01.112/50.007 Machine Learning

Homework 2

Clustering and SVM

Berrak Sisman

Assistant Professor, ISTD Pillar, SUTD

Graded by Zongyang Du

Clustering

Question 1. 1 [10 pts]

Please indicate whether the following statements are true (T) or false (F)

a) The goal of unsupervised learning is to uncover useful structure in the labeled data such as classifying cars or predicting house prices. **False (2pts)**

b) In clustering, the choice of which distance metric to use is important as it will determine the type of clusters you will find. **True (2pts)**

c) K-Means clustering algorithm is not sensitive to outliers as it uses the mean of cluster data points to find the cluster center. **False (2pts)**

d) Each iteration of the k-means and k-medoids algorithms lower the cost. Therefore, they always converge to the optimal (global) solution. **False (2pts)**

e) The quality of the clustering of k-medoids does not depend on the initialization. **False (2pts)**

Question 1. 2 [30 pts]

Question 1.2 [30 pts]
Suppose that you want to perform k-means clustering on the data given below:

$$S = \{10, 11, 2, 4, 12, 3\}$$

and the initial clustering sets are given as follows:

$$C_1 = \{3, 4, 11, 12\}$$
$$C_2 = \{2, 10\}$$

where C_1 and C_2 represent cluster 1 and cluster 2.
a) Compute the centroid of C_1, which is denoted as μ_1.
b) Compute the centroid of C_2, which is denoted as μ_2.
c) Compute the new clusters formed by the centroids μ_1 and μ_2. Label the cluster associated with μ_1 as D_1, and the cluster associated with μ_2 as D_2.
d) Calculate the new centroids of D_1 and D_2.
e) Is this clustering stable, in other words, did the k-means algorithm converge? Please explain your answer.

Question 1. 2 [30 pts]

a) Compute the centroid of C_1, which is denoted as μ_1.

$$\mu_1 = \frac{3+4+11+12}{4} = 7.5$$ **(2.5 pts)**

b) Compute the centroid of C_2, which is denoted as μ_2.

$$\mu_2 = \frac{2+10}{2} = 6$$ **(2.5 pts)**

Next slide

Question 1. 2 [30 pts]

c) Compute the new clusters formed by the centroids μ_1 and μ_2. Label the cluster associated with μ_1 as D_1, and the cluster associated with μ_2 as D_2.

Datapoint	Distance1	Distance2	New Cluster
10	2.5	4	D_1
11	3.5	5	D_1
2	5.5	4	D_2
4	3.5	2	D_2
12	4.5	6	D_1
3	4.5	3	D_2

Distance1: the distance between each point and μ_1.
Distance2: the distance between each point and μ_2.

$D_1 = 10,11,12$ **(5 pts)**

$D_2 = 2,3,4$ **(5 pts)**

Next slide

Question 1. 2 [30 pts]

d) Calculate the new centroids of D_1 and D_2.

$$\mu_1' = \frac{10+11+12}{3} = 11 \quad \textbf{(2.5 pts)}$$

$$\mu_2' = \frac{2+3+4}{3} = 3 \quad \textbf{(2.5 pts)}$$

Next slide

Question 1. 2 [30 pts]

e) Is this clustering stable, in other words, did the k-means algorithm converge? Please explain your answer.

Datapoint	Distance1	Distance2	Cluster
10	1	7	D_1
11	0	8	D_1
2	9	1	D_2
4	7	1	D_2
12	1	9	D_1
3	8	0	D_2

Answer: Yes, it is stable. **(5 pts)**

Explanation: As shown in the table, we compute the new clusters formed by the centroids $\mu_1^{'}$ and $\mu_2^{'}$. And we find the clustering with respect to $\mu_1^{'}$ and $\mu_2^{'}$ remain the same. Therefore, we can say this clustering is stable and the k-means algorithm converge. **(5 pts)**

Support Vector Machines

Question 2. 1 [10 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

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

[5 pts]

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

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

[5 pts]

Question 2. 2 [15 pts]

The primal problem of SVM with soft margin is given below:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \\ \text{subject to} \quad & d_i(w^T x_i + b) - 1 + \xi_i \geq 0 \,, \quad \xi_i \geq 0 \end{aligned}$$

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. [*10pts*]

2) Explain in which cases we would prefer to use soft margin rather than hard margin. [*5pts*]

Question 2. 2 [15 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. [*10pts*]

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

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

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

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

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

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

$$\beta_i\xi_i = 0$$

$$\alpha_i \geq 0$$

$$\beta_i \geq 0$$

[5 pts] for writing all KKT conditions;

Very important!

Next slide

Question 2. 2 [15 pts]

[5 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 2. 2 [15 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 10 points.

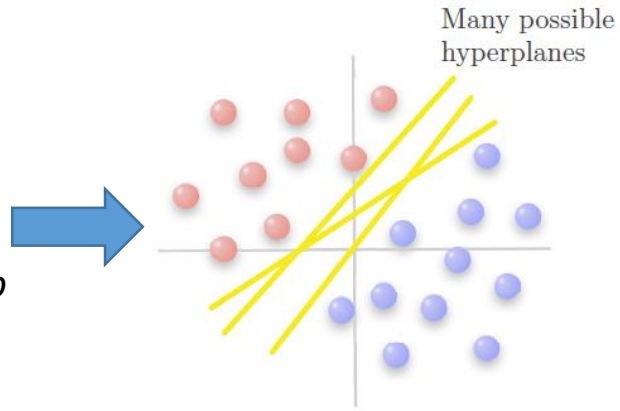
Question 2. 2 [15 pts]

2) Explain in which cases we would prefer to use soft margin rather than hard margin. [*5pts*]

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)



[5 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 2.3: Hands-on [35 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)

[28 pts] for correct accuracy

Question 2.3: Hands-on [35 pts]

Answer:

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

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) **Best performance!**

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