# Homework 4

Automated Learning and Data Analysis
Dr. Thomas Price

Spring 2022

## Instructions

**Due Date:** April, 25 2022 at 11:45 PM
**Total Points**: 60 for CSC 522; 53 for CSC 422
**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.
- Submit the programming portion of the homework *individually* through JupyterHub.

# 1 SVM (16 points CSC 522 / 9 points CSC 422) [Jianxun WANG]

1a) 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 2. 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 $x2$.

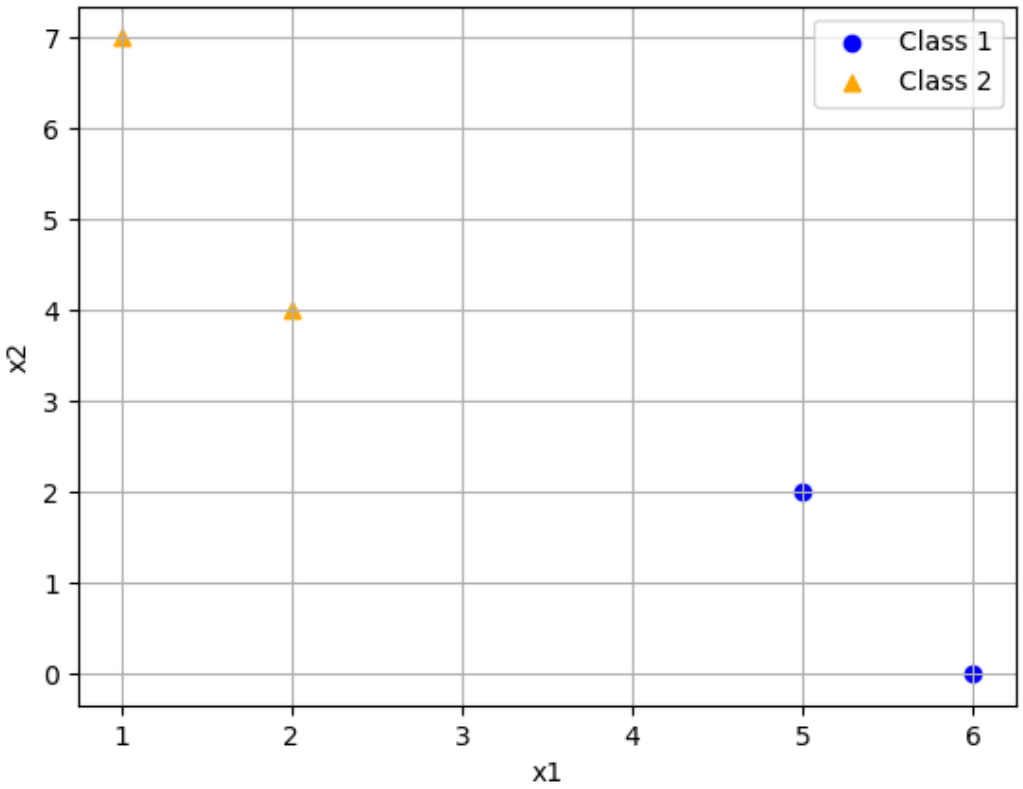



Figure 1: SVM data points for 1(a)

i) (*5 points*) Find the weight vector **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.

**Solution**: Support vectors have coordinates $(2, 4), (5, 2)$
The line goes through both points is $2x_1 + 3x_2 - 16 = 0$
The perpendicular line has form $3x_1 - 2x_2 + w_0 = 0$
The perpendicular line should go through $(3.5, 3)$
The boundary line would be $3x_1 - 2x_2 - 4.5 = 0$
or $6x_1 - 4x_2 - 9 = 0$

Another potential solution: For support vectors, $y_i(w^T x_i + w_0) - 1 = 0$
. Therefore, $2w_1 + 4w_2 + w_0 - 1 = 0$
$-(5w_1 + 2w_2 + w_0) - 1 = 0$
$3w_1 - 2w_2 + 2 = 0$ In addition, $M = \sqrt{(5-2)^2 + (2-4)^2} = \frac{2}{\sqrt{w_1^2+w_2^2}}$
$w_1^2 + w_2^2 = \frac{4}{13}$. Solve for solution, there are $w_1 = -\frac{6}{13} \approx -0.4615$, $w_2 = \frac{4}{13} \approx 0.3076$ and $w_0 = \frac{9}{13} \approx 0.6923$. (Need exact value). So the hyperplane would be $6x_1 - 4x_2 - 9 = 0$

ii) (*4 points*) Circle the support vectors and draw the decision boundary.

