Homework 4

Automated Learning and Data Analysis
Dr. Thomas Price

Spring 2021

Instructions

Due Date: April, 22 2021 at 11:45 PM
Total Points: 65 for CSC522; 58 for CSC422.
Submission checklist:

- Clearly list each team member's names and Unity IDs at the top of your submission.
- Your submission should be a single PDF file containing your answers. **Name your file**: G(homework group number)_HW(homework number), e.g. G1_HW4.
- If a question asks you to explain or justify your answer, **give a brief explanation** using your own ideas, not a reference to the textbook or an online source.
- Submit your PDF through Gradescope under the HW4 assignment (see instructions on Moodle). **Note**: Make sure to add you group members at the end of the upload process.
- In addition to your group submission, please also *individually* submit your Programming portion via our JupyterHub site *and* Moodle.

Problems

1. SVM Theory (16 points CSC 522 / 9 points CSC 422) [**Krishna Gadiraju**].

   (a) Support vector machines (SVM) learn a decision boundary leading to the largest margin between classes. In this question, you will train a SVM on a tiny dataset with 4 data points, shown in Figure 1. This dataset consists of two points with Class 1 (y = -1) and two points with Class 2 (y = 1). Each data point has two non-class attributes: $x1$ and $x2$.

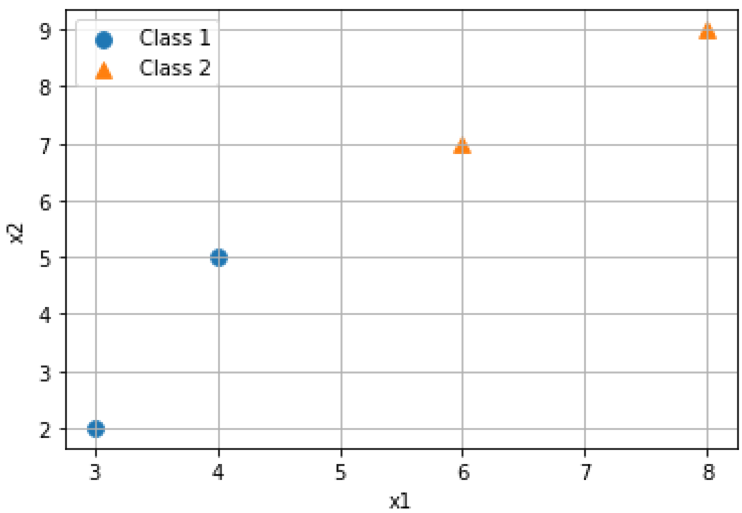

Figure 1: SVM data points for 1(a)

i. Find the weight vector $\mathbf{w}$ and bias w_0 for the decision boundary of a hard-margin SVM. What is the equation corresponding to this decision boundary? Show your work, including the equations you used to derive your answer.

ii. Circle the support vectors and draw the decision boundary.

(b) (12 points) Required for CSC522, Extra Credit for CSC 422): You are given 1-dimensional data points $X_i, i \in [1, 2, 3, 4, 5, 6, 7]$ as shown in Table 1 ,also shown in Figure 2 in this question.

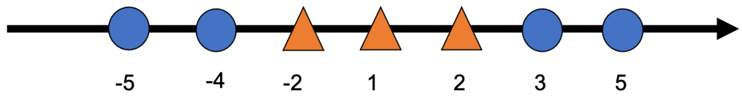

Figure 2: SVM data points for 1(b)

Data ID	x	y
X_1	-5	-1
X_2	-4	-1
X_3	-1	1
X_4	1	1
X_5	2	1
X_6	3	-1
X_7	5	-1

Table 1: Six Data Points

Use this data to answer the following questions:

i. Calculate the equation for the decision boundary of a *hard-margin* SVM, or if this is not possible, explain why in 1-2 sentences.

ii. If you were to train a *soft-margin* SVM on this data, would you select a C value where $C \to 0$ or $C \to \infty$. Explain why in 1 sentence.

iii. Imagine you want to transform the 7 given data points to a higher dimensional space. You decide to use the kernel function $K(X_i, X_j) = (1 + 8X_iX_j)^2$, which is equal to $\phi(X_i) \cdot \phi(X_j)$. What is the function $\phi(X)$? How many dimensions is the transformed data?

iv. Use the function $\phi(X)$ to calculate $\phi(X_i)$ for $i \in [1, 2, 3, 4, 5, 6, 7]$. Graph these data points in the higher-dimensional space. (**Hint**: If the data is more than 2-dimensional, can you simplify your visualization to show it in 2D?)

v. Is it possible to linearly separate the data in the higher-dimensional space? If so, draw the decision boundary in your graph. If not, explain why. **Note**: You do not have to calculate the weights, just draw the decision boundary.

