Elective Module:

Advanced ML Techniques

Chapter 11

Support Vector Machine

Dual SVM and Kernel SVM

SVM Classifier

- Recall, the problem to determine classifier with **maximum margin** is

maximizes margin

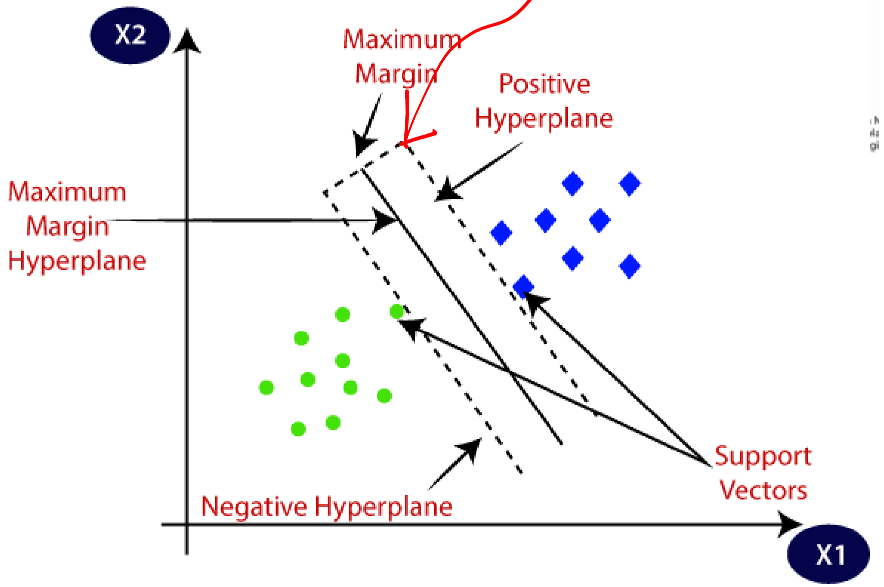
$$\min \|\bar{a}\|$$

$$\text{s.t. } C_0: \bar{a}^T \bar{x}_i + b \geq 1 \quad i = 1, .., M$$

$$C_1: \bar{a}^T \bar{x}_i + b \leq -1 \quad i = M+1, .., 2M$$

Positive hyperplane

Negative Hyperplane



SVM Classifier

- Recall, the problem to determine classifier with **maximum margin** is

Convex optimization

$$\min \|\bar{\mathbf{a}}\|_2$$

$$\mathcal{C}_{\mathbf{0}}: \bar{\mathbf{a}}^T \bar{\mathbf{x}}_i + b \geq 1, 1 \leq i \leq M$$

$$\mathcal{C}_{\mathbf{1}}: \bar{\mathbf{a}}^T \bar{\mathbf{x}}_i + b \leq -1, M+1 \leq i \leq 2M$$

SVM Response

- Let the response be defined as

$$\mathcal{C}_0 : \; y_i = 1, \quad i = 1, \ldots, M$$
$$\mathcal{C}_1 : \; y_i = -1, \quad i = M+1, \ldots, 2M$$

SVM Response

- Let the response be defined as

$$\mathcal{C}_0 : y_i = 1, 1 \leq i \leq M$$

$$\mathcal{C}_1 : y_i = -1, M+1 \leq i \leq 2M$$

SVM Constraints

- The constraints can be expressed as

$\mathcal{C}_0:$

$$1(\bar{a}^T\bar{x}_i + b) \geq 1$$

$$\Rightarrow y_i(\bar{a}^T\bar{x}_i + b) \geq 1$$

$\mathcal{C}_1:$

$$\bar{a}^T\bar{x}_i + b \leq -1$$

$$\Rightarrow -1(\bar{a}^T\bar{x}_i + b) \geq 1$$

$$\Rightarrow y_i(\bar{a}^T\bar{x}_i + b) \geq 1$$

constraints become same for both classes!

