Assignment 3

Benny Chen

November 30, 2023

1 Kernels for SVM

1.1

In the lecture of Support Vector Machine, we have discussed the kernel function, which can map the data points into a new high-dimensional space.
Let $k_1(x, y)$ and $k_2(x, y)$ be the *valid kernels*, that is, each of these two kernel functions can be decomposed into the product of two feature mapping functions $\Phi(x), i.e, k_1(x, y) = \Phi_1(x)^T \Phi_1(y)$ and $k_2(x, y) = \Phi_2(x)^T \Phi_2(y)$.
Please show that $k(x, y) = k_1(x, y) + k_2(x, y)$ is also a valid kernel function by decomposing it into product of two feature mapping functions.

Answer:

$$
\begin{aligned}
k(x,y) &= k_1(x,y) + k_2(x,y) \\
&= \Phi_1(x)^T \Phi_1(y) + \Phi_2(x)^T \Phi_2(y) \\
&= \begin{bmatrix} \Phi_1(x)^T & \Phi_2(x)^T \end{bmatrix} \begin{bmatrix} \Phi_1(y) \\ \Phi_2(y) \end{bmatrix} \\
&= \begin{bmatrix} \Phi_1(x)^T & \Phi_2(x)^T \end{bmatrix} \begin{bmatrix} \Phi_1(y) \\ 0 \end{bmatrix} + \begin{bmatrix} \Phi_1(x)^T & \Phi_2(x)^T \end{bmatrix} \begin{bmatrix} 0 \\ \Phi_2(y) \end{bmatrix} \\
&= \begin{bmatrix} \Phi_1(x)^T & 0 \end{bmatrix} \begin{bmatrix} \Phi_1(y) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & \Phi_2(x)^T \end{bmatrix} \begin{bmatrix} 0 \\ \Phi_2(y) \end{bmatrix} \\
&= \begin{bmatrix} \Phi_1(x)^T & 0 \end{bmatrix} \begin{bmatrix} \Phi_1(y) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & \Phi_2(x)^T \end{bmatrix} \begin{bmatrix} 0 \\ \Phi_2(y) \end{bmatrix} \\
&= \begin{bmatrix} \Phi_1(x)^T & \Phi_2(x)^T \end{bmatrix} \begin{bmatrix} \Phi_1(y) \\ \Phi_2(y) \end{bmatrix} \\
&= \Phi(x)^T \Phi(y)
\end{aligned}
$$

The product of two feature mapping functions is $\Phi(x)^T \Phi(y)$.

2 K-Means and Cluster Validation

2.1

Consider the following set of one-dimensional points:

$$\{0.1, 0.2, 0.45, 0.55, 0.8, 0.9\}$$

All the points are located in the range between [0,1]. Suppose we apply k-means clustering to obtain three clusters, A, B, and C. If the initial centroids are located at 0, 0.4, 1, respectively, please find:

(a) The cluster assignments (fill in either A, B, or C for each data point);

(b) Locations of the centroids (coordinate) after the first three iterations by filling out the following table.

Answer:

Iter	0.10	0.25	0.45	0.55	0.80	0.90	A	B	C
0	—	—	—	—	—	—	0.00	0.40	1.00
1	A	A	B	B	C	C	0.15	0.50	0.85
2	A	A	B	B	C	C	0.15	0.50	0.85
3	A	A	B	B	C	C	0.15	0.50	0.85

2.2

The following table shows the clustering results in a land cover classification dataset that consists of many pieces of land. The number provided in the table is the number of objects (pieces of land) that are clustered into each cluster that belongs to each category. For example, the number in the forest column and cluster 1 row means that 10 forest items are clustered into cluster 1.

	Forest	Farm	Shrubland	Urban	Water
Cluster 1	20	10	10	10	950
Cluster 2	400	100	400	50	50
Cluster 3	50	50	500	200	200
Cluster 4	200	250	150	200	200

Which cluster has the smallest entropy? Which cluster has the largest entropy?