**Solution**:

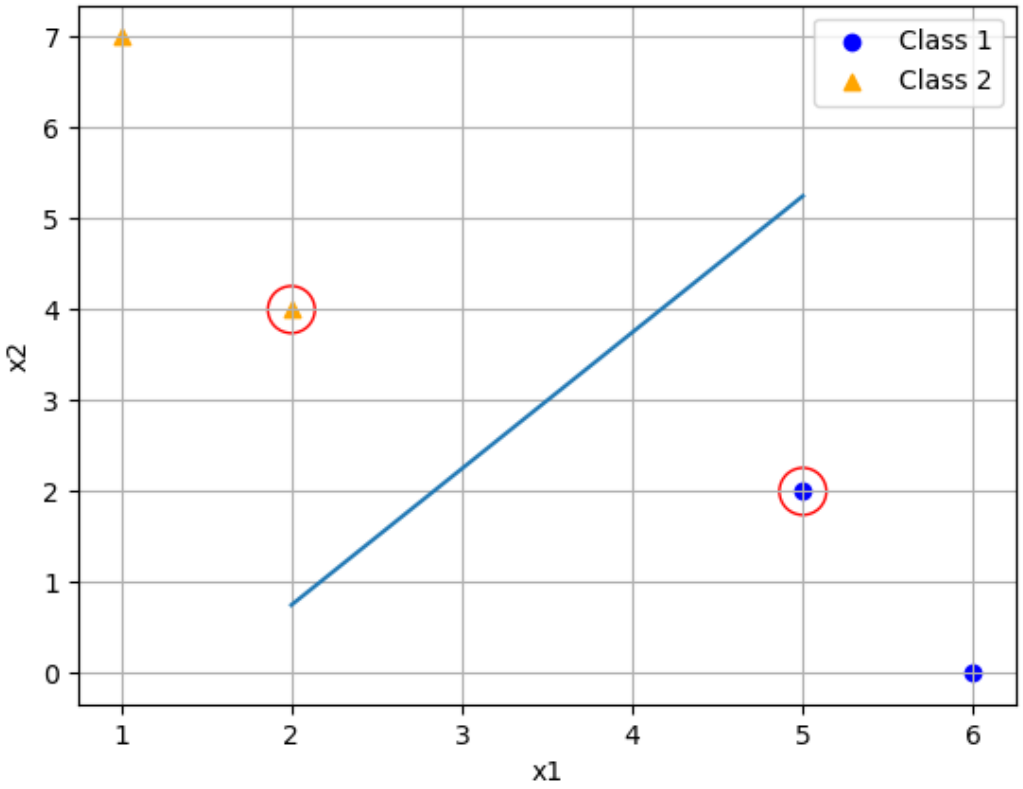



Figure 2: SVM data points for 1(a)

1b) **This whole question is 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 3 in this question.

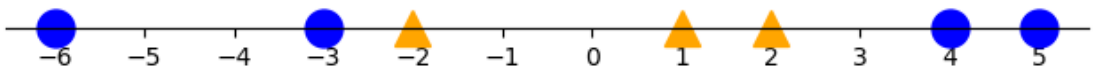



Figure 3: SVM data points for 1(b)

| Data ID | $x$ | $y$ |
|---|---|---|
| $X_1$ | -6 | -1 |
| $X_2$ | -3 | -1 |
| $X_3$ | -2 | 1 |
| $X_4$ | 1 | 1 |
| $X_5$ | 2 | 1 |
| $X_6$ | 4 | -1 |
| $X_7$ | 5 | -1 |

Table 1: Six Data Points

Use this data to answer the following questions:

i) (*1 point*) Calculate the equation for the decision boundary of a *hard-margin* SVM, or if this is not possible, explain why in 1-2 sentences.
**Solution**: It is not possible since the dataset is not linearly separable.

