50.007 Machine Learning

Homework 3

Classification with Support Vector Machines

Berrak Sisman

Assistant Professor, ISTD Pillar, SUTD

Graded by TA

Question 1. 1 [20 pts]

Given the mapping

$$\mathrm{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T \mapsto \varphi(\mathrm{x}) = \begin{bmatrix} 1 & x_1^2 & \sqrt{2}x_1x_2 & x_2^2 & \sqrt{2}x_1 & \sqrt{2}x_2 \end{bmatrix}^T$$

(i) Determine the kernel $K(\mathrm{x}, \mathrm{y})$

(ii) Calculate the value of the kernel if $\mathrm{x} = [1\ 2]^T$ and $\mathrm{y} = [3\ 4]^T$

Solution: (i) The kernel defined by this mapping is

$$\begin{aligned} K(\mathrm{x}, \mathrm{y}) &= \varphi^T(\mathrm{x})\,\varphi(\mathrm{y}) \\ &= \begin{bmatrix} 1 & x_1^2 & \sqrt{2}x_1x_2 & x_2^2 & \sqrt{2}x_1 & \sqrt{2}x_2 \end{bmatrix} \begin{bmatrix} 1 \\ y_1^2 \\ \sqrt{2}y_1y_2 \\ y_2^2 \\ \sqrt{2}y_1 \\ \sqrt{2}y_2 \end{bmatrix} \\ &= 1 + x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2 + 2x_1y_1 + 2x_2y_2 \\ &= \left(1 + \mathrm{x}^T\mathrm{y}\right)^2 \end{aligned}$$

[10 pts]

(ii) With $\mathrm{x} = [1\ 2]^T$ and $\mathrm{y} = [3\ 4]^T$, the value of the kernel is

$$\begin{aligned} K(\mathrm{x}, \mathrm{y}) &= 1 + x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2 + 2x_1y_1 + 2x_2y_2 \\ &= 1 + 1^2 \cdot 3^2 + 2 \cdot 1 \cdot 2 \cdot 3 \cdot 4 + 2^2 \cdot 4^2 + 2 \cdot 1 \cdot 3 + 2 \cdot 2 \cdot 4 \\ &= 1 + 9 + 48 + 64 + 6 + 16 = 144 \end{aligned}$$

[10 pts]

Question 1. 2 [30 pts]

Question 1.2 [30 pts]
The primal problem of SVM with soft margin is given below:

$$\text{minimize} \quad \frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i$$

$$\text{subject to} \quad d_i(w^T x_i + b) - 1 + \xi_i \geq 0\,, \quad \xi_i \geq 0$$

1) Using Lagrange multipliers and KKT conditions, can you derive the formulation of dual problem with soft margin? Please note that the dual form is already provided in slides, so we expect you to go through the mathematical steps. [*20pts*]
2) Explain in which cases we would prefer to use soft margin rather than hard margin. [*10pts*]

Question 1. 2 [30 pts]

1) Using Lagrange multipliers and KKT conditions, can you derive the formulation of dual problem with soft margin? Please note that the dual form is already provided in slides, so we expect you to go through the mathematical steps. [*20pts*]

Solution

Let α_i and β_i be the Lagrange multipliers. Then

$$
\begin{aligned}
& L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
= \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left(d_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1 + \xi_i\right) - \sum_{i=1}^{N}\beta_i\xi_i \\
= \quad & \frac{\mathbf{w}^T\mathbf{w}}{2} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i d_i \mathbf{w}^T\mathbf{x}_i - b\sum_{i=1}^{N}\alpha_i d_i + \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}\alpha_i\xi_i - \sum_{i=1}^{N}\beta_i\xi_i
\end{aligned}
$$

The KKT conditions are

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N}\alpha_i\, d_i\, \mathbf{x}_i = 0$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N}\alpha_i\, d_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0$$

$$d_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1 + \xi_i \geq 0$$

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

$$\beta_i\xi_i = 0$$

$$\alpha_i \geq 0$$

$$\beta_i \geq 0$$

[10 pts] for writing all KKT conditions;

Very important!