SVM Constraints

- The constraints can be expressed as

$$\mathcal{C}_{\mathbf{0}}: y_i(\bar{\mathbf{a}}^T \bar{\mathbf{x}}_i + b) \geq 1, 1 \leq i \leq M$$

$$\mathcal{C}_{\mathbf{1}}: y_i(\bar{\mathbf{a}}^T \bar{\mathbf{x}}_i + b) \geq 1, M + 1 \leq i \leq 2M$$

SVM Constraints

- The constraints can be combined as

$i = 1, 2, \ldots, 2M.$ For all i

$$y_i(\bar{a}^T\bar{x}_i + b_i) \geq 1$$

$$\Rightarrow -y_i(\bar{a}^T\bar{x}_i + b_i) \leq -1$$

$$\Rightarrow -(y_i(\bar{a}^T\bar{x}_i + b_i) - 1) \leq 0$$

Valid for all i

Final version of the constraint

SVM Constraints

- The constraints can be combined as

$$y_i(\bar{\mathbf{a}}^T\bar{\mathbf{x}}_i + b) \geq 1, 1 \leq i \leq 2M$$

$$\Rightarrow -y_i(\bar{\mathbf{a}}^T\bar{\mathbf{x}}_i + b) \leq -1$$

$$\Rightarrow -(y_i(\bar{\mathbf{a}}^T\bar{\mathbf{x}}_i + b) - 1) \leq 0$$

SVM Classifier

- The SVM classifier problem can be recast as

CONVEX

$$\min \|\bar{a}\| \equiv \min \frac{1}{2}\|\bar{a}\|^2$$

$$\text{s.t.} \quad -\left(y_i\left(\bar{a}^T\bar{x}_i + b\right) - 1\right) \leq 0$$

for all i

Equivalent SVM optimization problem

SVM Classifier

- The SVM classifier problem can be recast as

$$\min \frac{1}{2} \|\bar{\mathbf{a}}\|^2$$

subject to

$$-(y_i(\bar{\mathbf{a}}^T \bar{\mathbf{x}}_i + b) - 1) \leq 0, 1 \leq i \leq 2M$$

Lagrangian

- The **Lagrangian** for this problem is

Constrained Optimization problem

$$\frac{1}{2}\|\bar{a}\|^2 + \sum_{i=1}^{2M} \lambda_i \left(-\left(y_i (\bar{a}^T \bar{x}_i + b) - 1 \right) \right) \Big\} \text{ Lagrangian Function}$$

$$= \frac{1}{2}\|\bar{a}\|^2 - \sum_{i=1}^{2M} \lambda_i \left(y_i (\bar{a}^T \bar{x}_i + b) - 1 \right)$$

Lagrangian

- The **Lagrangian** for this problem is

$\frac{1}{2} \cdot \bar{a}^T \bar{a}$ $\quad \nabla\left(\frac{1}{2} \cdot \bar{a}^T \bar{a}\right) = \bar{a}$

$\lambda_i y_i \bar{a}^T \bar{x}_i \xrightarrow{\nabla} \lambda_i y_i \bar{x}_i$ $\quad \nabla = 0$

$$\frac{1}{2}\|\bar{\mathbf{a}}\|^2 - \sum_{i=1}^{2M} \lambda_i (y_i (\bar{\mathbf{a}}^T \bar{\mathbf{x}}_i + b) - 1)$$

Lagrangian

KKT

Karush Kuhn Tucker

- Setting gradient wrto $\bar{\mathbf{a}}$ to zero

$$\nabla_{\bar a}\left(\frac{1}{2}\bar a^T \bar a - \sum_i \lambda_i \left(y_i(\bar a^T \bar x_i + b) - 1\right)\right) = 0$$