vi. You train a hard-margin SVM on the higher-dimensional data using a library and it gives you the following Lagrange multiplier for your data[1]: $\alpha_2 = 0.1$, $\alpha_5 = 0.2$, $\alpha_6 = 0.1$. What are the remaining Lagrange multipliers, α_1, α_4 and α_7? Justify your answer in 1-2 sentences. (**Hint**: This should not require any math to calculate.)

vii. Recall that the SVM's prediction (using the Kernel transformation) for a data point $\mathbf{Z}$ can be defined with the following equation:

$$f(\mathbf{Z}) = \text{sign}(\sum_i \alpha_i y_i (\phi(\mathbf{X}_i) \cdot \phi(\mathbf{Z})) + w_0) \tag{1}$$

You are given w_0 = -2/3. You are now asked to classify a new test data point, Z, using the SVM defined earlier by the Lagrange multipliers. You do not know what Z's attributes are, but you do know: $K(X_2, Z) = 25$, $K(X_5, Z) = 16$, $K(X_6, Z) = 49$. Classify Z using the SVM. (**Hint**: If you find yourself trying to solve for Z's x value, you are doing it wrong.)

[1]Note, these are not the actual Lagrange multipliers for the SVM, but assume they are for the purposes of this question.

2. K-Means Clustering (14 points) [**Chengyuan Liu**] Use the K-means clustering algorithm with *Euclidean Distance* to cluster the 10 data points in Figure 3 into 3 clusters. Suppose that the initial seed centroids are at points: C, I and H. The data are also given in tabular format in Table 2.

   (a) After each iteration of k-means, report the coordinates of the new centroids and which cluster each data point belongs to. **Stop when the algorithm converges and clearly label on the graph where the algorithm converges.** To report your work, give your answer in tabular format with the following attributes: **Round** (e.g. Round 1, 2, etc), **Points** (e.g. {A, B, C}), and **Cluster_ID** (order does not matter). Also report the **centroids** for each cluster after each round. Please followed the example table format in Table 3

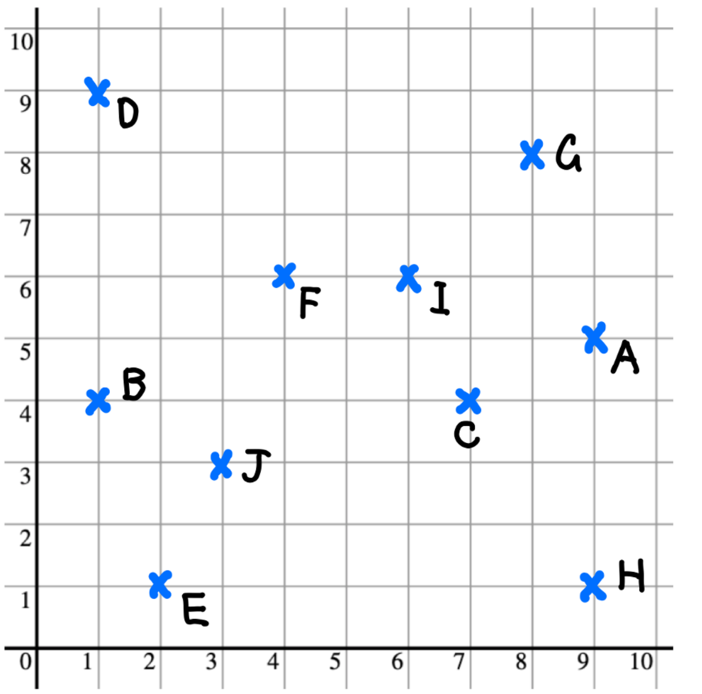

Figure 3: K-means Clustering (a)

Point	**x**	**y**
A	9	5
B	1	4
C	7	4
D	1	9
E	2	1
F	4	6
G	8	8
H	9	1
I	6	6
J	3	3

Table 2: K-means Clustering (b)

Round	Points	Cluster	Centroid
1	A, B, C	1	[0, 1]
	D	2	[1, 0]
	E, F, G , H	3	[2, 0]
2	A, B, C	1	[3, 1]
	D, E	2	[1, 3]
	F	3	[4, 0]

Table 3: K-means Solution Example

(b) How many rounds are needed for the K-means clustering algorithm to converge?