Next slide

Question 1. 2 [30 pts]

[10 pts] for solving the KKT conditions and obtaining the dual form;

Solution

Since $\mathbf{w} = \sum_{i=1}^{N} \alpha_i \, d_i \, \mathbf{x}_i$. We have

$$\mathbf{w}^T\mathbf{w} = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \, d_i \, \alpha_j \, d_j \, \mathbf{x}_i^T \mathbf{x}_j$$

$$\sum_{i=1}^{N} \alpha_i \, d_i \, \mathbf{w}^T\mathbf{x}_i = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \, d_i \, \alpha_j \, d_j \, \mathbf{x}_i^T \mathbf{x}_j$$

Moreover, from the KKT conditions we also have

$$C = \alpha_i + \beta_i$$

This sets an upper bound for α_i!
Remember that $\beta_i \geq 0$, and $\alpha_i = C - \beta_i$
$\mathbf{0} \leq \boldsymbol{\alpha_i} \leq \mathbf{C}$

Hence,

$$\begin{aligned}
& L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
= & \frac{\mathbf{w}^T\mathbf{w}}{2} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i d_i \mathbf{w}^T\mathbf{x}_i - b\sum_{i=1}^{N}\alpha_i d_i + \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}\alpha_i\xi_i - \sum_{i=1}^{N}\beta_i\xi_i \\
= & -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i d_i \alpha_j \, d_j \mathbf{x}_i^T\mathbf{x}_j + \sum_{i=1}^{N}\underbrace{(\alpha_i + \beta_i)}_{C}\xi_i + \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}\alpha_i\xi_i - \sum_{i=1}^{N}\beta_i\xi_i \\
= & -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i \, d_i \, \alpha_j \, d_j \, \mathbf{x}_i^T\mathbf{x}_j + \sum_{i=1}^{N}\alpha_i
\end{aligned}$$

Next slide

Question 1. 2 [30 pts]

Solution

Dual problem (with soft margin)

Find : α_i

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

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

Please note that $Q(\alpha)$ is same as that for the dual problem without soft margin.

If the previous 2 pages are correct, then the student gets 20 points.

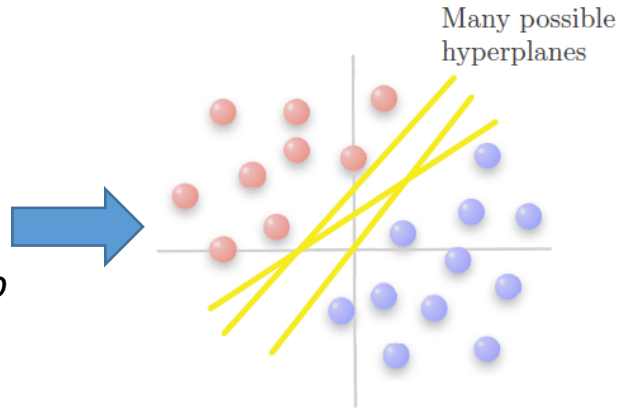
Question 1. 2 [30 pts]

2) Explain in which cases we would prefer to use soft margin rather than hard margin. [*10pts*]

Solution

If our data is linearly separable, we can use hard margin. As discussed in lectures, hard margin is successful to handle such data. An example from our lecture notes is given below:

linearly separable
(Students do not need to include this)



[10 pts] students need to mention "not linearly separable" case

If our data is **not linearly separable**, we can use soft margin to allow a few points to be on the wrong side.

Question 1.3: Hands-on [50 pts]

Answer:

Kernel 0 (linear) accuracy = 79.3651% (50/63) (classification)

Kernel 1 (polynomial) accuracy = 55.5556% (35/63) (classification)

Kernel 2 (RBF) accuracy = 87.3016% (55/63) (classification)

Kernel 3 (Sigmoid) accuracy = 82.5397% (52/63) (classification)

[40 pts] for correct accuracy

Question 1.3: Hands-on [50 pts]

Answer:

Kernel 0 (linear) accuracy = 79.3651% (50/63) (classification)

Kernel 1 (polynomial) accuracy = 55.5556% (35/63) (classification)

Kernel 2 (RBF) accuracy = 87.3016% (55/63) (classification)

Kernel 3 (Sigmoid) accuracy = 82.5397% (52/63) (classification)

[10 pts] for picking RBF and stating that it is the best.

Best performance!