Classification - A Two-Step Process

Model Construction: Describing a set of predetermined classes.

* Each Tuple/Sample is assumed to belong to a predefined class, as determined by the class label attribute.
* The set of tuples used for model construction is training set
* The model is represented as classification rules, decision trees or mathematical formula

Model Usage: for classifying future or unknown objects

* Estimate Accuracy of the Model
  + The known label of test sample is compared with the classified result from the model
  + **Accuracy rate is the percentage of test set samples that are correctly classified by the model**
  + Test set is independent of training set, otherwise overfitting will occur
* If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known

Classification Algorithms

|  |  |
| --- | --- |
| Marks | Pass/Fail |
|  |  |
|  |  |

Example:

A Classification Problem: predict the grades for students taking this course

Key Steps:

1. Data: What “Past Experience” can we rely on?
2. Assumptions: what can we assume about the students or the course?
3. Representation: How do we “summarize” a student?
4. Estimation: how do we construct a map from students to grades?
5. Evaluation: how well are we predicting?
6. Model Selection: perhaps we can do even better?

<======= Lecture Missed =======>

**What is Entropy? How to calculate entropy?**It is a measure of uncertainty.

What is Information Gain? Formula for Information Gain.

**SAMPLE DATA TABLE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **Outlook** | **Temperature** | **Humidity** | **Play** |
| A | Sunny | Hot | High | No |
| B | Sunny | Hot | High | No |
| C | Overcast | Hot | High | Yes |
| D | Rainy | Mild | High | Yes |
| E | Rainy | Cool | Normal | Yes |
| F | Rainy | Cool | Normal | No |
| G | Overcast | Cool | Normal | Yes |
| H | Sunny | Mild | High | No |
| I | ––Sunny | Cool | Normal | Yes |
| J | Rainy | Mild | Normal | Yes |
| K | Sunny | Mild | Normal | Yes |
| L | Overcast | Mild | High | Yes |
| M | Overcast | Hot | Normal | Yes |
| N | Rainy | Mild | High | No |

entropy (Play Tennis)

**Step 1**: Calculate entropy of (X)  
 entropy of (X) = P1 (Play Tennis=YES) = 9/14  
 P2 (Play Tennis=No) = 5/14

**Step 2**: Use Formula (Considering the undivided Values)

P1 (-(log P1)) + P2 (-(log P2))

**Step 3**: entropy of Outlook Column

P(Rain)=0.971, P(Sunny)=0.971, P(Overcast)= 0

\*How to calculate LOG Value using Scientific Calculator.

**Step 4**: Information= Entropy (X) – Entropy (X/Y)

Calculate Information Gain. Maximum Information Gain Variable is going to be the Root of the Decision Tree.

Decision Tree will be like:

YES

YES

NO

NO

YES

Sunny

Overcast

Rainy

High

Normal

False

True

Derive Decision Tree Rules like:

* **IF** Outlook = Overcast then Play = YES
* **IF** Outlook = Sunny **and** Humidity= High **then** Play = No
* **IF** Outlook = Sunny **and** Humidity=Normal **then** Play = Yes
* **IF** Outlook =Windy **and** Humidity=True **then** Play = No
* **IF** Outlook =Windy **and** Humidity=False **then** Play = Yes

**Q.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Weekend** | **Weather** | **Parents** | **Money** | **Decision** |
| W1 | Sunny | Yes | Rich | Cinema |
| W2 | Sunny | No | Rich | Tennis |
| W3 | Windy | Yes | Rich | Cinema |
| W4 | Rainy | Yes | Poor | Cinema |
| W5 | Rainy | No | Rich | Stay In |
| W6 | Rainy | Yes | Poor | Cinema |
| W7 | Windy | No | Poor | Cinema |
| W8 | Windy | No | Rich | Shopping |
| W9 | Windy | Yes | Rich | Cinema |
| W10 | Sunny | No | Rich | Tennis |

Answers:

Entropy (x) = 1.571

And we need to determine the best of:  
Entropy (S sunny) = 0.918  
Entropy (S windy) = 0.811  
Entropy (S rainy) = 0.918  
Entropy (S weather) = 0.070

<<answer remaining>>

<<Flowchart remaining>>

**Pruning**: It is a technique in machine learning that reduces the size of decision trees by removing selections of the tree that provides little power to classify instances.

The goal of pruning is:

1. To reduce complexity of the final classifier as well as better predictive accuracy by the reduction of overfitting.
2. Removal of sections of a classifier that may be based on noisy or erroneous data.

Pruning should reduce the size of a learning tree without reducing predictive accuracy as measured by a test set or using cross-validation.

How to Address Overfitting

Pre-Pruning (Early Stopping Rule)

* Stop the algorithm before it becomes a fully-grown tree
* Stop if all instances belong to the same class
* Stop if all the attribute values are the same

Post-Pruning

* Grow decision tree to its entirety.
* Trim the notes of the decision tree in a bottom-up fashion.

Q. Find out Advantages and Disadvantages of Decision Tree.

Disadvantages of using ID3.