Answer:

(a) Cluster 1: $20 + 10 + 10 + 10 + 950 = 1000$
Entropy:

$$-\frac{20}{1000}\log_2(\frac{20}{1000}) - 3(\frac{10}{1000}\log_2(\frac{10}{1000})) - \frac{950}{1000}\log_2(\frac{950}{1000}) = 0.38249$$

(b) Cluster 2: $400 + 100 + 400 + 50 + 50 = 1000$
Entropy:

$$-2(\frac{400}{1000}\log_2(\frac{400}{1000})) - 2(\frac{50}{1000}\log_2(\frac{50}{1000})) - \frac{100}{1000}\log_2(\frac{100}{1000}) = 1.82193$$

(c) Cluster 3: $50 + 50 + 500 + 200 + 200 = 1000$
Entropy:

$$-2(\frac{50}{1000}\log_2(\frac{50}{1000})) - \frac{500}{1000}\log_2(\frac{500}{1000}) - 2(\frac{200}{1000}\log_2(\frac{200}{1000})) = 1.52193$$

(d) Cluster 4: $200 + 250 + 150 + 200 + 200 = 1000$
Entropy:

$$-3(\frac{200}{1000}\log_2(\frac{200}{1000})) - \frac{250}{1000}\log_2(\frac{250}{1000}) - \frac{150}{1000}\log_2(\frac{150}{1000}) = 2.30370$$

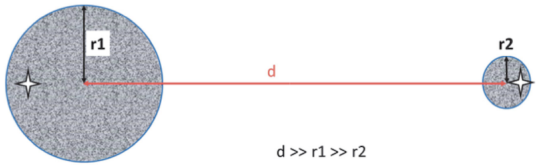
Cluster 1 has the smallest entropy, and Cluster 4 has the largest entropy.

2.3

Please provide True or False, followed by the Short Answer on your explanations.

(a) When clustering a dataset using K-means, whenever SSE decreases, cohesion increases.

(b) With the given initial centroid in the following diagram: When the k-means algorithm completes, each shaded circle will have one cluster centroid at its center.



(c) With the given initial centroids: When the k-means algorithms completes, there will be one cluster centroid in the center of each of the two shaded regions, and each of the two final clusters will consist only of points from one of the shaded regions. In other words, none of the two final clusters will have points from both shaded regions.

Answer:

(a) True.

(b) True.

(c) False.

2.4

The distance matrices (Euclidean distance) below (**Figures A, B, and C**) are sorted according to cluster labels, and correspond, in some order, to the sets of points (Figures D, E, and F).
Differences in color distinguish between clusters, and each set of points contains 100 points and four clusters, each of equal size. In the distance matrix, blue indicates the smallest distances, and red indicates the largest distances.

(a) Match the distance matrices (**Figures A, B, and C**) with the right sets of points (**Figures D, E, and F**). Please note that the colors of the data points in Figures D-F DO NOT match the colors in the distance matrices in Figures A-C.

(b) For the symmetric matrix given in **Figure B**, match the four rows to the corresponding clusters (characterized by four colors — red, blue, green, and magenta) in the dataset that you match with it in the previous question.

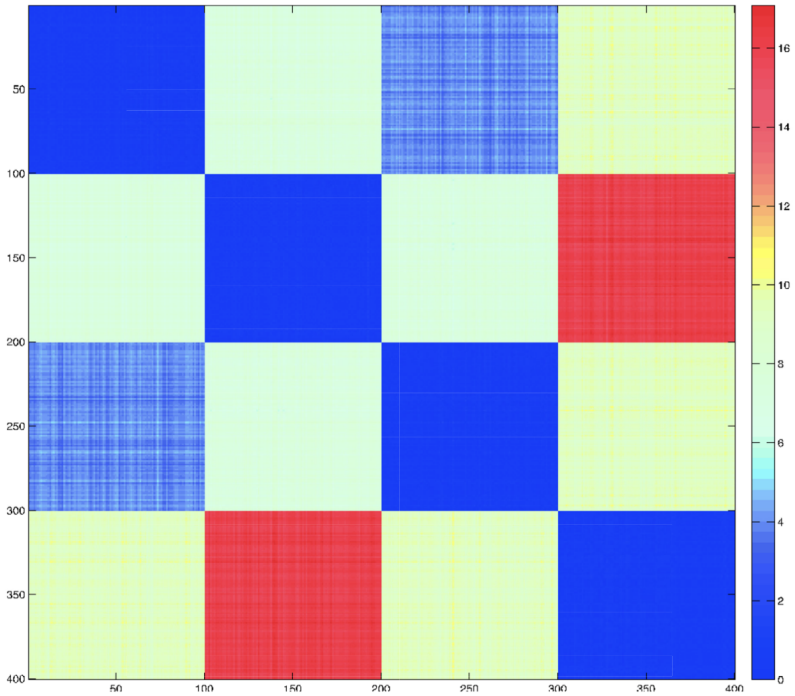
Answer:

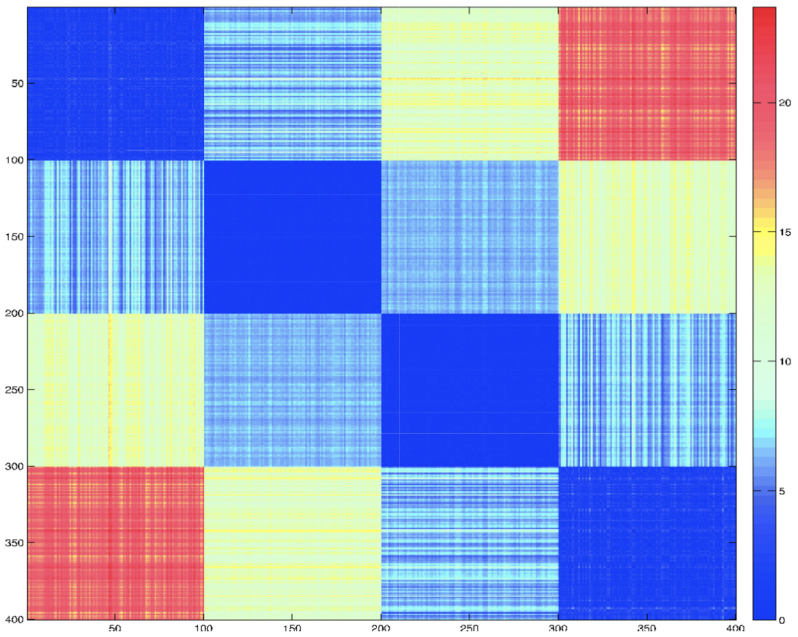
Matrices	Datasets
1(a)	F
1(b)	D
1(c)	E

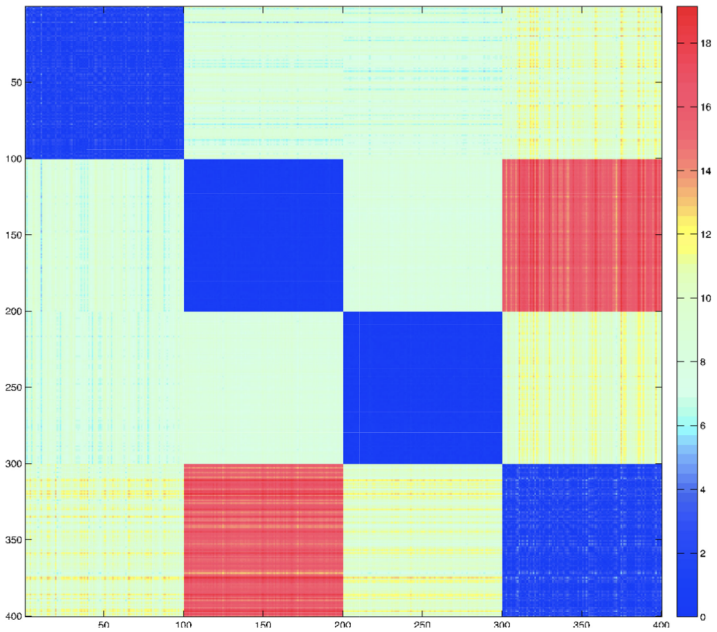
(a)

