Exam 2

Benny Chen

November 17, 2023

1 True or False plus explanation

1. In the soft-margin linear support vector machine classification problem (the one with the slack variables), it is not possible to control the relative importance of maximizing the margin and minimizing the error.

2. In soft-margin SVM, suppose slack variable $\epsilon_i = 0.2$ for some i. Then the corresponding datapoint x_i for this slack variable is a support vector since it is classified incorrectly.

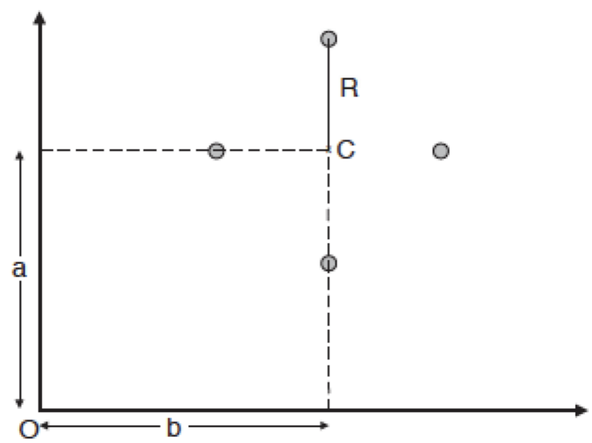
Solutions

1. False. The relative importance of maximizing the margin and minimizing the error can be controlled by C. A large C will result in a small margin and a small C will result in a large margin.

2. False. The datapoint x_i is not a support vector since it is classified correctly. This is because the slack variable ϵ_i is greater than 0 and less than 1. If ϵ_i was equal to 1, then x_i would be a support vector.

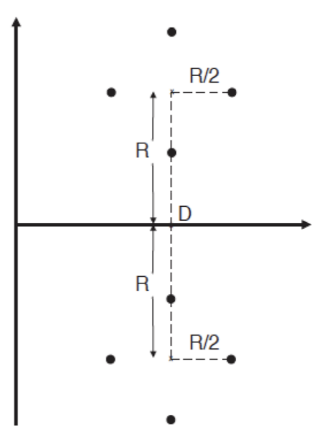
2 K-Means Clustering

Consider the 4 data points shown in the following figure. The distance between each data point to the centroid C is R.



Example of 4 data points in 2-dimensional space.

1. Compute the total SSE of the data points to the centroid, C.
2. Compute the total SSE of the data points to the origin, O.
3. Using parts (a) and (b), compute the SSE for the 8 data points shown below with respect to the centroid, D. Note that points each group (cluster) of data points lie on a circle of radius R/2. Also, the figure is symmetric with respect to the horizontal line running through D.



Example of 8 data points in 2-dimensional space.

Solutions

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x) \tag{1}$$

1.
$$SSE = R^2 + R^2 + R^2 + R^2 = 4R^2 \tag{2}$$

2.
$$dist^2(O, R_1) = (b - R)^2 + (a)^2 \tag{3}$$
$$dist^2(O, R_2) = (b)^2 + (a - R)^2 \tag{4}$$
$$dist^2(O, R_3) = (b + R)^2 + (a)^2 \tag{5}$$
$$dist^2(O, R_4) = (b)^2 + (a + R)^2 \tag{6}$$
$$SSE = (b - R)^2 + (a)^2 + (b)^2 + (a - R)^2 + (b + R)^2 + (a)^2 + (b)^2 + (a + R)^2 \tag{7}$$
$$SSE = 4a^2 + 4b^2 + 4R^2 \tag{8}$$

3.
$$dist^2(D, R_1) = (\frac{R}{2})^2 \tag{9}$$
$$dist^2(D, R_2) = (\frac{3R}{2})^2 \tag{10}$$
$$dist^2(D, R_3) = (R)^2 + (\frac{R}{2})^2 \tag{11}$$
$$dist^2(D, R_4) = (R)^2 + (\frac{R}{2})^2 \tag{12}$$
$$SSE = (\frac{R}{2})^2 + (\frac{3R}{2})^2 + (R)^2 + (\frac{R}{2})^2 + (R)^2 + (\frac{R}{2})^2 \tag{13}$$
$$SSE = 5R^2 \tag{14}$$

Since the figure is symmetric with respect to the horizontal line running through D, the SSE for the 8 data points is $5R^2 + 5R^2 = 10R^2$.