$$\Rightarrow \bar a - \sum_i \lambda_i y_i \bar x_i = 0$$

$$\Rightarrow \bar a = \sum_i \lambda_i y_i \bar x_i$$

Lagrangian

- Setting gradient wrto $\bar{\mathbf{a}}$ to zero

$$\nabla_{\bar{\mathbf{a}}}\left(\frac{1}{2}\|\bar{\mathbf{a}}\|^2 - \sum_{i=1}^{2M} \lambda_i\left(y_i\left(\bar{\mathbf{a}}^T\bar{\mathbf{x}}_i + b\right) - 1\right)\right) = 0$$

$$\Rightarrow \bar{\mathbf{a}} - \sum_{i=1}^{2M} \lambda_i y_i\, \bar{\mathbf{x}}_i = 0 \Rightarrow \bar{\mathbf{a}} = \sum_{i=1}^{2M} \lambda_i y_i\, \bar{\mathbf{x}}_i$$

Lagrangian

$$\bar{\mathbf{a}} = \sum_{i=1}^{2M} \overset{\alpha_i}{\lambda_i y_i} \bar{\mathbf{x}}_i$$

$$\sum_i \alpha_i \bar{x}_i$$

- Thus, $\bar{\mathbf{a}}$ can be expressed as linear combination of $\bar{\mathbf{x}}_i$.

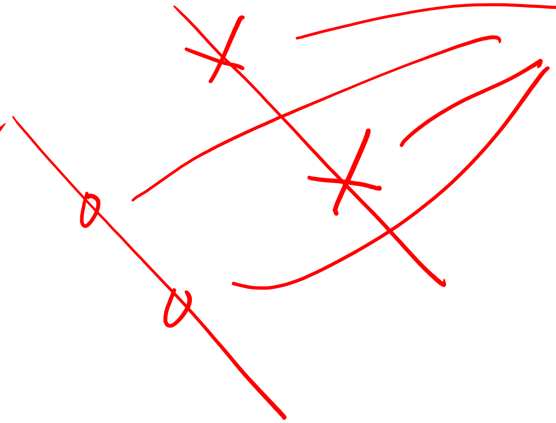
Lagrangian

$$\bar{\mathbf{a}} = \sum_{i=1}^{2M} \lambda_i y_i \bar{\mathbf{x}}_i$$

$\alpha_i \neq 0$

- The points for which $\lambda_i \neq 0$ are termed the **support vectors**.

Support vectors.



Linear combination of support vectors since $a_i \neq 0$.

Lagrangian

- Setting gradient wrto b to zero

$$\nabla_b \left(\underbrace{\frac{1}{2}\,\bar{a}^T\bar{a}} - \sum_i \lambda_i \left(y_i \left(\bar{a}^T \bar{x}_i + b \right) - 1 \right) \right) = 0$$

$$\sum_i \underbrace{\lambda_i y_i \bar{a}^T \bar{x}_i}_{\nabla = 0} + \underbrace{\lambda_i y_i b}_{\nabla = \lambda_i y_i}$$

$$\Rightarrow \sum_i \lambda_i y_i = 0$$

Lagrangian

- Setting gradient wrto b to zero

$$\nabla_b \left(\frac{1}{2} \|\bar{\mathbf{a}}\|^2 - \sum_{i=1}^{2M} \lambda_i (y_i (\bar{\mathbf{a}}^T \bar{\mathbf{x}}_i + b) - 1) \right) = 0$$

$$\Rightarrow \sum_{i=1}^{2M} \lambda_i y_i = 0$$

Lagrangian

$\sum_i \lambda_i y_i \bar{x}_i$

- The expression for $\bar{\mathbf{a}}$ can be substituted in the Lagrangian
- The resulting expression can be simplified as shown next

Original problem: Primal problem → min

Dual objective: maximization

Lagrangian

- Consider the **Lagrangian** given as

$$\frac{1}{2}\cdot\bar{a}^T\bar{a} - \sum_i \lambda_i\left(y_i\left(\bar{a}^T\bar{x}_i + b\right) - 1\right)$$