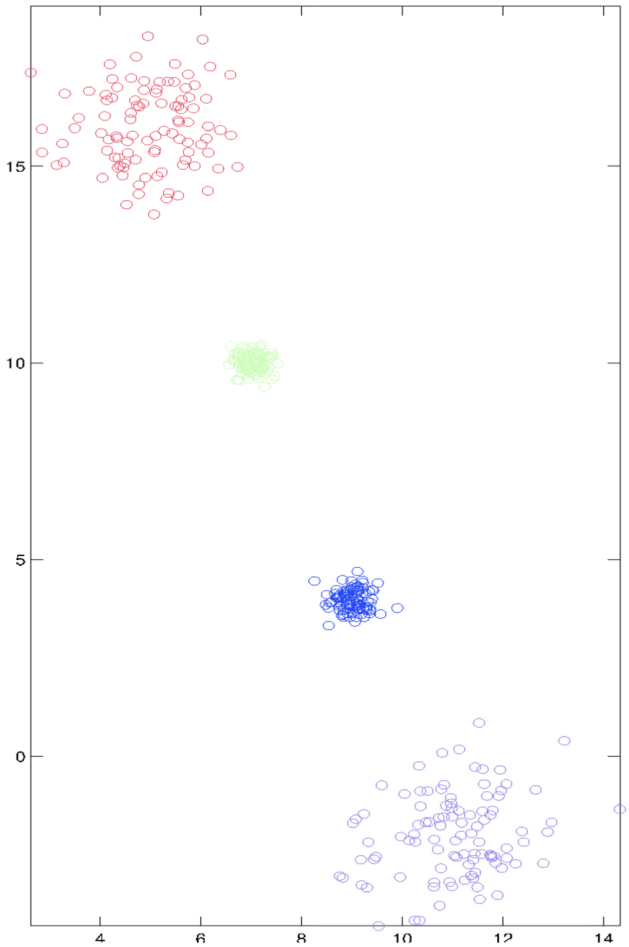
Matrix Rows	Cluster (Represented by Colors)
1st (1–100)	Red
2nd (101–200)	Green
3rd (201–300)	Blue
4th (301–400)	Magenta

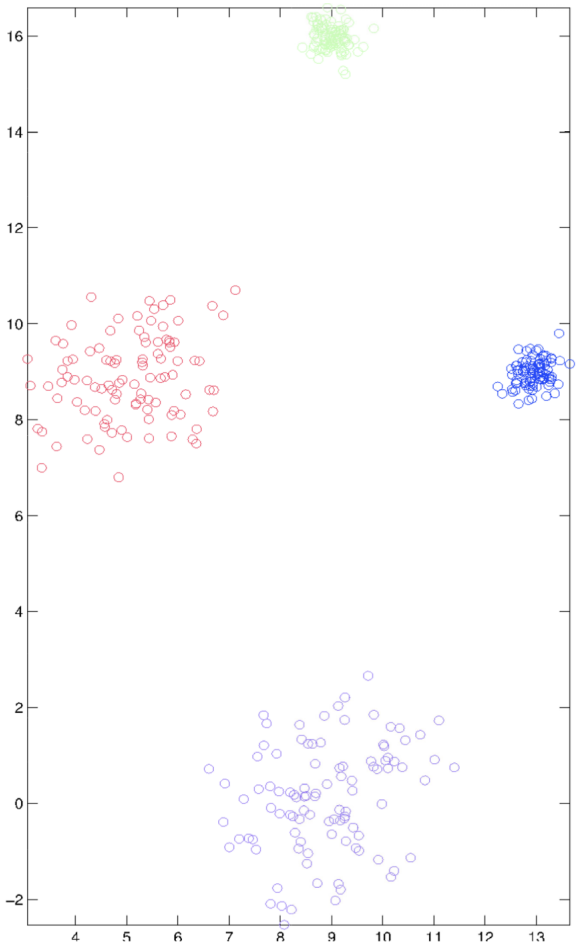
(b)

