# Machine Learning

- SCSA1601

Name: Mohnish Devaraj

Reg NO: 39110636

section: C1

## Assignment-1

## PART-A

① Machine Learning

Machine Learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine Learning focuses on the development of computer programs that can access data and use it to learn for themselves.

② There are four types of learning:

-> Supervised learning

-> Unsupervised Learning

-> Semi-supervised Learning

-> Reinforcement Learning

③ The following are the application of Machine learning:

-> Automatic Language Translation

-> Medical Diagnosis

-> Stock Market Trading

-> Online Fraud detection

-> Virtual Personal Assistant

-> Email spam and Malware Filtering

-> Self Driving Cars

-> Product Recommendation

-> Traffic Prediction

-> Speech Recognition

-> Image Recognition

④ An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. the analysis of outlier data is referred to as outlier analysis or outlier mining. Most data mining methods discard outlier noise or exceptions, however, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occuring one and hence, the outlier analysis becomes important in such case.

5) Cross Validation

It is generally used to analyze the test error while a machine learning model is being fitted over several training samples of data (model assessment). It then further helps us select the appropriate model (model selection) based on its complexity.

PART-B

1) (D) It discovers causal relationship

2) (A) High variance

3) Unsupervised machine learning

4) Reinforcement Learning

5) supervised machine learning

PART-C

1)

| Tid | List of item IDs |
|---|---|
| T100 | I1, I2, I5 |
| T200 | I2, I4 |
| T300 | I2, I3 |
| T400 | I1, I2, I4 |
| T500 | I1, I3 |
| T600 | I2, I3 |
| T700 | I1, I3 |
| T800 | I1, I2, I3, I5 |
| T900 | I1, I2, I3 |

- Suppose minimum support count is 2

- Let Minimum confidence is 60%.

step-1:

K=1

- Create a table containing sup-count of each item present in dataset - C1 (Candidate set)

| Itemset | sup-count |
|---|---|
| I1 | 6 |
| I2 | 7 |
| I3 | 6 |
| I4 | 2 |
| I5 | 2 |

⟶ C1

- Compare candidate set items support count with minimum support count (given min_support =2). This gives us itemset L1.

| Itemset | sup-count |
|---|---|
| I1 | 6 |
| I2 | 7 |
| I3 | 6 |
| I4 | 2 |
| I5 | 2 |

⟶ L1

step-2:

k=2

- Generate candidate set C2 using L1 (called as join step). Condition of joining $L_{k-1}$ and $L_{k-1}$ is that it should have (k-2) elements in common.

| Itemset | sup-count |
|---|---|
| I1, I2 | 4 |
| I1, I3 | 4 |
| I1, I4 | 1 |
| I1, I5 | 2 |
| I2, I3 | 4 |
| I2, I4 | 2 |
| I2, I5 | 2 |
| I3, I4 | 0 |
| I3, I5 | 1 |
| I4, I5 | 0 |

→ C2

- check all subsets of an item set are frequent remove that item set. Now find support count of these itemsets by searching in dataset. Compare candidate (c2) support count with minimum support count, this gives us itemset L2.

| Itemset | Sup-count |
|---|---|
| I1, I2 | 4 |
| I1, I3 | 4 |
| I1, I5 | 2 |
| I2, I3 | 4 |
| I2, I4 | 2 |
| I2, I5 | 2 |

→ L2

step-3:

K=3

- Generate candidate set C3 using L2 (join step). Condition of joining $L_{k-1}$ and $L_{k-1}$ is that should have (k-2) elements in common.

| Itemset |
|---|
| I1, I2, I3 |
| I1, I2, I5 |
| I1, I2, I4 |
| I1, I3, I5 |
| I2, I3, I4 |
| I2, I3, I5 |
| I2, I4, I5 |

→ C3

- For L2, first element should match. Check if all subsets of these itemsets are frequent or not and if not, then remove that itemset.

(Here subset of {I1, I2, I3} are {I1, I2} {I2, I3} {I1, I3} which are frequent. For {I2, I3, I4}, subset {I3, I4} is not frequent so remove it. Similarly check for every itemset)

find support count of these remaining itemset by searching in dataset.

| Item set | Sup-count |
|---|---|
| I1, I2, I3 | 2 |
| I1, I2, I5 | 2 |

→ L3

step-4:

- Generate candidate set C4 using L3 (join step). Condition of joining $L_{k-1}$ and $L_{k-1}$ ($k=4$) is that, they should have ($k-2$) elements in common. So here, for L3, first 2 elements (items) should match.

- check all subsets of these are frequent or not (Here itemset formed by joining L3 is {I1, I2, I3, I5} so its subset contains {I1, I3, I5} which is not frequent).

So not itemset in C4

stop, because no frequent itemsets are found further.

Generating Association Rule

$$\text{Confidence } (A \Rightarrow B) = P(B/A) = \frac{\text{support\_count } (A \cup B)}{\text{support\_count } (A)}$$

a) Itemset {I1, I2, I3} from L3

| | |
|---|---|
| {I1, I2} => I5, | confidence = 2/4 = 50% |
| {I1, I5} => I2, | confidence = 2/2 = 100% |
| {I2, I5} => I1, | confidence = 2/2 = 100% |
| I1 => {I2, I5}, | confidence = 2/6 = 66% |
| I2 => {I1, I5}, | confidence = 2/7 = 29% |
| I5 => {I1, I2}, | confidence = 2/2 = 100% |

b) Itemset {I1, I2, I3} from L3

{I1, I2} => $I_3$, Confidence = 2/4 = 50%

{I2, I3} => I1, Confidence = 2/4 = 50%

{I1, I3} => I2, Confidence = 2/4 = 50%

I3 => {I1, I2}, Confidence = 2/5 = 40%

I1 => {I2, I3}, Confidence = 2/6 = 33.33%

I2 => {I1, I3}, Confidence = 2/7 = 28%

As the taken threshold or minimum confidence is 60%, no rules can be considered as the strong association rules for the given problem.