* Irrelevant attributes may affect badly the construction of a decision tree. Eg: ID Numbers.
* Small variations in the data can imply that very different looking trees are generated.
* Data may be over fitted or over classified, if a small sample is tested.
* Only one attribute at a time is tested for making a decision.
* Not good for predicting the values of a continuous class attribute.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AGE** | INCOME | STUDENT | CREDIT\_RATING | BUYS\_COMPUTER |
| <=30 | High | No | Fair | No |
| <=30 | High | No | Excellent | No |
| 31-40 | High | No | Fair | Yes |
| >40 | Medium | No | Fair | Yes |
| >40 | Low | Yes | Fair | Yes |
| >40 | Low | Yes | Excellent | No |
| 34-40 | Low | Yes | Excellent | Yes |
| <=30 | Medium | No | Fair | No |
| <=30 | Low | Yes | Fair | Yes |
| >40 | Medium | Yes | Fair | Yes |
| <=30 | Medium | Yes | Excellent | Yes |
| 31-40 | Medium | No | Excellent | Yes |
| 31-40 | High | Yes | Fair | Yes |
| >40 | Medium | No | Excellent | No |
| 31-40 | Low | Yes | Fair | ?????? |
| >40 | Medium | Yes | Excellent | ?????? |
| 31-40 | High | No | Excellent | ?????? |

Calculate Entropy, Information Gain and draw Decision Tree.

**Baye’s Theorem**

Let X be a data tuple and H be hypothesis, such that **X belongs to a specific class C.**

Probability of a hypothesis h on X, P(X|P)

**P(Ci | X) = P(X | Ci) P(Ci)**

Q.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AGE** | INCOME | STUDENT | CREDIT\_RATING | BUYS\_COMPUTER |
| <=30 | High | No | Fair | No |
| <=30 | High | No | Excellent | No |
| 31-40 | High | No | Fair | Yes |
| >40 | Medium | No | Fair | Yes |
| >40 | Low | Yes | Fair | Yes |
| >40 | Low | Yes | Excellent | No |
| 34-40 | Low | Yes | Excellent | Yes |
| <=30 | Medium | No | Fair | No |
| <=30 | Low | Yes | Fair | Yes |
| >40 | Medium | Yes | Fair | Yes |
| <=30 | Medium | Yes | Excellent | Yes |
| 31-40 | Medium | No | Excellent | Yes |
| 31-40 | High | Yes | Fair | Yes |
| >40 | Medium | No | Excellent | No |

Given the following training data set, find the outcome for the following test case.

Data sample : x = (age<=30,

Income=medium,

Student=yes,

Credit\_rating=fair) -> buys\_computer?

Answer:

Class: c1= buys\_computer = yes = (9/14) = **0.643**

C2= Buys\_computer = no = (5/14) = **0.357**

Base on the data sample, we need to compare each probability with the last column.

Compute P(X|Ci)

**P(** Age<=30 ------ buys\_computer**)** = yes ( 2/9 ) = 0.22   
 ------ buys\_computer **)** = no ( 3/5 ) = 0.6

**P(** Income=medium ------ buys\_computer **)** = yes ( 4/9 ) = 0.44  
 ------ buys\_computer **)** = no ( 2/5 ) = 0.40

**P(** Student=yes ------ buys\_computer **)** = yes ( 6/9 ) = 0.66  
 ------ buys\_computer **)** = no ( 1/5 ) = 0.20

**P(** Credit\_rating=fair ------ buys\_computer **)** = yes ( 6/9 ) = 0.66  
 ------ buys\_computer **)** = no ( 2/5 ) = 0.40

Calculating **P(X|Ci):**

Probability of YES = Multiplying all YES = 0.22 x 0.44 x 0.66 x 0.66 = **0.042**

Probability of NO = Multiplying all NO = 0.60 x 0.40 x 0.20 x 0.40 = **0.019**

Calculating **P(X|Ci)\* P(Ci):**

0.042 x 0.643 = 0.027 (YES)

0.019 x 0.657 = 0.012 (NO)

**Therefore X belongs to the class (buys\_computer=’yes’)**

**Naïve Bavesian Classifier: Comments**

* Advantages:
  + Easy to Implement
  + Good results obtained in most of the cases
* Dis advantages
  + Assumption of class conditional independence, therefore loss of accuracy
  + Practically dependencies exist among variables
  + Eg: hospitals: patients, Profile: age, Family history etc.

Classifier Accuracy Measures

* Accuracy of a Classifier M, acc(M): percentage of test set tuples that are correctly classified by the model M.
  + Error Rate (misclassification rate) of M
  + Given m classes, CMi,k, an entry in a confusion matrix, indicates # of tuples in class i that are labeled by the classifier as class j.

True Positive Rate: Actual Class and predicted class are same.

False Positive Rate: Predicted to be in class but does not belong to that class (Nos of -ve examples predicted as +ve)

False Negative: Number of positive examples wrongly predicted as negative by classification model.

True Negative: Number of negative examples correctly by classification model.

