Association Rules

Recap

Rule Evaluation Metrics

1. Support:

   Measure the relative frequency or occurrence of an itemset (a set of items) in a dataset.

   Support(X) = (Number of transactions containing X) / (Total number of transactions)

2. Support Count:

   The actual count of the number of transactions in a dataset that contains a specific itemset.

   Support Count(X) = Number of transactions containing X

3. Confidence
   Indicates the likelihood that an itemset Y appears in a transaction given that another itemset X is present in the same transaction.
   Confidence(X -> Y) = (Number of transactions containing both X and Y) / (Number of transactions containing X)

Apriori Algorithm

An efficient algorithm for frequent itemset mining and association rule learning in large datasets.

- **Frequent Itemset**

  An itemset whose support is greater than or equal to a **minsup threshold.**

- **Anti-Monotone Property of Frequent Itemsets**

  Any subset of a frequent itemset must be also frequent.

Question

Market Basket Analysis (MBA) is one of the key techniques used by large retailers to increase sales by better understanding customer purchasing patterns. Association Rules are widely used to analyse retail basket or customer purchasing data. You have been given the following transaction database that consists of items bought in a store by customers.

Consider the transactions shown in the below table.

Transaction ID	Products
1	Chips, Coke, Hot Dogs, Ketchup, Pizza

2	Chips, Coke, Hot Dogs, Ketchup, Lime
3	Chips, Coke, Donuts, Ketchup, Pizza
4	Chips, Coke, Hot Dogs, Submarine
5	Donuts, Hot Dogs, Ketchup, Pizza
6	Cake, Chips, Coke, Ketchup, Pizza
7	Burgers, Chips, Coke
8	Chips, Soda
9	Buns, Cake, Chips, Coke, Hot Dogs, Ketchup
10	Buns, Chips, Coke, Donuts, Ketchup

Answer the following questions using the information above. Show all your steps clearly.

i. Give the definitions for Support of item A and, Confidence in buying item B when item A has already bought. (1 Mark)

Support(A)= freq(A) / Total number of transactions

Confidence (A=>B) = freq(A, B) / freq(A)

ii. Find the support values for buying Chips, Coke, Lime and Pizza individually. (2 Marks)

Sup (Chips) = 0.9
Sup (Coke) = 0.8
Sup (Lime) = 0.1
Sup (Pizza) = 0.4

iii. Find the confidence values for buying Chips, given that Pizza is already bought. (1 Mark)

Conf (Pizza -> Chips) = 3/4 =0.75

iv. Find the support and confidence for the associations given below. (3 Marks)

a. {Coke} → {Chips, Pizza}
b. {Chips, Cake} → {Hot Dogs}
c. {Chips, Coke, Pizza} → {Hot Dogs, Ketchup}

Rule	Support	Confidence
{Coke} → {Chips, Pizza}	0.3	3/8 = 0.375
{Chips, Cake} → {Hot Dogs}	0.1	½ = 0.5
{Chips, Coke, Pizza} → {Hot Dogs, Ketchup}	0.1	1/3 = 0.33

v. Find all the frequent itemsets (sets of 1 item, 2 items, 3 items, etc.) Assume that minimum support count > 4. (10 Marks)

Find itemsets of size-01

Itemset	Support Count
Buns	2
Burgers	1
Cake	2
Chips	9
Coke	8
Donuts	3
Hot Dogs	5
Ketchup	7
Lime	1
Pizza	4
Soda	1
Submarine	1

Here the support count of Buns, Burgers Cake, Donuts, Lime, Soda and Submarine are less than the minimum support count. Therefore, all the other elements are candidates for itemsets of size-02.

Find itemsets of size-02

Itemset	Support Count
Chips, Coke	8
Chips, Hot Dogs	4
Chips, Ketchup	6
Chips, Pizza	3
Coke, Hot Dogs	4
Coke, Ketchup	6
Coke, Pizza	3
Hot Dogs, Ketchup	4
Hot Dogs, Pizza	2
Ketchup, Pizza	4

7, 2-itemsets have a support count greater than the minimum support count. Therefore, they will be candidates for the itemsets of size 3.