Lagrangian

- Consider the **Lagrangian** given as

$$\frac{1}{2}\|\bar{\mathbf{a}}\|^2 - \sum_{i=1}^{2M} \lambda_i (y_i(\bar{\mathbf{a}}^T \bar{\mathbf{x}}_i + b) - 1)$$

$$= \frac{1}{2}\bar{\mathbf{a}}^T \bar{\mathbf{a}} - \sum_{i=1}^{2M} \lambda_i (y_i(\bar{\mathbf{a}}^T \bar{\mathbf{x}}_i + b) - 1)$$

$$\sum_i \lambda_i y_i \bar{x}_i$$

Lagrangian

- Substitute $\bar{\mathbf{a}} = \sum_{i=1}^{2M} \lambda_i y_i \bar{\mathbf{x}}_i$

$$
\begin{aligned}
\text{Lagrangian} &= \frac{1}{2}\left(\sum_i \lambda_i y_i \bar{x}_i\right)^T \left(\sum_j \lambda_j y_j \bar{x}_j\right) \\
&\quad - \sum_i \lambda_i \left(y_i\left(\left(\sum_j \lambda_j y_j \bar{x}_j\right)^T \bar{x}_i + b\right) - 1\right)
\end{aligned}
\quad \Bigg\} \; \underline{\text{Messy!}}
$$

Lagrangian

- Substitute $\bar{\mathbf{a}} = \sum_{i=1}^{2M} \lambda_i y_i \,\bar{\mathbf{x}}_i$

$$= \frac{1}{2}\left(\sum_{i=1}^{2M} \lambda_i y_i \,\bar{\mathbf{x}}_i\right)^T \left(\sum_{i=1}^{2M} \lambda_j y_j \,\bar{\mathbf{x}}_j\right)$$

$$- \sum_{i=1}^{2M} \lambda_i \left(y_i \left(\left(\sum_{j=1}^{2M} \lambda_j y_j \,\bar{\mathbf{x}}_j\right)^T \bar{\mathbf{x}}_i + b \right) - 1 \right)$$

Lagrangian

$$= \sum_i \lambda_i - \frac{1}{2} \sum_{i,j=1}^{2M} \lambda_i \lambda_j y_i y_j \bar{x}_i^T \bar{x}_j - b \underbrace{\sum_i \lambda_i y_i}_{0}$$

$$\underbrace{= \sum_i \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \bar{x}_i^T \bar{x}_j}_{\text{Dual Objective:}} \Bigg\} \text{max.}$$

Lagrangian

$$= \sum_{i=1}^{2M} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{2M} \lambda_i \lambda_j y_i y_j \bar{\mathbf{x}}_i^T \bar{\mathbf{x}}_j - b \sum_{i=1}^{2M} \lambda_i y_i$$

$$= \sum_{i=1}^{2M} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{2M} \lambda_i \lambda_j y_i y_j \bar{\mathbf{x}}_i^T \bar{\mathbf{x}}_j$$

Dual SVM

- Therefore, the **dual problem** can be formulated as

Dual Problem

$$\max. \sum_i \lambda_i - \sum_i \sum_j \lambda_i \lambda_j y_i y_j \bar{x}_i^T \bar{x}_j$$

$$\text{s.t.} \quad \lambda_i \geq 0$$: Lagrange multipliers ≥ 0 non-negative

$$\sum_i \lambda_i y_i = 0$$

Dual SVM

Variables: λ_i

Lagrange multipliers.

- Therefore, the **dual problem** can be formulated as

$$\max \sum_{i=1}^{2M} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{2M} \lambda_i \lambda_j y_i y_j \bar{\mathbf{x}}_i^T \bar{\mathbf{x}}_j$$