ii) (*1 point*) 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. **Solution**: The bigger $C$, the smaller margin it is, so we need to make $C$ small enough to include the edge points such as $x = -3$, but not too small that makes every point a support vector

iii) (*1 point*) Imagine you want to transform the 7 given data points to a higher dimensional space. You decide to use the kernel function $K(X_i, X_j) = (1 + \frac{9}{2} X_i X_j)^2$, which is equal to $\phi(X_i) \cdot \phi(X_j)$. What is the function $\phi(X)$? How many dimensions is the transformed data? **Solution**: Denote the 1-dimensional vectors as: $X_i = [x_i]$, $X_j = [x_j]$.
$K(X_i, X_j) = (1 + \frac{9}{2} \cdot X_i \cdot X_j)^2 = (1 + \frac{9}{2} \cdot x_i \cdot x_j)^2 =$
$1 + 9x_i x_j + \frac{81}{4}(x_i x_j)^2$
So the function is $\phi(X) = [1, 3x, \frac{9}{2}x^2]$.

iv) (*1 point*) 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?) **Solution**: The transformed 7 points are:
$[1, -18, 162], [1, -9, 40.5], [1, -6, 18], [1, 3, 4.5], [1, 6, 18], [1, 12, 72], [1, 15, 112.5]$
Since the first dimension is always 1, we can show the graph like:

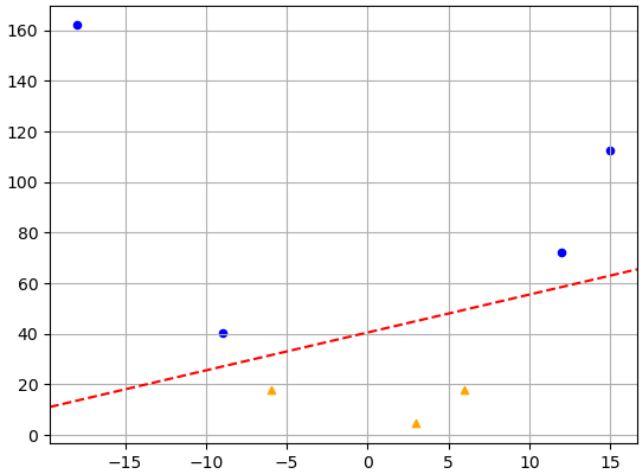



Figure 4: SVM projection result

v) (*1 point*) 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. **Solution**: Yes. See the red line above.

vi) (*1 point*) 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]: $\alpha_2 = 0.1$, $\alpha_3 = 0.2$, $\alpha_6 = 0.1$. What are the remaining Lagrange multipliers, $\alpha_1$, $\alpha_4$, $\alpha_5$ and $\alpha_7$? Justify your answer in 1-2 sentences. (**Hint**: This should not require any math to calculate.) **Solution**: $\alpha_1 = \alpha_4 = \alpha_5 = \alpha_7 = 0$. Since only support vectors have $\alpha \neq 0$, and from the graph it is clear that neither point 1, 4, 5 or 7 are support vectors.

vii) (*1 point*) 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 $w_0 = -2$. 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: $K(X_2, Z) = 36$, $K(X_3, Z) = 12$, $K(X_6, Z) = 9$. Classify $Z$ using the SVM. (**Hint**: If you find yourself trying to solve for $Z$'s $x$ value, you are doing it wrong.) **Solution**: From $K(X_2, Z) = 36$, $K(X_3, Z) = 12$, $K(X_6, Z) = 9$ and $K(X_i, X_j) = \phi(X_i) \cdot \phi(X_j)$, we apply the take the values of the four equations to the $f(\mathbf{Z})$, getting $f(\mathbf{Z}) = sign(-0.1*36+0.2*12-0.1*9-2) = -1$. Thus $\mathbf{Z}$ is classified as $-1$.

[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 6 data points in Figure 5 into 3 clusters. Suppose that the initial seed centroids are at points: C, A and E. The data are also given in tabular format in Table 2.

