Assignment 2:

Danil Shalagin
B19-DS-01
d.shalagin@innopolis.university

Innopolis University

1 Theoretical question on K-means Clustering

Initially we have two datasets: $D_1 = \{-2, ..., -2\}$ $D_2 = \{0, ..., 0, a\}$, where D_1 has m -2 values and D_2 has m 0 and one a.

Let us consider other dataset, which may have lower sum-of-squared error. Given two sets has sum-of-squared error J. $D_1 = \{-2, ..., -2, 0, ..., 0\} D_2 = \{a\}$, where D_1 has one a values and D_2 has m -2 and m 0. This dataset will have low or 0 sum-of-squared error for dataset D_2 and low sum-of-squared error for D_1. This datasets has sum-of-squared error J‘.

Now J and J‘ will be computed:

First calculate J.

Calculate μ_1 and μ_2:

$\mu_1 = \sum_{x \in D_1}(x/m) = -2m/m = -2$

$\mu_2 = \sum_{x \in D_2}(x/m) = a/(m+1) + 0/(m+1) = a/(m+1)$

Now calculate J: $J = \sum\sum_{i:D_i \neq \emptyset, x \in D_i}(x - \mu_i)^2 = 0 + (0-a)^2 * m/(m+1)^2 + (a - a/(m+1))^2 = a^2m/(m+1)$

Secondly calculate J‘.

Calculate μ_1 and μ_2:

$\mu_1 = \sum_{x \in D_1}(x/m) = (-2m + 0m)/2m = -1$

$\mu_2 = \sum_{x \in D_2}(x/m) = a/1 = a$

Now calculate J‘: $J‘ = \sum\sum_{i:D_i \neq \emptyset, x \in D_i}(x - \mu_i)^2 = (-2+1)^2 * m + (0-1)^2 * m = 2m$

Thus, our initial dataset will have minimum sum-of-squared error J<J‘. So $a^2m/(m+1) < 2m$, thus $a^2 < 2(m+1)$

2 Theoretical question on SVM.

(I):

a: No Given margin doesn't have support vector from negative dataset. According to the definition, margin is computed based on perpendicular distances from both sides of the decision boundary. Thus, given margin cannot be produced by SVM described by formula (I).

b: No Formula (I) has condition: $y_t\theta^T x_t > 1$ which means that $\theta_0 = 0$. Therefore, margin goes through the origin. As we see on the plot b, margin b doesn't go through the origin. So plot be cannot be generated by formula (I).

c: No There are two points from "+" dataset which doesn't satisfy condition $y_t\theta^T x_t \geq 1$. Those points has distance less then 1. Thus, margin c cannot be constructed.

(II):

a: No Given margin doesn't have support vector from negative dataset. According to the definition, margin is computed based on perpendicular distances from both sides of the decision boundary. Thus, given margin cannot be produced by SVM described by formula (II).

b: Yes

c: No There are two points from "+" dataset which doesn't satisfy condition $y_t(\theta^T x_t + \theta_0) \geq 1$. Those points has distance less then 1. Thus, margin c cannot be constructed.

(III):

a: No Given margin doesn't have support vector from negative dataset. According to the definition, margin is computed based on perpendicular distances from both sides of the decision boundary. Thus, given margin cannot be produced by SVM described by formula (I).

b: No Formula (III) has condition $y_t\theta^T x_t \geq 1 - \epsilon_t$ which means that constant $\theta_0 = 0$, thus margins that are created by formula (III) goes through the origin. Margin on plot b doesn't satisfy this condition. Thus, b is not created by formula (III)

c: Yes