3. Hierarchical Clustering (17 points) [**Chengyuan Liu**] We will use the same dataset as in Question 2 (shown in Figure 3) for Hierarchical Clustering. The *Euclidean Distance* matrix between each pair of the datapoints is given in Figure 4 below:

   (a) Perform *single* link hierarchical clustering. Show your work at each iteration by giving the inter-cluster distances. Report your results by drawing a corresponding dendrogram. The dendrogram should clearly show the order and the height in which the clusters are merged. If possible, use a program to construct your dendrogram (e.g. PowerPoint, LucidChart[2], or VisualParadigm[3]). Scanned hand drawings will also be accepted if they are very clear. **NOTE**: There may be ties (i.e. two clusters have the same distance). In this case, you can choose any order to merge in, and ensure that this is reflected in your dendrogram.

   (b) Perform *complete* link hierarchical clustering on the dataset. As above, show your calculations and report the corresponding dendrogram.

   (c) If we assume there are *two* clusters, will the *single* or *complete* link approach give a better clustering? Justify your answer.

   (d) Consider the single-link hierarchical clustering with **3 clusters**. Compare the quality of this single link clustering with your final K-means clustering in Question 2. To evaluate the quality of each clustering, calculate its corresponding Sum of Squared Error (SSE). Based on this measure, which clustering (k-means, single link), is better? **Please show the SSE value for both clustering.** Note: you may want to write some code to help speed up these calculations, which you can include in lieu of showing your work.

	A	B	C	D	E	F2	G	H	I	J
A	0	8.06	2.24	8.94	8.06	5.1	3.16	4	3.16	6.32
B	8.06	0	6	5	3.16	3.61	8.06	8.54	5.39	2.24
C	2.24	6	0	7.81	5.83	3.61	4.12	3.61	2.24	4.12
D	8.94	5	7.81	0	8.06	4.24	7.07	11.31	5.83	6.32
E	8.06	3.16	5.83	8.06	0	5.39	9.22	7	6.4	2.24
F	5.1	3.61	3.61	4.24	5.39	0	4.47	7.07	2	3.16
G	3.16	8.06	4.12	7.07	9.22	4.47	0	7.07	2.83	7.07
H	4	8.54	3.61	11.31	7	7.07	7.07	0	5.83	6.32
I	3.16	5.39	2.24	5.83	6.4	2	2.83	5.83	0	4.24
J	6.32	2.24	4.12	6.32	2.24	3.16	7.07	6.32	4.24	0

Figure 4: Hierarchical Clustering Dataset

4. Association Rule Mining (6 points) [**Chengyuan Liu**]. Consider the following market basket transactions shown in the Table 4 below.

[2] https://www.lucidchart.com/
[3] https://online.visual-paradigm.com/features/dendrogram-software/

Transaction ID	**Bread**	**Milk**	**Butter**	**Eggs**	**Beer**	**Cola**
1	1	1	0	0	0	1
2	0	0	0	1	0	1
3	0	0	1	0	1	1
4	1	1	0	0	0	1
5	1	1	0	1	0	0
6	0	1	1	1	0	1
7	0	1	0	1	0	0
8	1	1	1	0	0	1
9	1	0	1	1	1	0
10	1	0	1	0	1	0

Table 4: For each transaction (row), a 1 indicates that a given item was present in that transaction, and a 0 indicates that it was not.

(a) What is the maximum number of unique itemsets that can be extracted from this data set (including itemsets that have 0 support)? Briefly explain your answer in 1-2 sentences.

(b) What is the maximum number of association rules that can be extracted from this data set (including rules that have zero support)? Briefly explain your answer in 2-3 sentences.

(c) Compute the support of the itemset: $\{Eggs, Cola\}$?

(d) Compute the support and confidence of association rule: $\{Bread\} \rightarrow \{Butter\}$?

(e) Given min support = 0.3 and min confidence = 0.6, identify all valid association rules of the form $\{A, B\} \rightarrow \{C\}$.

(f) In a different dataset, the support of the rule $\{a\} \rightarrow \{b\}$ is 0.46, and the support of the rule $\{a, c\} \rightarrow \{b, d\}$ is 0.23. What can we say for sure about the support of the rule $\{a\} \rightarrow \{b, d\}$. Explain in 1-2 sentences.

5. Apriori algorithm (12 points) [**Chengyuan Liu**].

Consider the data set shown in Table 5 and answer the following questions using apriori algorithm.

TID	**Items**
t_1	A,C,D,E
t_2	A,B,C,D
t_3	C,D
t_4	A,B,C
t_5	A,B,C,E
t_6	A,D,E
t_7	A,B,D
t_8	A,B,E

Table 5: Apriori algorithm

(a) Show (compute) each step of frequent itemset generation process using the apriori algorithm, with a minimum support count of 3.

(b) Show the lattice structure for the data given in table above, and mark each node in the lattice as either **F**: Frequent, **IC**: Infrequent due insufficient support count, or **IP**: Infrequent due to pruning (we do not need to calculate the support count). (Scanned hand-drawing is acceptable as long as it is clear.)