Remaining elements: Chips, Coke, Hot Dogs, Ketchup, Pizza

Find itemsets of size-03

Itemset	Support Count	Comments
Chips, Coke, Hot Dogs	4	
Chips, Coke, Ketchup	6	
Chips, Coke, Pizza	3	Rejected, due to lower sup. count
Chips, Hot Dogs, Ketchup	3	Rejected, due to lower sup. count

Chips, Hot Dogs, Pizza		Rejected, subsets ({Chips, Pizza} and {Hot Dogs, Pizza} are not frequent
Chips, Ketchup, Pizza		Rejected, subset ({Chips, Pizza} is not frequent
Coke, Hot Dogs, Ketchup	3	
Coke, Hot Dogs, Pizza		Rejected, subsets (Coke, Pizza} and {Hot Dogs, Pizza} are not frequent
Coke, Ketchup, Pizza		Rejected, subset ({Coke, Pizza} is not frequent
Hot Dogs, Ketchup, Pizza		Rejected, subset ({Hot Dogs, Pizza} is not frequent

2, 3-itemsets have a support count greater than the minimum support count. Therefore, they will be candidates for the itemsets of size 4.

Remaining elements: Chips, Coke, Hot Dogs, Ketchup

Itemset	Support Count	Comments
Chips, Coke, Hot Dogs, Ketchup		Rejected, some subsets ({Chips, Hot Dogs, Ketchup}, {Coke, Hot Dogs, Ketchup}, etc.) are not frequent

None of the size 4 itemsets are frequent.

So, finally we have 5 1-itemsets, 7 2-itemsets, and 2 3-itemsets which are frequent.

Therefore, frequent itemsets: {Chips}, {Coke}, {Hot Dogs}, {Ketchup}, {Pizza}, {Chips, Coke}, {Chips, Hot Dogs}, {Chips, Ketchup}, {Coke, Hot Dogs}, {Coke, Ketchup}, {Hot Dogs, Ketchup}, {Ketchup, Pizza}, {Chips, Coke, Hot Dogs}, {Chips, Coke, Ketchup}

In summary, following are the frequent itemsets.

Chips	9	Chips, Coke	8	Chips, Coke, Hot Dogs	4
Coke	8	Chips, Hot Dogs	4	Chips, Coke, Ketchup	6
Hot Dogs	5	Chips, Ketchup	6		
Ketchup	7	Coke, Hot Dogs	4		
Pizza	4	Coke, Ketchup	6		
		Hot Dogs, Ketchup	4		
		Ketchup, Pizza	4		

vi. Using the frequent itemsets, find all the closed frequent itemsets which are not maximal. Show all the steps/results of your work clearly and justify any decision you have taken in your analysis. (8 Marks)

Closed frequent itemsets:

If we get {Coke} for an example, it's immediate superset {Chips, Coke} has the same support as itself: 8. Therefore, {Coke} is not a closed frequent itemset. Frequent itemset {Coke, Ketchup} also has one immediate superset - {Chips, Coke, Ketchup} which has the same support as itself. Therefore, {Coke, Ketchup} is also not a closed frequent itemset. Likewise, we can omit the frequent itemsets which has at least one immediate superset which has the same support as itself when finding the closed frequent itemsets.

In summary, following are the closed frequent itemsets we can get from the given dataset.

Chips	9	Chips, Coke	8	Chips, Coke, Hot Dogs	4
Hot Dogs	5	Hot Dogs, Ketchup	4	Chips, Coke, Ketchup	6
Ketchup	7	Ketchup, Pizza	4		

But we are interested in those closed itemsets which are not maximal ones. Let's try to find those which are maximal ones and remove them from the closed list. Then, the remaining ones will be the required answer.

Maximal frequent itemsets:

Hot Dogs, Ketchup	4
Ketchup, Pizza	4
Chips, Coke, Hot Dogs	4
Chips, Coke, Ketchup	6

As the above given closed frequent itemsets have no immediate supersets that are frequent, they are maximal frequent itemsets.

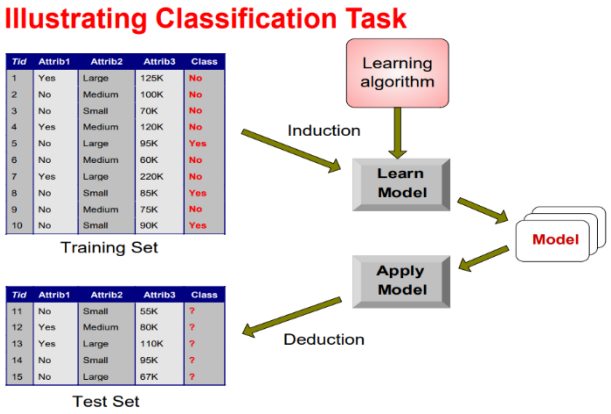
Therefore, closed frequent itemsets that are not maximal frequent itemsets are as follows:

Chips	9
Hot Dogs	5
Ketchup	7
Chips, Coke	8

Decision Trees

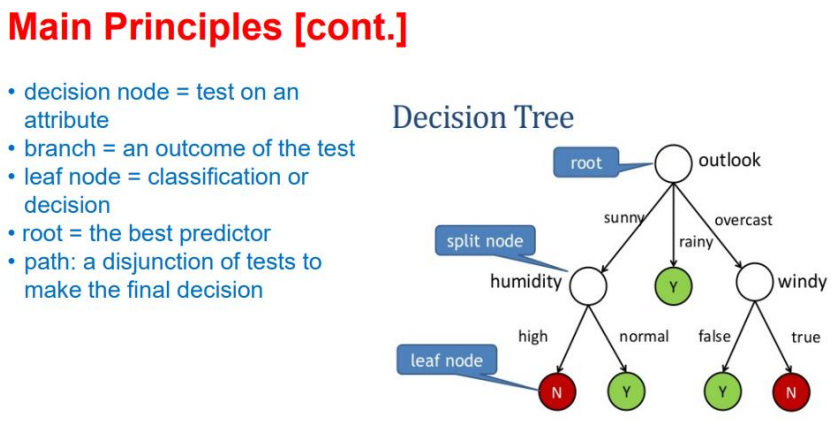
Recap

- Tree-based methods are one of the commonly used supervised learning algorithms in machine learning, both for regression and classification problems.
- Why?
  - Relatively fast compared to other classification models.
  - Obtain similar and sometimes better accuracy compared to other models.
  - Simple and easy to understand.
  - Can be converted into simple and easy to understand classification rules.



The training examples are used for choosing appropriate tests in the decision tree. Typically, a tree is built from top to bottom, where tests that maximize the information gain about the classification are selected first.

Main Principles



Classification on new instances is done by following a matching path from the root to a leaf node

The criterion used to select the best split?

Entropy

- Entropy measures the homogeneity (randomness in the data) of a dataset.

$$\text{Entropy} = \sum_i - p_i \log_2 p_i$$

p_i is the probability of class i

- In the context of binary classification problems, where there are two possible outcomes (e.g., positive and negative), entropy is calculated using the proportion of positive and negative samples in a dataset.
- Formula:

  E = -p_pos * log2(p_pos) - p_neg * log2(p_neg)

- The entropy for a completely pure set is 0 and is 1 for a set with equal occurrences for both the classes.

Information Gain

- Determine which attribute in a given set of training feature vectors is most useful for discriminating between the classes to be learned.
- It tells us how important a given attribute of the feature vectors is. We will use it to decide the ordering of attributes in the nodes of a decision tree.

Information Gain = entropy(parent) – [average entropy(children)]

Question

You have been appointed as a consultant to analyze the characteristics of Hogwarts house placements from a randomly selected set of students. Table H1b shows data collected from the students.

Magical Ability	Blood Status	House	Quidditch Player?
High	Half-Blood	Gryffindor	Yes
Low	Half-Blood	Hufflepuff	Yes
Medium	Half-Blood	Hufflepuff	No
High	Pure-Blood	Hufflepuff	Yes
Medium	Half-Blood	Gryffindor	No

High	Pure-Blood	Hufflepuff	Yes
Low	Half-Blood	Gryffindor	No
Low	Pure-Blood	Hufflepuff	Yes
Medium	Half-Blood	Gryffindor	No
High	Pure-Blood	Hufflepuff	Yes
Low	Pure-Blood	Gryffindor	Yes
High	Pure-Blood	Hufflepuff	Yes

Answer the following questions based on the information given in Table:

1. Calculate the overall entropy before splitting.
2. Calculate the entropy after splitting each attribute.
3. At which attribute should the decision tree split first? Explain why.
4. Draw the final classifier tree (decision tree), with house labels in its leaf nodes.

Naïve Bayes

Recap

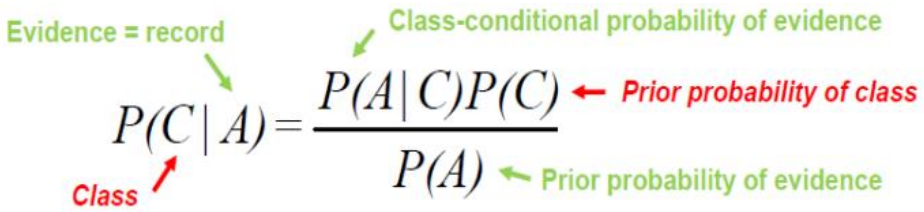
Conditional Probability

The conditional probability of (X|Y) is the probability that X will occur, given Y.

