SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

50.007 Machine Learning, Fall 2021

Homework 2: Clustering and Support Vector Machines

Due: 20 October 2021
This homework will be graded by Zongyang Du (TA).

1 Clustering: k-means and k-medoids

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.
b) In clustering, the choice of which distance metric to use is important as it will determine the type of clusters you will find.
c) K-means clustering algorithm is not sensitive to outliers as it uses the mean of cluster data points to find the cluster center.
d) Each iteration of the k-means and k-medoids algorithms lower the cost. Therefore, they always converge to the optimal (global) solution.
e) The quality of the clustering of k-medoids does not depend on the initialization.

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.

Point	3	4	11	12	2	10
C_1	4.5	3.5	3.5	4.5	4.5	2.5
C_2	3	2	5	6	4	4

2 Support Vector Machines

Question 2.1 [10 pts]
Given the mapping $x = [x_1 \;\; x_2]^T \rightarrow \varphi(x) = [1 \;\; x_1^2 \;\; \sqrt{2}x_1x_2 \;\; x_2^2 \;\; \sqrt{2}x_1 \;\; \sqrt{2}x_2]^T$
a) Determine the Kernel $K(x, y)$ [*5pts*]
b) Calculate the value of the Kernel $K(x, y)$ if $x = [1 \;\; 2]^T$ and $y = [3 \;\; 4]^T$ [*5pts*]

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 2.3: Hands-on [35 pts]
Download and install the widely used SVM implementation LIBSVM `https://github.com/cjlin1/libsvm`, or `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`. We expect you to install the package on your own – this is part of learning how to use off-the-shelf machine learning software. Read the documentation to understand how to use it.

Download fishorrock folder. In that folder there are training.txt and test.txt, which respectively contain 145 training examples and 63 test examples in LIBSVM format. Sonar systems are generally used underwater for range finding, classification and detection. In the fishing industry, a sonar is generally used to detect (or classify) fish, and the seafloor around the vessel. The task in this homework is to train an SVM model to discriminate between signals, and to classify if the object is a fish or rock given 60 attributes about the signal. Please note that this is a binary classification task.

Run LIBSVM to classify objects with the following kernels:

- linear
- polynomial
- radial basis function
- sigmoid

You should use default values for all other parameters. Please report the test accuracy for each kernel. Which kernel would you choose and why?

Question 1

Question 1.1

(a) False. Unsupervised learning does not have labeled data.

(b) True. The euclidean distance is one example. The further the euclidean distance of a data point is from one centroid as compared to the other centroid, the less likely the data point belongs to that cluster.

(c) False. As the average of all the points taken into consideration in the computation of the centroid, the outliers are thus considered.

(d) False. K-means converge to the local minima which does not mean global minima all the time.

(e) False. It depends on the decision of the number of clusters which affects the random assignment of the centroids. K-means can converge to different values depending on the random initialisation.

1.2 (a) $\mu_1 = \frac{3+4+11+12}{4} = \frac{30}{4} = 7.5$

(b) $\mu_2 = \frac{2+10}{2} = 6$

(c)

Point	3	4	11	12	2	10
C_1	4.5	3.5	3.5	4.5	4.5	2.5
C_2	3	2	5	6	4	4

$D_1 = \{10, 11, 12\}$

$D_2 = \{2, 3, 4\}$

(d) let the new centroids of D_1 and D_2 be μ_a and μ_b respectively.

$$\mu_a = \frac{10+11+12}{3} = 11$$

$$\mu_b = \frac{2+3+4}{3} = 3.$$

(e) Yes, the K-means algorithm converge.

	10	11	12	2	4	3
E_1	1	0	1	9	7	8
E_2	7	8	9	1	1	0

There is no change in the cluster membership.