Accuracy of Classifier = Correct Records/ Total No. of Records

|  |  |  |
| --- | --- | --- |
|  | Actual Class (Expectation) | |
| Predicted Classes (observation) | TP **(True Positive)** | FP **(FalsePositive)** |
| FN (False Negative) | TN (True Negative) |

**Nearest Neighbor Algorithm**

*K-Nearest Algorithm*

To Determine the class of a new example E:

* + Calculate the distance between E and all examples in the training set.
  + Select K-nearest examples to E in the training set.

Closeness is defined in terms of the *Euclidean* distance between two examples.

The Educlidean distance between X=(X1,X2,X3,…Xn) and Y=(Y1,Y2,Y3,…Yn) is defined as :

Example: Nearest Neighbors

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Customer** | **Age** | **Income** | **No. Credit Cards** | **Response** |
| Jhon | 35 | 35K | 3 | No |
| Rachel | 22 | 50K | 2 | Yes |
| Hannah | 63 | 200K | 1 | No |
| Tom | 59 | 170K | 1 | No |
| Nellie | 25 | 40K | 4 | Yes |
| David | 37 | 50K | 2 | ? |

Answer:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Customer** | **Age** | **Income** | **No. of CC** | **Response** | **Distance** |
| Jhon | 35 | 35K | 3 | No | Sqrt[ (35-37)2 + (35-50)2 + (3-2)2 ] = 15.16 |
| Rachel | 22 | 50K | 2 | Yes | Sqrt[ (22-37)2 + (50-50)2 + (2-2)2 ] = 15 |
| Hannah | 63 | 200K | 1 | No | Sqrt[ (63-37)2 + (200-50)2 + (1-2)2 ] = 152.2 |
| Tom | 59 | 170K | 1 | No | Sqrt[ (59-37)2 + (170-50)2 + (1-2)2 ] = 122 |
| Nellie | 25 | 40K | 4 | Yes | Sqrt[ (25-37)2 + (40-50)2 + (3-2)2 ] = 15.65 |
| David | 37 | 50K | 2 | ***YES*** |  |

David’s Record is nearest to Rachel because the distance between them is 15 (which is lowest among the remaining) therefore Response would be YES.

Strengths and Weaknesses:

Strengths:

* Simple to implement and use
* Very flexible, data-driven
* Comprehensible – easy to explain prediction
* Robust to noisy data by averaging K-nearest neighbors.
* Can also be used for continuous data.

Weaknesses:

* Need a lot of space to store all examples
* Take more time to classify a new example than with a model (need to calculate and compare distance from new example to all other examples)
* Beware of over fitting! Need a test set.

Distance between neighbors could be dominated by some attributes with relatively large numbers.

Normalize the data before calculating by dividing all the values of the one column with the highest value of that column.

**Associate Mining**

Given a data set, find the items in the data that are associated with each other.

Association is measured as frequency of occurrence in the same context.

The definition of context depends on the Application

**Frequently Used Terms**

* Item Set: A collection of one or more items.
* K-item set: An item set containing k items
* Frequent item set: An item set whose support is greater than a minimum **threshold (cut off value to define any item as frequent item)**

Every association rule has a **support and a confidence.**

**Support:** Fraction of transactions that contain an itemset.  
Eg: Database with transactions (customer\_#: item\_a1, item\_a2, …)

1. 1, 3, 5

2. 1, 8, 14, 17, 12

3. 4, 6, 8, 12, 9, 104

4. 2, 1, 8

Support {8,12} = 2 (or 50% ~ 2 of 4 customers)

Support {1,5}= 1 (or 25%~ 1 of 4 customers)

*\*support: No. of times set is called in given data set*

**Frequent:** An itemset is called frequent if its support is equal or greater than an agreed upon minimal value – the support threshold

Add to previous example: -

If threshold 50%  
then item sets {8,12} and {1} called frequent

**Confidence:** The confidence is the conditional probability that, given X present in a transition, Y will also be present.

*An association rule is of the form: X = > Y*

*X = > Y: if someone buys X, he also buys Y*

Confidence measure, by definition:

*Confidence(X=>Y) = support (x,y) / support (x)*

We should only consider rules derived from item sets with high support, and that also have high confidence.

**“A rule with low confidence is not meaningful”**

Rules don’t explain anything, they just point out hard facts in data volumes.

Q.

1 3,5,8

2 2,6,8

3 1,4,7,10

4 3,8,10

5 2,5,8

6 1,5,6

7 4,5,6,8

8 2,3,4

9 1,5,7,8

10 3,8,9,10

Support (5) = 5

Confidence (5,8) = 4/5 (no. of times [5,8] occur together / no. of times 5 occurred)

**= 80%**

Confidence (8,5) = 4/7 (no. of times [8,5] occur together / no. of times 8 occurred)