2a) (*8 points*) 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

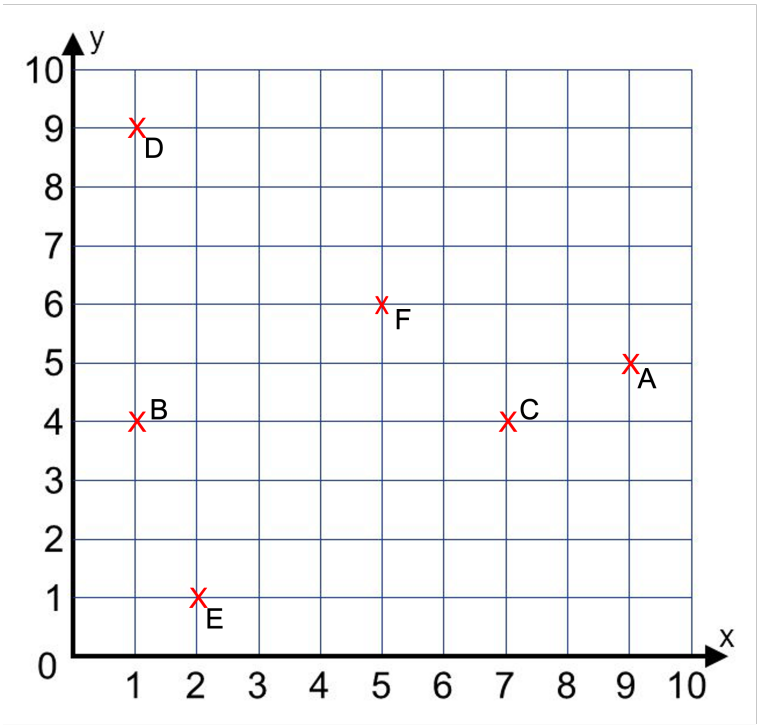



Figure 5: K-means Clustering (a)

| Point | x | y |
|---|---|---|
| A | 9 | 5 |
| B | 1 | 4 |
| C | 7 | 4 |
| D | 1 | 9 |
| E | 2 | 1 |
| F | 5 | 6 |

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

2b) (*2 points*) How many rounds are needed for the K-means clustering algorithm to converge?

2c) (*4 points*) Calculate the Sum of Squared Errors (SSE) for the k-means clustering.

**SOLUTIONS**:

| Round | Points | Cluster | Centroid |
|---|---|---|---|
| 1 | A | 1 | [9, 5] |
| | E, B | 2 | [1.5, 2.5] |
| | F, C, D | 3 | [4.33, 6.33] |
| 2 | A, C | 1 | [8, 4.5] |
| | E, B | 2 | [1.5, 2.5] |
| | F,D | 3 | [3, 7.5] |

Table 4: K-means Clustering Solution (c)

2a)

2b) According to Table 4, the algorithm needs 2 rounds to be converged. At round 3, the same clusters as round 2 are formed.

2c) SSE = 20

# 3 Hierarchical Clustering (10 points) [Chengyuan Liu]

We will use the same dataset as in Question 2 (shown in Figure 5) for Hierarchical Clustering. The *Euclidean Distance* matrix between each pair of the datapoints is given in Figure 6 below:

3a) (*4 points*) 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.

3b) (*4 points*) Perform *complete* link hierarchical clustering on the dataset. As above, show your calculations and report the corresponding dendrogram.

3c) (*2 points*) If we assume there are *two* clusters, will the *single* or *complete* link approach give a better clustering? Justify your answer. (**Hint**: you should not need to calculate SSE to answer the question.)

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0 | 8.06 | 2.24 | 8.94 | 8.06 | 4.12 |
| B | 8.06 | 0 | 6 | 5 | 3.16 | 4.47 |
| C | 2.24 | 6 | 0 | 7.81 | 5.83 | 2.83 |
| D | 8.94 | 5 | 7.81 | 0 | 8.06 | 5 |
| E | 8.06 | 3.16 | 5.83 | 8.06 | 0 | 5.83 |
| F | 4.12 | 4.47 | 2.83 | 5 | 5.83 | 0 |

Figure 6: Euclidean Distance Matrix For Hierarchical Clustering Dataset

**SOLUTIONS**:

3a) See dendrograms for Single and Complete link in Figures 7 and 8.