2.1(a) $K(x,y) = \varphi^T(x)\,\varphi(y)$

$$= \begin{bmatrix} 1 & x_1^2 & \sqrt{2}x_1x_2 & x_2^2 & \sqrt{2}x_1 & \sqrt{2}x_2 \end{bmatrix} \begin{bmatrix} 1 \\ y_1^2 \\ \sqrt{2}y_1y_2 \\ y_2^2 \\ \sqrt{2}y_1 \\ \sqrt{2}y_2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 + x_1^2 y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2 + 2x_1y_1 + \\ 2x_2y_2 \end{bmatrix}$$

2.1(b) $x = [1\ 2]^T$; $y = [3\ 4]^T$

(x_1, x_2 and y_1, y_2)

$$K(x,y) = [1 + 9 + 48 + 64 + 6 + 16]$$

$$= [144]$$

2.2(i) By lagrange multipliers,

$$L(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{m} \beta_i h_i(w)$$

(α_i: Lagrange mult; β_i: Lagrange mult)

Let's modify the constraint! (next page)

$$d_i(w^T x_i) + b - 1 + \xi_i \geq 0$$

$$d_i(w^T x_i) + b + \xi_i \geq 1$$

$$\Rightarrow 1 - (d_i(w^T x_i) + b + \xi_i) \leq 0$$

Lagrangian function: $q_i(w) \leq 0$; $h_i(w) = 0$

$$L(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i q_i(w) + \sum_{i=1}^{m} \beta_i h_i(w)$$

$$L(w, \beta) = f(w) + \sum_{i=1}^{m} \beta_i h_i(w)$$

$$L(w, b, \alpha, \xi) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i (d_i(w^T x_i + b) - 1 + \xi_i) + \sum_{i=1}^{N} \beta_i (-\xi_i)$$

$$= \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i d_i w^T x_i - b \sum_{i=1}^{N} \alpha_i d_i + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i \xi_i - \sum_{i=1}^{N} \beta_i \xi_i$$

For the KKT conditions,

$$\frac{\partial L(w,b,\alpha,\xi)}{\partial w} = 0 = w - \sum_{i=1}^{N} \alpha_i d_i x_i \Rightarrow w = \sum_{i=1}^{N} \alpha_i d_i x_i$$

$$\frac{\partial L(w,b,\alpha,\xi)}{\partial b} = 0 = \sum_{i=1}^{N} \alpha_i d_i$$

$$\frac{\partial L(w,b,\xi,\alpha)}{\partial \xi} = 0 = C - \alpha_i - \beta_i \Rightarrow C = \alpha_i + \beta_i$$

$$\frac{1}{2} w^T w = \frac{1}{2}\left[\sum_{i=1}^{N} \alpha_i d_i x_i^T \sum_{i=1}^{N} \alpha_i d_i x_i \right]$$

$$L(w,b,\xi,\alpha) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i d_i w^T x_i - b \sum_{i=1}^{N} \alpha_i d_i + \sum_{i=1}^{N} \alpha_i - \left(\sum_{i=1}^{N} \alpha_i \xi_i + \sum_{i=1}^{N} \beta_i \xi_i \right)$$

$$= \frac{1}{2}\left[\sum_{i=1}^{N} \alpha_i d_i x_i^T \sum_{i=1}^{N} \alpha_i d_i x_i \right] - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_j d_j x_i^T \alpha_i d_i x_j + \sum_{i=1}^{N} \alpha_i$$

$$= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_j d_j x_j \alpha_i d_i x_i^T + \sum_{i=1}^{N} \alpha_i = Q(\alpha)$$

For the dual problem

$$\max Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i d_i x_i^T \alpha_j d_j x_j$$

subj (1) $\sum_{i=1}^{N} \alpha_i d_i = 0$

(2) $0 \le \alpha_i \le C$

2.2(2) For non-linearly separable case, ie. no linear boundary, we opt for a soft margin whereas for linearly separable cases like $w^T x + b = 0$ hard margin is used. To prove further, in hard margin, it is highly sensitive to a data point being an outlier so soft margin is less likely to overfit when data is not linearly separable

2.3 On the attached code file.