**= 57%**

*IF Confidence is High, and Support is low, rule is still not meaningful.*

Remove Items with support 1 (below threshold)

|  |  |
| --- | --- |
| **TID** | **ITEMS** |
| 100 | 1 3 4 |
| 200 | 2 3 5 |
| 300 | 1 2 3 5 |
| 400 | 2 5 |

|  |  |
| --- | --- |
| **ITEM** | **Support** |
| (1) | 2 |
| (2) | 3 |
| (3) | 3 |
| (4) | 1 |
| (5) | 3 |

|  |  |
| --- | --- |
| **ITEM** | **Support** |
| (1) | 2 |
| (2) | 3 |
| (3) | 3 |
| (5) | 3 |

|  |  |
| --- | --- |
| **ITEM** | **Support** |
| (1,2) | 1 |
| (1,3) | 2 |
| (1,5) | 1 |
| (2,3) | 2 |
| (2,5) | 3 |
| (3,5) | 2 |

|  |  |
| --- | --- |
| **ITEM** | **Support** |
| (1,3) | 2 |
| (2,3) | 2 |
| (2,5) | 3 |
| (3,5) | 2 |

Combine each item with remaining one and eliminate duplicate (eg: 1,2 and 2,1) Remove Items with support 1 (below threshold)

Combine each item with remaining one and eliminate duplicate (eg: 1,3 and 2,3 = {1,2,3}) Remove Items with support 1 (below threshold)

|  |  |
| --- | --- |
| **ITEM** | **Support** |
| (1,3,2) | 1 |
| (1,3,2,5) | 1 |
| (1,3,5) | 1 |
| (2,3,5) | 2 |

|  |  |
| --- | --- |
| **ITEM** | **Support** |
| (2,3,5) | 2 |

Confidence (2,3,5) = 2/3 = **66.66%**

**Q1.**

|  |  |
| --- | --- |
| **TID** | **ITEM** |
| T1 | {M,O,N,K,E,Y} |
| T2 | {D,O,N,K,E,Y} |
| T3 | {M,A,K,E} |
| T4 | {M,U,C,K,Y} |
| T5 | {C,O,K,I,E} |

|  |  |
| --- | --- |
| **ITEM** | **SUPPORT** |
| M | 3 |
| O | 3 |
| N | 2 |
| K | 5 |
| E | 4 |
| Y | 3 |
| D | 1 |
| A | 1 |
| U | 1 |
| C | 2 |
| I | 1 |

|  |  |
| --- | --- |
| **ITEM** | **SUPPORT** |
| M | 3 |
| O | 3 |
| K | 5 |
| E | 4 |
| Y | 3 |

|  |  |
| --- | --- |
| **ITEM** | **SUPPORT** |
| M,O | 1 |
| M, K | 3 |
| M, E | 2 |
| M, Y | 2 |
| O, K | 3 |
| O, E | 3 |
| O,Y | 2 |
| K, E | 4 |
| K,Y | 3 |
| E,Y | 2 |

|  |  |
| --- | --- |
| **ITEM** | **SUPPORT** |
| M, K | 3 |
| O, K | 3 |
| O, E | 3 |
| K, E | 4 |
| K,Y | 3 |

|  |  |
| --- | --- |
| **ITEM** | **SUPPORT** |
| M K O | 1 |
| M K O E | 1 |
| M K E | 2 |
| M K Y | 1 |
| O K E | 3 |
| O K Y | 2 |
| K E Y | 2 |

|  |  |
| --- | --- |
| **ITEM** | **SUPPORT** |
| O K E | 3 |

**Ans:** OKE -> 3

Confidence **:**

1. O,K -> (3/3) = 1 = **100%**
2. O,E -> (3/3) = 1 = **100%**
3. K,E -> ( 4/5 ) =0.8 = **80%**
4. O,K,E -> (3/3) = 1 = **100%**

**Clustering**

* Unsupervised Learning
* No Predefined Classes
* Finds “natural” grouping of instances given unlabeled data

Clustering is the process of grouping a set of objects into classes/clusters of similar objects.

Objects should be similar to one another within the same cluster and dissimilar with those in other clusters.

**Intra** **Clusters**- objects within the same cluster sharing similar properties. Distance is minimized.

**Inter** **Clusters**- two different clusters of different clusters. Distance is maximized.

A Good clustering method will produce high quality clusters which:

* The intra-class (that is, intra-cluster) similarity is high.
* The inter-class similarity is low
* The quality of a clustering result also depends on both the similarity measure used by the method and its implementation.

Where to use clustering?

**Marketing:** Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.

**Land use**: Identification of areas of similar land use in an earth observation database.

Clustering method

Algorithms also vary by:

1. Measures of Similarity
2. Linkage Methods
3. Computational Efficiency

Distance Measures

Assume a K-Dimentional Euclidean space, the distance between two points, X=[x1,x2,…,xn]

Main Categories of Clustering Methods

1. **Partitioning Algorithms**: Construct various partitions and then evaluate them by some criterion.
   1. K-Means
   2. K-Medoids
2. **Hierarchy Algorithms**: Create a hierarchical decomposition of the set of data (or objects) using some criterion.
   1. Agglomerative
   2. Divisive
3. **Density-based**: based on connectivity and density functions
4. **Grid-based**: based on a multiple-level granularity structure
5. **Model-based**: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other.