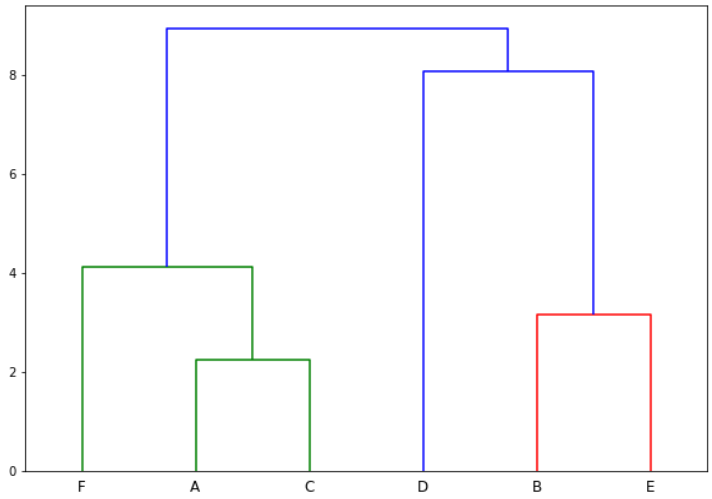



Figure 8: Complete Link

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

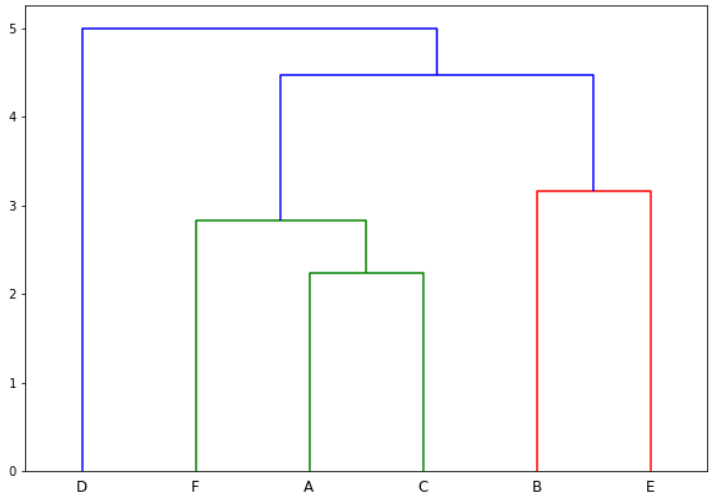



Figure 7: Single Link

3b)

3c) From Figures 7 and 8, we can see that the clusters and constituent points for single link and complete link are different. Complete link with two clusters would have same numbers of points in each cluster, while single link would have 1 data point in one cluster and the remaining points in the other. This suggests complete link creates a more meaningful 2-cluster solution.

# 4 Association Rule Mining (10 points) [Benyamin Tabarsi]

Consider the following market basket transactions shown in the Table 5 below.

| Transaction ID | Cheese | Milk | Butter | Chicken | Wine | Cookie |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 1 | 0 | 1 |
| 2 | 0 | 0 | 1 | 1 | 0 | 0 |
| 3 | 1 | 1 | 0 | 1 | 0 | 0 |
| 4 | 1 | 1 | 1 | 1 | 0 | 1 |
| 5 | 0 | 0 | 1 | 1 | 1 | 0 |
| 6 | 0 | 0 | 0 | 1 | 1 | 1 |
| 7 | 0 | 1 | 1 | 0 | 1 | 1 |
| 8 | 1 | 0 | 1 | 0 | 1 | 0 |
| 9 | 0 | 0 | 0 | 1 | 0 | 1 |
| 10 | 1 | 0 | 0 | 1 | 0 | 0 |

Table 5: For each transaction (row), a 1 indicates that a given item was present in that transaction, and a 0 indicates that it was not.

4a) (*1 point*) What is the maximum number of unique itemsets that can be extracted from this data set (including itemsets that have zero support)? Briefly explain your answer in 1-2 sentences.

4b) (*1 point*) 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.

4c) (*1 point*) Compute the support of the itemset: $\{Chicken, Cookie\}$?

4d) (*2 points*) Compute the support and confidence of association rule: $\{Cheese\} \rightarrow \{Butter\}$?

4e) (*3 points*) Given min support = 0.2 and min confidence = 0.8, identify all valid association rules of the form $\{A, B\} \rightarrow \{C\}$.

4f) (*2 points*) In a different dataset, the support of the rule $\{a\} \rightarrow \{b\}$ is 0.62, and the support of the rule $\{a, c\} \rightarrow \{b, d\}$ is 0.31. What can we say for sure about the support of the rule $\{a\} \rightarrow \{b, d\}$. Explain in 1-2 sentences.

**Solution**:

4a) The maximum number of unique itemsets is $2^6 - 1 = 63$, since there are 6 items which can either be in or not in each itemset.

4b) There are 6 items in the data set. Each item can either be on the left side, the right side, or not in the rule ($3^6$). However, we have to have at least 1 item on the left and right side, so we remove the rules that would break this constraint (there are $2^6$ ways to have no left-side item, $2^6$ to have no right-side item, and one completely empty rule that would overlap them). Therefore the total number of rules is $3^6 - 2^{(6+1)} + 1 = 602$.