$$\text{subject to } \lambda_i \geq 0$$

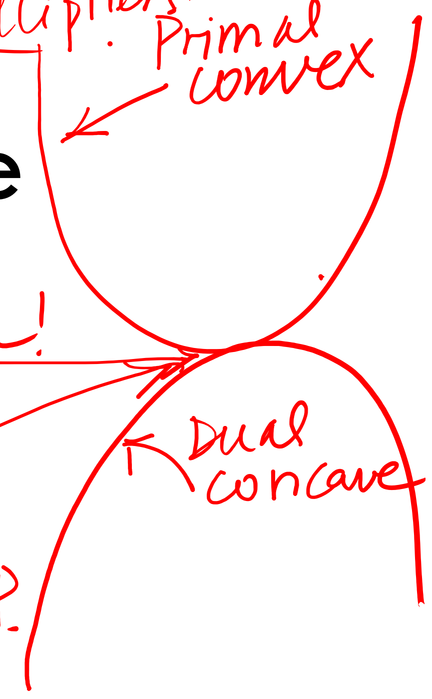
$$\sum_{i=1}^{2M} \lambda_i y_i = 0$$

DUAL PROBLEM for SVM.

Dual objective

Same solution!

Duality Gap =0.



Dual SVM

- How to calculate b?
- For any point for which $\lambda_i \neq 0$

$\lambda_i > 0$

$$y_i(\bar{a}^T \bar{x}_i + b) = 1$$

complementary slackness

Solving this one can determine b.

Dual SVM

- How to calculate b?
- For any point for which $\lambda_i \neq 0$

$$y_i(\bar{\mathbf{a}}^T \bar{\mathbf{x}}_i + b) - 1 = 0$$

Dual SVM

- Note that the quantity $\bar{\mathbf{x}}_i^T \bar{\mathbf{x}}_j$ denotes the inner product.
- This can be represented as

$$\langle \bar{x}_i , \bar{x}_j \rangle$$

inner product.

Dual SVM

- Note that the quantity $\bar{\mathbf{x}}_i^T \bar{\mathbf{x}}_j$ denotes the inner product.
- This can be represented as

$$\langle \bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j \rangle$$

Dual SVM

- Using this notation, the **dual SVM** problem can be defined as

Depends only on inner products.

inner product

$$\max. \quad \sum_i \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j \langle \bar{x}_i , \bar{x}_j \rangle$$

Subject to.

$$\left.\begin{aligned} &\lambda_i \geq 0 \\ &\sum_i \lambda_i y_i = 0 \end{aligned}\right\} \text{constraints.}$$

Dual SVM

- Using this notation, the **dual SVM** problem can be defined as

$$\max \sum_{i=1}^{2M} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{2M} \lambda_i \lambda_j y_i y_j \langle \bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j \rangle$$

$$\text{subject to } \lambda_i \geq 0$$

$$\sum_{i=1}^{2M} \lambda_i y_i = 0$$

Dual SVM

- For any new point $\bar{\mathbf{x}}$, y can be calculated as

$$\bar{a} = \sum_i \lambda_i y_i \bar{x}_i$$

$$\bar{a}^T \bar{x} + b = \left(\sum_i \lambda_i y_i \bar{x}_i\right)^T \bar{x} + b$$

$$= \sum_i \lambda_i y_i \langle \bar{x}_i, \bar{x} \rangle + b$$

innerproduct.

Dual SVM

- For any new point $\bar{\mathbf{x}}$, y can be calculated as

$$\bar{\mathbf{a}}^T \bar{\mathbf{x}}_i + b = \left(\sum_{i=1}^{2M} \lambda_i y_i \, \bar{\mathbf{x}}_i \right)^T \bar{\mathrm{x}} + b$$

$$= \sum_{i=1}^{2M} \lambda_i y_i \, \langle \bar{\mathbf{x}}_i, \bar{\mathbf{x}} \rangle + b$$

Kernel

Kernel. Kernel SVM.

- One can now replace $\langle \bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j \rangle$ by a kernel

$$K(\bar{x}_i, \bar{x}_j) = \phi(\bar{x}_i)^T \phi(\bar{x}_j)$$

$\bar{x}_i^T \bar{x}_j$: Linear

Feature mapping

Non-Linear

Kernel

- One can now replace $\langle \bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j \rangle$ by a kernel

$$K\left(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j\right) = \phi^T(\bar{\mathbf{x}}_i)\phi(\bar{\mathbf{x}}_i)$$

Feature mapping

- The quantity $\phi(\bar{\mathbf{x}}_i)$ is termed as a **feature mapping**.
- This can be used to model ***non-linear features***.