K-means Algorithm

Input: K: the number of clusters

D: the training data set

Output: A set of k Clusters

1. Select k objects as the initial clusters centroids
2. Repeat {
3. Assign all objects in D to the nearest centroids
4. Update centroid for each cluster, ie., compute the mean value of objects for each cluster

}

1. Until no change to all centroids or Maximum iteration has been reached.

Example:

Problem: Cluster the following eight points (with (x,y) representing locations) into three clusters

A(2,10) A2(2,5) A3(8,4) A4(5,8) A5(7,5) A6(6,4)

A7(1,2) A8(4,9)

Initial Cluster centers are: A1(2,10), A4(5,8) and A7(1,2).

The distance function between two points a(x1,y1) and b(x2,y2) is defined as:

P(a,b) = |x2-x1| + |y2-y1|

Use K – means algorithm to find the three clusters centers after the second iteration.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | (X2,Y2) | (2,10) | (5,8) | (5,8) | Cluster |
|  | Point (X1, Y1) | Dist. Mean 1 | Dist. Mean 2 | Dist. Mean 3 |  |
| A1 | 2,10 |  |  |  |  |
| A2 | 2,5 |  |  |  |  |
| A3 | 8,4 |  |  |  |  |
| A4 | 5,8 |  |  |  |  |
| A5 | 7,5 |  |  |  |  |
| A6 | 6,4 |  |  |  |  |
| A7 | 1,2 |  |  |  |  |
| A8 | 4,9 |  |  |  |  |

Q. Given : {2,4,10,12,3,20,30,11,25}

K=2, Randomly assign means: m1=2, m2=4

Solve the rest

|  |  |  |  |
| --- | --- | --- | --- |
|  | **C1** | **C2** |  |
|  | **2** | **4** | **Cluster** |
| **2** | 0 | 2 | C1 |
| **4** | 2 | 0 | C2 |
| **10** | 8 | 6 | C2 |
| **12** | 10 | 8 | C2 |
| **3** | 1 | 1 | C1 |
| **20** | 18 | 16 | C2 |
| **30** | 28 | 26 | C2 |
| **11** | 9 | 7 | C2 |
| **25** | 23 | 21 | C2 |

K1 = {2,3}

K2 = {4,10,12,20,30,11,25}

New C1 and C2 = avg of K1 and K2 respectively

|  |  |  |  |
| --- | --- | --- | --- |
|  | **C1** | **C2** |  |
|  | **2.5** | **16** | **Cluster** |
| **2** | 0.5 | 14 | C1 |
| **4** | 1.5 | 12 | C1 |
| **10** | 7.5 | 6 | C2 |
| **12** | 9.5 | 4 | C2 |
| **3** | 0.5 | 13 | C1 |
| **20** | 17.5 | 4 | C2 |
| **30** | 27.5 | 14 | C2 |
| **11** | 8.5 | 5 | C2 |
| **25** | 22.5 | 14 | C2 |

K1 = {2,4,3}

K2 = {10,12,20,30,11,25}

New C1 and C2 = avg of K1 and K2 respectively

|  |  |  |  |
| --- | --- | --- | --- |
|  | **C1** | **C2** |  |
|  | **3** | **18** | **Cluster** |
| **2** | 1 | 16 | C1 |
| **4** | 1 | 14 | C1 |
| **10** | 7 | 8 | C1 |
| **12** | 8 | 6 | C2 |
| **3** | 0 | 15 | C1 |
| **20** | 17 | 2 | C2 |
| **30** | 27 | 12 | C2 |
| **11** | 8 | 7 | C2 |
| **25** | 22 | 7 | C2 |

K1 = {2,4,10,3}

K2 = {12,20,30,11,25}

New C1 and C2 = avg of K1 and K2 respectively

|  |  |  |  |
| --- | --- | --- | --- |
|  | **C1** | **C2** |  |
|  | **4.75** | **19.6** | **Cluster** |
| **2** | 2.75 | 17.6 | C1 |
| **4** | 0.75 | 15.6 | C1 |
| **10** | 5.25 | 9.6 | C1 |
| **12** | 7.25 | 7.6 | C1 |
| **3** | 1.75 | 16.6 | C1 |
| **20** | 15.25 | 0.4 | C2 |
| **30** | 25.25 | 10.4 | C2 |
| **11** | 6.25 | 8.6 | C1 |
| **25** | 20.25 | 5.4 | C2 |

K1={2,4,10,12,3,11)

K2={20,30,25}

|  |  |  |  |
| --- | --- | --- | --- |
|  | **C1** | **C2** |  |
|  | **7** | **25** | **Cluster** |
| **2** | 5 | 23 | C1 |
| **4** | 3 | 21 | C1 |
| **10** | 3 | 15 | C1 |
| **12** | 5 | 13 | C1 |
| **3** | 4 | 22 | C1 |
| **20** | 13 | 5 | C2 |
| **30** | 23 | 5 | C2 |
| **11** | 4 | 14 | C1 |
| **25** | 18 | 0 | C2 |

K1={2,4,10,12,3,11)

K2={20,30,25}

**Advantages** of the Simple K-Means Algorithm:

* It is easy to implement and work with any of the standard norms
* It allows straightforward parallelization
* It is insensitive with respect to data ordering

**Disadvantages** with K-Means:

* The results strongly depend on the initial guess of centroids
* A local optimum (computed for a cluster) does not need to be a global optimum (overall clustering of a data set)
* It is not obvious what a good number K is in each case
* The process is sensitive with respect to outliers
* Resulting clusters can be unbalanced (even empty!!; cf. Forgy, 1965)

**Hierarchical Clustering**

* A Hierarchical clustering method works by grouping objects into a tree of clusters.
* Hierarchical clustering methods can be further classified as :
  + Agglomerative
    - It merges clusters iteratively
    - Bottom-up manner
  + Divisive
    - Starting with all objects in one cluster and subdividing them into smaller pieces.
    - Top-down manner.