4c) For $\{Chicken, Cookie\}$, the support is 4/10=0.4

4d) For $\{Cheese\} \rightarrow \{Butter\}$, support is 2/10=0.2, confidence is 2/4=0.5

4e) $\{Cheese, Milk\} \rightarrow \{Chicken\}$, $\{Milk, Butter\} \rightarrow \{Cookie\}$, $\{Milk, Cookie\} \rightarrow \{Butter\}$, $\{Butter, Cookie\} \rightarrow \{Milk\}$

4f) $0.31 \leq Support(\{a\} \rightarrow \{b, d\}) \leq 0.62$

# 5 Apriori algorithm (10 points) [Benyamin Tabarsi]

Consider the data set shown in Table 6 and answer the following questions using apriori algorithm.

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

Table 6: Apriori algorithm

5a) (*5 points*) Show (compute) each step of frequent itemset generation process using the apriori algorithm, with a minimum support count of 3.

5b) (*5 points*) 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. Tip (optional): Applications like CamScanner and Adobe Scan improve the quality of photos and reduce their size. )

**Solution**:

(a) For the 1-itemsets:

| Itemset | A | B | C | D |
|---|---|---|---|---|
| Count | 7 | 5 | 5 | 4 |

For the 2-itemsets:

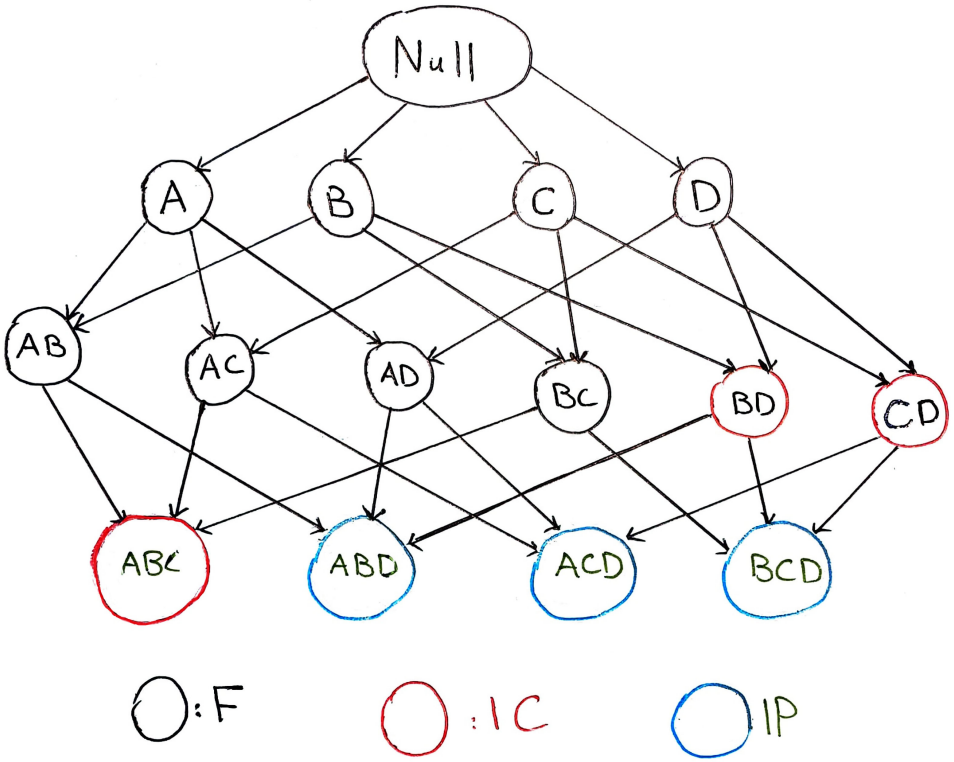
| Itemset_A | AB | AC | AD |
|---|---|---|---|
| Count | 4 | 4 | 4 |
| Itemset_B | BC | BD | |
| Count | 3 | 2 | |
| Itemset_C | CD | | |
| Count | 2 | | |

For the 3-itemsets:

| Itemset_AB | ABC |
|---|---|
| Count | 2 |

Based on the above tables, frequent itemsets could be collected as:
$\{A\}, \{B\}, \{C\}, \{D\}, \{AB\}, \{AC\}, \{AD\}, \{BC\}$

(b) The lattice structure of the data is:


Null
A
B
C
D
AB
AC
AD
BC
BD
CD
ABC
ABD
ACD
BCD
: F
: IC
IP