Kernel SVM

- Using this notation, the **kernel SVM** problem can be defined as

Kernel.

$$\min. \sum_i \lambda_i - \frac{1}{2}\sum_i \sum_j \lambda_i \lambda_j y_i y_j K(\bar{x}_i, \bar{x}_j)$$

s.t. $\lambda_i \geq 0$

$$\sum_i \lambda_i y_i = 0$$

Kernel SVM

- Using this notation, the **kernel SVM** problem can be defined as

Kernel.

$$\max \sum_{i=1}^{2M} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{2M} \lambda_i \lambda_j y_i y_j K(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j)$$

$$\text{subject to } \lambda_i \geq 0$$

$$\sum_{i=1}^{2M} \lambda_i y_i = 0$$

Feature mapping

- For example, for $N = 3$,

3 dimensional Feature vectors.

$$\bar{x}_i = \begin{bmatrix} x_i(1) \\ x_i(2) \\ x_i(3) \end{bmatrix}$$

$$\phi(\bar{x}_i) = \begin{bmatrix} x_i(1)\,x_i(1) \\ x_i(1)\,x_i(2) \\ x_i(1)\,x_i(3) \\ x_i(2)\,x_i(1) \\ x_i(2)\,x_i(2) \\ x_i(2)\,x_i(3) \\ x_i(3)\,x_i(1) \\ x_i(3)\,x_i(2) \\ x_i(3)\,x_i(3) \end{bmatrix}$$

Non Linear Feature map.

Feature mapping

- For example, for $N = 3$,

$$\phi(\bar{\mathbf{x}}_i) = \begin{bmatrix} x_i(1)x_i(1) \\ x_i(1)x_i(2) \\ x_i(1)x_i(3) \\ x_i(2)x_i(1) \\ x_i(2)x_i(2) \\ x_i(2)x_i(3) \\ x_i(3)x_i(1) \\ x_i(3)x_i(2) \\ x_i(3)x_i(3) \end{bmatrix}$$

Feature mapping

- For this kernel,

$$K(\bar{x}_i, \bar{x}_j) = \phi^T(\bar{x}_i)\,\phi(\bar{x}_j)$$

$$= (\bar{x}_i^T \bar{x}_j)^2 \Big\} \text{ Non-linear}$$

Feature mapping

- For this kernel,

$$\underbrace{K(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j)} = \phi^T(\bar{\mathbf{x}}_i)\phi(\bar{\mathbf{x}}_j) = (\bar{\mathbf{x}}_i^T \bar{\mathbf{x}}_j)^2$$

Need to know only Kernel!

Gaussian Kernel

$\cancel{\phi^T(\bar{x}_i)\,\phi(\bar{x}_j)}$

Dont Need!

- Another interesting kernel is the **Gaussian kernel** defined as,

Directly evaluate Kernel!

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

Gaussian Kernel.

Gaussian Kernel

- Another interesting kernel is the **Gaussian kernel** defined as,

$$K(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j) = \exp\left(-\frac{\left\|\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j\right\|^2}{2\sigma^2}\right)$$

Gaussian Kernel

Good Performance!

- Example- **Handwritten digit recognition**, from 16×16 images
- **Gaussian kernel** SVMs yield very good performance!

Handwritten Digits Recognition

7 6

Instructors may use this white area (14.5 cm / 25.4 cm) for the text.
Three options provided below for the font size.

Font: Avenir (Book), Size: 32, Colour: Dark Grey

Font: Avenir (Book), Size: 28, Colour: Dark Grey

Font: Avenir (Book), Size: 24, Colour: Dark Grey

Do not use the space below.