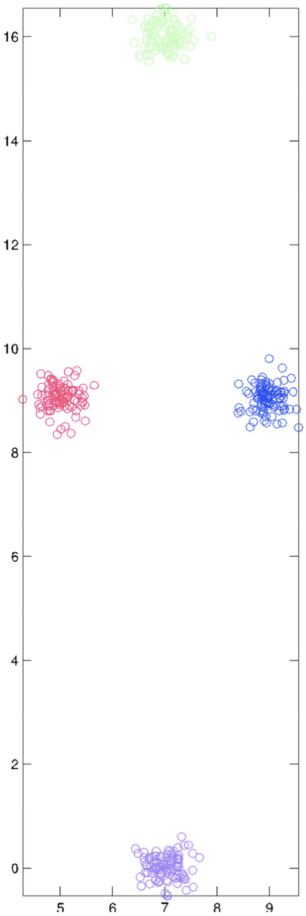



50
100
150
200
250
300
350
400
50
100
150
200
250
300
350
400
0
5
10
15
20


50
100
150
200
250
300
350
400
50
100
150
200
250
300
350
400
0
2
4
6
8
10
12
14
16
18


15
10
5
0
4
6
8
10
12
14

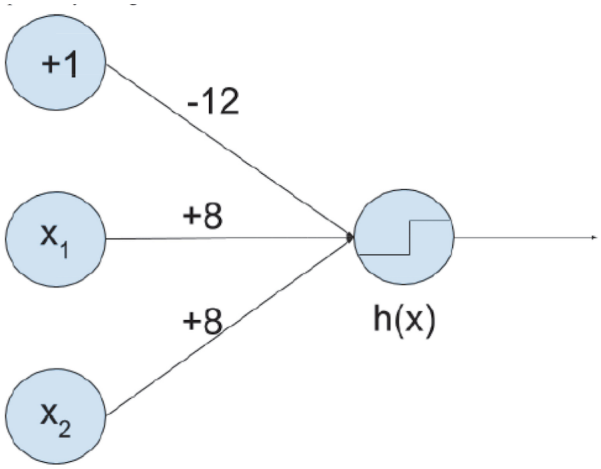

16
14
12
10
8
6
4
2
0
−2
4
5
6
7
8
9
10
11
12
13



3 Neural Network

3.1

You are given the following neural networks which take two binary valued inputs $x_1, x_2 \in \{0, 1\}$ and the activation function is the threshold function ($h(x) = 1$ if $x > 0$; 0 otherwise). Which of the following logical functions does it compute, OR, AND, NAND? Please provide your argument.



Answer:

This is a AND function due to x_1 and x_2 both being 1 or true would result in the output being greater than 0. Vice versa, if either x_1 or x_2 is 0 or false, the output would be less than 0 and the result of the network would be 0.

3.2

We have a function which takes a two-dimensional input $x = (x_1, x_2)$ and has two parameters $w = (w_1, w_2)$ given by

$$f(x; w) = \sigma(\sigma(x_1 w_1) \cdot w_2 + x_2) \tag{1}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$. We use backpropagation to estimate the right parameter values. We start by setting both the parameters to be 0. Assume that we are given a training point $x_1 = 2, x_2 = 0, y = 5$. Given above information, please find the value of $\frac{\partial f}{\partial w_2}$ based on the chain rule.

Answer:

$$\frac{\partial \sigma(x)}{\partial x} = \sigma(x)(1 - \sigma(x))$$

$$\frac{\partial(\sigma(x_1 w_1) \cdot w_2 + x_2)}{\partial w_2} = \sigma(x_1 w_1)$$

$$\frac{\partial f}{\partial w_2} = (\sigma(\sigma(x_1 w_1) \cdot w_2 + x_2)(1 - \sigma(\sigma(x_1 w_1)) \cdot w_2 + x_2)) \cdot \sigma(x_1 w_1)$$

$x = (2, 0)$, $w = (0, 0)$

$$\frac{\partial f}{\partial w_2} = (\sigma(\sigma(2 \cdot 0) \cdot 0 + 0)(1 - \sigma(\sigma(2 \cdot 0)) \cdot 0 + 0)) \cdot \sigma(2 \cdot 0) = 0.25$$