Distance Measurement

I **single-linkage clustering** (also called the connectedness or minimum method), consider the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster.

In **complete-linkage clustering** (also called the diameter or maximum method), consider the distance between one cluster and another cluster to be equal to the greatest distance from any member of one cluster to any member of the other cluster.

In **average-linkage clustering,** consider the distance between one luster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster.

Q. Construct agglomerative clustering hierarchies:

Consider a set of six objects: A, B, C, D, E, F. Let the following be a dissimilarity matrix between these objects.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** | **F** |
| **A** | 0.0 | 1.0 | 5.0 | 9.0 | 10.0 | 2.0 |
| **B** | 1.0 | 0.0 | 3.5 | 8.0 | 7.0 | 5.5 |
| **C** | 5.0 | 3.5 | 0.0 | 3.0 | 4.0 | 6.5 |
| **D** | 9.0 | 8.0 | 3.0 | 0.0 | 0.5 | 4.5 |
| **E** | 10.0 | 7.0 | 4.0 | 0.5 | 0.0 | 2.5 |
| F | 2.0 | 5.5 | 6.5 | 4.5 | 2.5 | 0.0 |

**Single Linkage Clustering**

* Find the minimum distance and consider their main objects. (in Our Case: D, E (0.5))
* Compare Col and Row values of D and E and put the number which is lower. We will get:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D E** | **F** |
| **A** | 0.0 | 1.0 | 5.0 | 9.0 | 2.0 |
| **B** | 1.0 | 0.0 | 3.5 | 7.0 | 5.5 |
| **C** | 5.0 | 3.5 | 0.0 | 3.0 | 6.5 |
| **DE** | 9.0 | 7.0 | 3.0 | 0.0 | 2.5 |
| **F** | 2.0 | 5.5 | 6.5 | 2.5 | 0.0 |

* Again find the minimum distance (other than 0.0) (in this case A and B). Repeat the same step until we get only one row and column

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **AB** | **C** | **D E** | **F** |
| **AB** | 0.0 | 3.5 | 7.0 | 2.0 |
| **C** | 3.5 | 0.0 | 3.0 | 6.5 |
| **DE** | 7.0 | 3.0 | 0.0 | 2.5 |
| **F** | 2.0 | 6.5 | 2.5 | 0.0 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **ABF** | **C** | **D E** |
| **ABF** | 0.0 | 3.5 | 2.5 |
| **C** | 3.5 | 0.0 | 3.0 |
| **DE** | 2.5 | 3.0 | 0.0 |

|  |  |  |
| --- | --- | --- |
|  | **ABFDE** | **C** |
| **ABFDE** | 0.0 | 3.0 |
| **C** | 3.0 | 0.0 |

Q. Hierarchical Clustering of Distances in miles between U.S. Cities.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** |
| **A** | 0 | 206 | 129 | 1504 | 963 | 2976 | 3095 | 2979 | 1949 |
| **B** | 206 | 0 | 233 | 1308 | 802 | 2815 | 2934 | 2786 | 1771 |
| **C** | 129 | 233 | 0 | 1075 | 671 | 2684 | 2799 | 2631 | 1616 |
| **D** | 1504 | 1308 | 1075 | 0 | 1329 | 3273 | 3053 | 2687 | 2037 |
| **E** | 963 | 802 | 671 | 1329 | 0 | 2013 | 2142 | 2054 | 996 |
| **F** | 2976 | 2815 | 2684 | 3273 | 2013 | 0 | 808 | 1131 | 1307 |
| **G** | 3095 | 2934 | 2799 | 3053 | 2142 | 808 | 0 | 379 | 1235 |
| **H** | 2979 | 2786 | 2631 | 2687 | 2054 | 1131 | 379 | 0 | 1059 |
| **I** | 1949 | 1771 | 1616 | 2037 | 996 | 1307 | 1235 | 1059 | 0 |

Step 1:

