2 v_2$$
$$= (-\alpha_2 K_{1,1} - \alpha_2 K_{2,2} + 2\alpha_2 K_{1,2}) + \zeta y_2 K_{1,1} - \zeta y_2 K_{1,2} - y_1 y_2 + y_2 v_1 - y_2 v_2 + 1$$
$$= -\alpha_2 (K_{1,1} + K_{2,2} - 2K_{1,2}) + \zeta y_2 K_{1,1} - \zeta y_2 K_{1,2} - y_1 y_2 + y_2 (v_1 - v_2) + 1$$

Replace with old α

$$= -\alpha_2 (K_{1,1} + K_{2,2} - 2K_{1,2}) + \zeta y_2 K_{1,1} - \zeta y_2 K_{1,2} - y_1 y_2 + y_2 [\ f_\phi(x_1) - f_\phi(x_2) - \zeta K_{1,1} + \zeta K_{1,2} + (K_{1,1} + K_{2,2} - 2K_{1,2}) \alpha_2^{old} y_2\ ] + 1$$
$$= -(K_{1,1} + K_{2,2} - 2K_{1,2}) \alpha_2 + (K_{1,1} + K_{2,2} - 2K_{1,2}) \alpha_2^{old} + y_2 (f_\phi(x_1) - f_\phi(x_2) + y_2 - y_1)$$

Let η and E_i be

$$\eta = K_{1,1} + K_{2,2} - 2K_{1,2}, \quad E_i = f_\phi(x_i) - y_i$$
$$\frac{\partial \mathcal{L}_d(\alpha)}{\partial \alpha_2} = -\eta \alpha_2 + \eta \alpha_2^{old} + y_2 (E_1 - E_2)$$

Since we want to minimize the gradient, let the gradient be 0.

$$-\eta \alpha_2 + \eta \alpha_2^{old} + y_2 (E_1 - E_2) = 0$$

Then we can update α_2 as following

$$\alpha_2 = \alpha_2^{old} + \frac{y_2 (E_1 - E_2)}{\eta}$$

Step 2. Clip with Constraint

$$\alpha_1 y_1 + \alpha_2 y_2 = \zeta, \quad 0 \leq \alpha_i \leq C$$

Case 1: Inequality

When $y_1 \neq y_2$, the equation is either $\alpha_1 - \alpha_2 = k$ or $\alpha_1 - \alpha_2 = -k$ where k is a positive constant.

The upper bound can be written as

$$B_U = \min(C, C + \alpha_2^{old} - \alpha_1^{old})$$

and the lower bound is

$$B_L = \max(0, \alpha_2^{old} - \alpha_1^{old})$$

Case 2: Equality

When $y_1 = y_2$, the equation is either $\alpha_1 + \alpha_2 = k$ or $\alpha_1 + \alpha_2 = -k$ where k is a positive constant.

The upper bound can be written as

$$B_U = \min(C, \alpha_2^{old} + \alpha_1^{old})$$

and the lower bound is

$$B_L = \max(0, \alpha_2^{old} - \alpha_1^{old} - C)$$

Clip The Value

According the bound we've derived, we need **clip** the updated variable α_2^{new} to satisfy the constraint.

$$\alpha_2^* = CLIP(\alpha_2^{new}, B_L, B_U)$$

Update α_1

$$\alpha_1^* y_1 + \alpha_2^* y_2 = \alpha_1^{old} y_1 + \alpha_2^{old} y_2 = \zeta$$

$$\alpha_1^* = \frac{\alpha_1^{old} y_1 + \alpha_2^{old} y_2 - \alpha_2^* y_2}{y_1}$$

$$\alpha_1^* = \alpha_1^{old} + y_1 y_2 (\alpha_2^{old} - \alpha_2^*)$$

Step 3. Update Bias

Pseudo Code

Random Feature For Kernel Approximation

Based on the paper **Random Features for Large-Scale Kernel Machines** on NIPS'07.

Experiments

Reference

- Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines
- NIPS'07 - Random Features for Large-Scale Kernel Machines
- 現代啟示錄 - Karush-Kuhn-Tucker (KKT) 條件
- 現代啟示錄 - Lagrange 乘數法
- 之乎 - 机器学习算法实践-SVM中的SMO算法
- 之乎 - Python · SVM（四）· SMO 算法
- Machine Learning Techniques (機器學習技法)