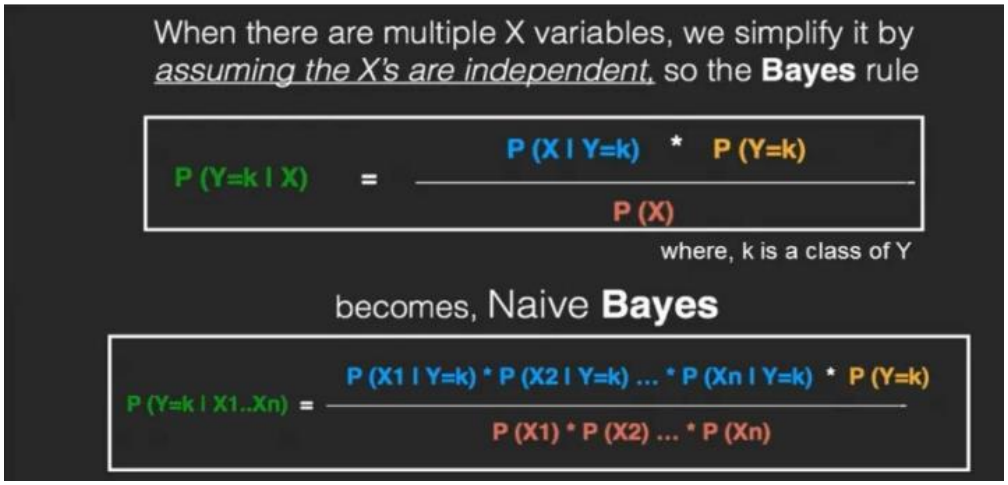
Law of Conditional Probability

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{P(X) \cdot P(Y|X)}{P(Y)}$$

- The second formula is known as Bayes Theorem.
- We can use it for single evidence classifications.



Naïve Bayes (Multiple Attributes)



Question

Given the training data in the table below (Buy Computer dataset) predict the class (yes or no) of the following example using Naïve Bayes classification: **age<=30, income=medium, student=yes, credit-rating=fair** (this is our target)

RID	age	income	student	credit_rating	Class: buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31 . . . 40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31 . . . 40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31 . . . 40	medium	no	excellent	yes
13	31 . . . 40	high	yes	fair	yes
14	>40	medium	no	excellent	no

Class:

C1:buys_computer = 'yes'

C2:buys_computer = 'no'

Data to be classified: (our objective)

E = (age <=30, Income = medium, Student = yes, Credit_rating = Fair)

Prior probability P(Ci):

P(buys_computer = "yes") = 9/14 = 0.643

P(buys_computer = "no") = 5/14= 0.357

Let's decompose our objective

E1 is age<=30, E2 is income=medium, E3 is student=yes, E4 is credit-rating=fair

We need to compute P(yes|E) and P(no|E) and compare them.

Compute P(Ei|Ci) for each class

P(E1 |buys_computer = “yes”) = 2/9 = 0.222

P(E1 |buys_computer = “no”) = 3/5 = 0.6

P(E2 | buys_computer = “yes”) = 4/9 = 0.444

P(E2 | buys_computer = “no”) = 2/5 = 0.4

P(E3 | buys_computer = “yes) = 6/9 = 0.667

P(E3 | buys_computer = “no”) = 1/5 = 0.2

P(E4 | buys_computer = “yes”) = 6/9 = 0.667

P(E4 | buys_computer = “no”) = 2/5 = 0.4

P(E|Ci) :

P(E|buys_computer = “yes”) = 0.222 x 0.444 x 0.667 x 0.667 = 0.044

P(E|buys_computer = “no”) = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

P(E|Ci)*P(Ci) :

P(E|buys_computer = “yes”) * P(buys_computer = “yes”) = 0.028

P(E|buys_computer = “no”) * P(buys_computer = “no”) = 0.007

Since Red > Blue here, E belongs to class (“buys_computer = yes”)

Check: Here the denominator P(E) is not calculated, as it is the same in both cases, so does not have a practical effect in the result. How much is P(E) however? Consult the slides to find it out!