Merging A and B

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **AB** | **C** | **D** | **E** | **F** | **G** | **H** | **I** |
| **AB** | 0 | 233 | 1308 | 802 | 2815 | 2934 | 2786 | 1771 |
| **C** | 233 | 0 | 1075 | 671 | 2684 | 2799 | 2631 | 1616 |
| **D** | 1308 | 1075 | 0 | 1329 | 3273 | 3053 | 2687 | 2037 |
| **E** | 802 | 671 | 1329 | 0 | 2013 | 2142 | 2054 | 996 |
| **F** | 2815 | 2684 | 3273 | 2013 | 0 | 808 | 1131 | 1307 |
| **G** | 2934 | 2799 | 3053 | 2142 | 808 | 0 | 379 | 1235 |
| **H** | 2786 | 2631 | 2687 | 2054 | 1131 | 379 | 0 | 1059 |
| **I** | 1771 | 1616 | 2037 | 996 | 1307 | 1235 | 1059 | 0 |

Merging AB and C

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **ABC** | **D** | **E** | **F** | **G** | **H** | **I** |
| **ABC** | 0 | 1075 | 671 | 2684 | 2799 | 2631 | 1616 |
| **D** | 1075 | 0 | 1329 | 3273 | 3053 | 2687 | 2037 |
| **E** | 671 | 1329 | 0 | 2013 | 2142 | 2054 | 996 |
| **F** | 2684 | 3273 | 2013 | 0 | 808 | 1131 | 1307 |
| **G** | 2799 | 3053 | 2142 | 808 | 0 | 379 | 1235 |
| **H** | 2631 | 2687 | 2054 | 1131 | 379 | 0 | 1059 |
| **I** | 1616 | 2037 | 996 | 1307 | 1235 | 1059 | 0 |

Merging G and H

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **ABC** | **D** | **E** | **F** | **GH** | **I** |
| **ABC** | 0 | 1075 | 671 | 2684 | 2631 | 1616 |
| **D** | 1075 | 0 | 1329 | 3273 | 2687 | 2037 |
| **E** | 671 | 1329 | 0 | 2013 | 2054 | 996 |
| **F** | 2684 | 3273 | 2013 | 0 | 808 | 1307 |
| **GH** | 2631 | 2687 | 2054 | 808 | 0 | 1059 |
| **I** | 1616 | 2037 | 996 | 1307 | 1059 | 0 |

Merging ABC and E

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **ABCE** | **D** | **F** | **GH** | **I** |
| **ABCE** | 0 | 1075 | 2013 | 2054 | 996 |
| **D** | 1075 | 0 | 3273 | 2687 | 2037 |
| **F** | 2013 | 3273 | 0 | 808 | 1307 |
| **GH** | 2054 | 2687 | 808 | 0 | 1059 |
| **I** | 996 | 2037 | 1307 | 1059 | 0 |

Merging GH and F

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **ABCE** | **D** | **GHF** | **I** |
| **ABCE** | 0 | 1075 | 2013 | 996 |
| **D** | 1075 | 0 | 2687 | 2037 |
| **GHF** | 2013 | 2687 | 0 | 1059 |
| **I** | 996 | 2037 | 1059 | 0 |

Merging ABCE and I

|  |  |  |  |
| --- | --- | --- | --- |
|  | **ABCEI** | **D** | **GHF** |
| **ABCEI** | 0 | 1075 | 1059 |
| **D** | 1075 | 0 | 2687 |
| **GHF** | 1059 | 2687 | 0 |

Merging ABCEI and GHF

|  |  |  |
| --- | --- | --- |
|  | **ABCEIGHF** | **D** |
| **ABCEIGHF** | 0 | 1075 |
| **D** | 1075 | 0 |

Merging ABCEIGHF and D

{ ABCEIGHFD }

**What are Outliers?**

* Outlier: A data object that deviates significantly from normal objects as if it were generated by a different mechanism.
* Outliers are different from the noise data
  1. Noise is random error or variance in a measured variable
  2. Noise should be removed before outlier detection
* Outliers are interesting: it violates the mechanism that generates the normal data
* Outlier detection vs novelty detection: early stage, outlier; but later merged into the model
* Applications:
  1. Credit card fraud detection
  2. Telecom fraud detection
  3. Customer segmentation
  4. Medical analysis

**Types of Outliers**

There are three kinds of Outliers: Global, Contextual and Collective outliers

* Global Outlier (or point anomaly)
  + Objects in O if it significantly deviates from the rest of the data set
  + Ex. Intrusion detection in computer network
  + Issue: Find an appropriate measurement of deviation
* Contextual Outlier (or conditional outlier)
  + Object is O if it deviates significantly based on a selected context
  + Temp 80f: outlier? (depending on summer or winter?)
  + Attributes of data objects should be divided into two groups
    - Contextual Attribute: defines the context, eg: time and location
    - Behavioral attributes: characteristics of the object, used in outlier evaluation eg: temperature
* Collective Outliers
  + A subset of data object collectively deviates significantly from the whole data set, even if the individual data objects many not be outliers
  + Applications: eg: intrusion detection
    - When a number of computers keep sending denial-of-service packages to each other
  + Detection of collective outliers
    - Consider not only behavior of individual objects, but also that of groups of objects
    - Need to have the background knowledge on the relationship among data objects, such as a distance or similarity measure on objects
  + A data set may have multiple types of outliers
  + One object may belong to more than one type of outlier

Challenges of Outlier Detection

* Modeling normal objects and outliers properly
  + Hard to enumerate all possible normal behaviors in an application
  + The border between normal and outlier objects is often a gray area
* Application-specific outlier detection
  + Choice of distance measure among objects and the model of relationship among objects are often application-dependent
  + Eg: clinic data: a small deviation could be an outlier; while in marketing analysis, larger fluctuations
* Handling noise in outlier detection
  + Noise may distort the normal object and blur the distinction between normal objects and outliers. It may help hide outliers and reduce the effectiveness of outlier detection
* Understandability
  + Understand why these are outliers: justification of the detection
  + Specify the degree of an outlier: the unlikelihood of the object being generated by a normal mechanism

Outlier Detection 1: Supervised Methods

* Two ways to categorize outlier detection methods:
  + Based on whether user-*labeled* example of outliers can be obtained:
    - Supervised, semi-supervised VS unsupervised methods
  + Based on assumption about normal data and outliers:
    - Statistical, proximity-based, and clustering-based methods
* Outlier Detection 1: Supervised Methods (data is labeled, training data is given)
  + Modeling outlier detection as a classification problem
    - Samples examined by domain experts used for training and testing
  + Methods for learning a classifier for outlier detection effectively:
    - Model normal objects and report those not matching the model as outliers, or
    - Model outliers and treat those not matching the model as normal
  + Challenges
    - Imbalanced classes, i.e., outliers are rare: Boost the outlier class and make up some artificial outliers
    - Catch as many outliers as possible, i.e., recall is more important than accuracy (i.e., not mislabeling normal objects as outliers)

Outlier Detection II: Unsupervised Methods

* Assume the normal objects are somewhat “clustered” into multiple groups, each having some distinct features
* An outlier is expected to be far away from any groups of normal objects
* Weakness: Cannot detect collective outlier effectively
  + Normal objects may not share any strong patterns, but the collective outliers may share high similarity in a small area
* Ex: in some intrusions or virus detection, normal activities are diverse
  + Unsupervised methods may have a high false positive rate but still miss many real outliers.
  + Supervised methods can be more effective, eg: identify attacking some key resources
* Many Clustering methods can be adapted for unsupervised methods
  + Find clusters, then outliers: not belonging to any cluster
  + Problem 1: hard to distinguish noise from outliers
  + Problem 2: costly since first clustering: but far less outliers than normal objects
    - Newer methods: tackle outliers directly

Outlier Detection III: Semi-supervised methods

* Situation: in many applications, the number of labeled data is often small: Labels could be on outliers only, normal objects only, or both
* Semi-supervised outlier detection: Regarded as applications of semi-supervised learning
* If some labeled normal objects are available
  + Use the labeled examples and the proximate unlabeled objects to train a model for normal objects
  + Those not fitting the model of normal objects are detected as outliers
* If only some labeled outliers are available, a small number of labeled outliers may not cover the possible outliers well
  + To improve the quality of outlier detection, one can get help from models from normal objects learned from unsupervised methods

Outlier detection: Statistical Methods

* Statistical methods (also known as model-based methods) assume that thenormal data follow some statistical model (a stochastic model)

Outlier detection: Proximity Methods

* An object in an outlier if the nearest neighbors of the object are far away, i.e., the proximity of the object is significantly deviates from the normal range of objects.

Q. Write a short note on outlier.

Q. what do you mean by outliers.

Q. Explain different types of outliers.

Q. Explain supervised, unsupervised and semi-supervised approach for outlier detection.

Q. Give different examples (applications of the outliers).

**Big Data**

Big Data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing application.

Big data spans three dimensions: Volume, Velocity and Variety

* Volume: Enterprise are awash with ever-growing data of all types, easily amassing terabytes – even petabytes of information
  + Turn 12 terabytes of tweets created each day into improved product sentiment analysis
  + Convert 350 billion annual meter readings to better predict power consumption
* Velocity: Sometimes 2 minutes is too late. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.
  + Scrutinize 5 million trade events created each day to identify potential fraud
  + Analyze 500 million daily call details records in real-time to predict customer churn faster
* Varienty: Variety refers to the many sources and types of data both structured and unstructured. We used to store data from sources like spreadsheets and databases.

Q. What is web mining?

Q. What is Hadoop?

Q. What is BigData and 3V’s of Big Data?

Precision: The closeness of repeated measurements to one another. It is calculated as the ratio of the number of relevant records retrived to the total number of irrelevant and relevant records retrived.

**Precision= True Positive/(True Positive + False Positive)**

**Recall** is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

**F-measure:** measure of a test’s accuracy. Scores reaches its best value at 1 and worst scores at 0.

F = 2 x (Precision x recall / precision + recall)

Precision can be seen as a measure of exactness or quality,

Recall is a measure of completeness or quantity

ROC (Receiver Operating Characteristics) curves: for visual comparison of classification models

* Shows the trade-off between the true positive rate and the false positive rate
* The area under the ROC curves is a measure of the accuracy of the model.

Implementing Closedown

1. Getting delivery acceptance from the customer.
2. Shutting down resources and releasing to new uses.
3. Reassigning project team member.
4. Closing accounts and paying all bills.