

BUSINESS INTELLIGENCE



LECTURE 3

OUTLINE

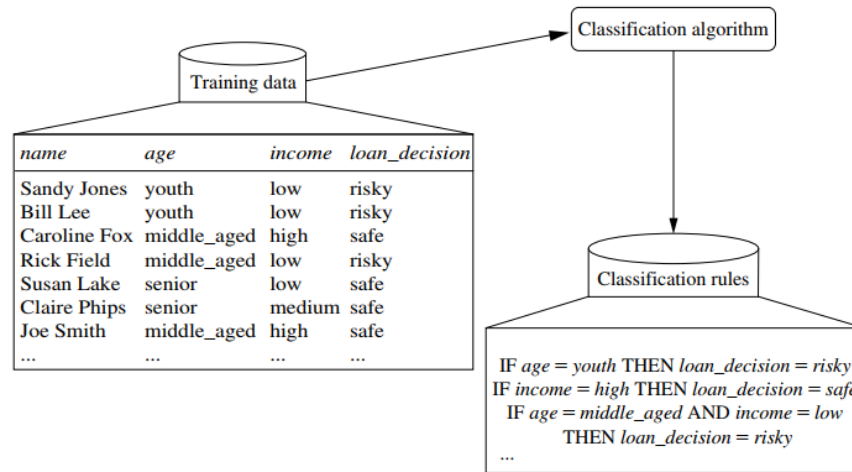
- Introduction
- Data mining overview
- Data pre-processing
- **Classification techniques**
- **Clustering techniques**
- **Association rules**
- **Prediction**
- References

CLASSIFICATION TECHNIQUES

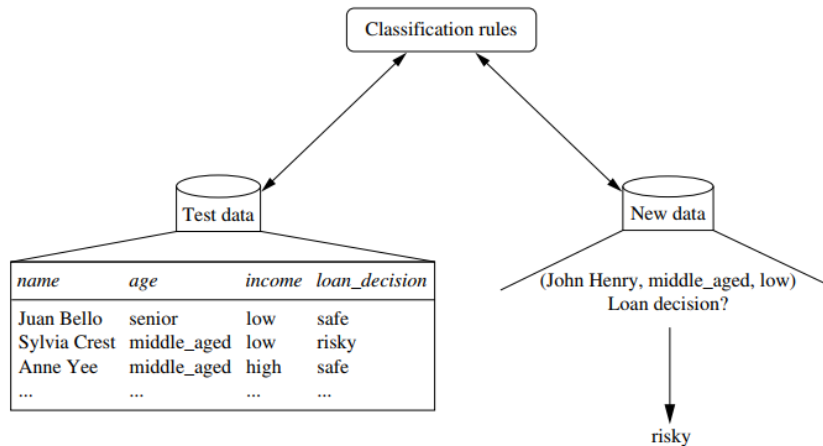
CLASSIFICATION (1 / 2)

- Data classification is a two-step process:
 - A learning step where a classification model is constructed
 - A classification step where the model is used to predict class labels for given data
- E.g.,
 - Loan applicants are safe or risky
 - A customer profile to buy a computer
 - One of the treatment a patient should receive

CLASSIFICATION (2/2)



(a)



(b)

DECISION TREES (1/3)

- ID3
- C4.5
- CART

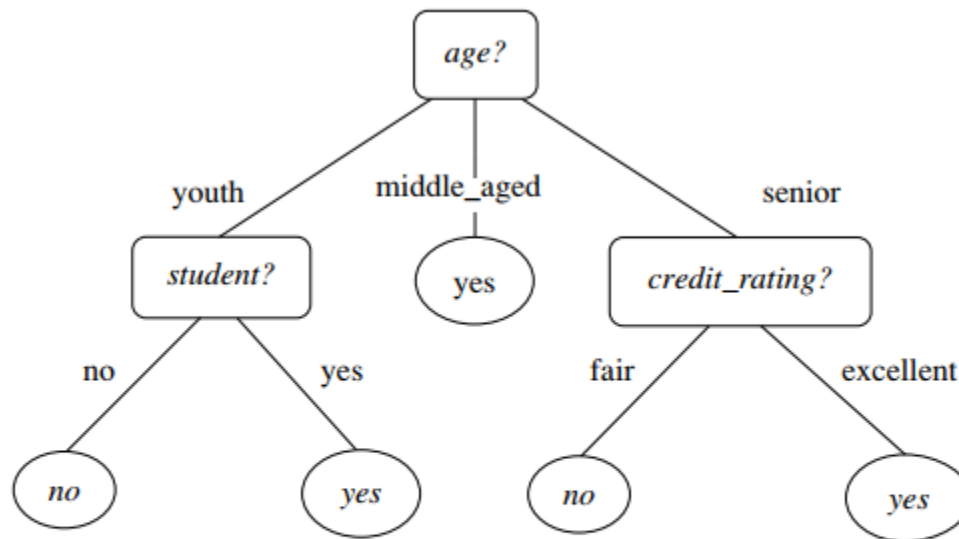


Figure 8.2 A decision tree for the concept *buys_computer*, indicating whether an *AllElectronics* customer is likely to purchase a computer. Each internal (nonleaf) node represents a test on an attribute. Each leaf node represents a class (either *buys_computer* = *yes* or *buys_computer* = *no*).

DECISION TREES (2/3)

The expected information needed to classify a tuple in D is given by

■ E.g.,

$$Info(D) = - \sum_{i=1}^m p_i \log_2(p_i), \quad (8.1)$$

$$Info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times Info(D_j). \quad (8.2)$$

Table 8.1 Class-Labeled Training Tuples from the *AlIElectronics* Customer Database

<i>RID</i>	<i>age</i>	<i>income</i>	<i>student</i>	<i>credit_rating</i>	<i>Class: buys_computer</i>
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

DECISION TREES (3/3)

- E.g., $Gain(age) = Info(D) - Info_{age}(D) = 0.940 - 0.694 = 0.246$ bits.

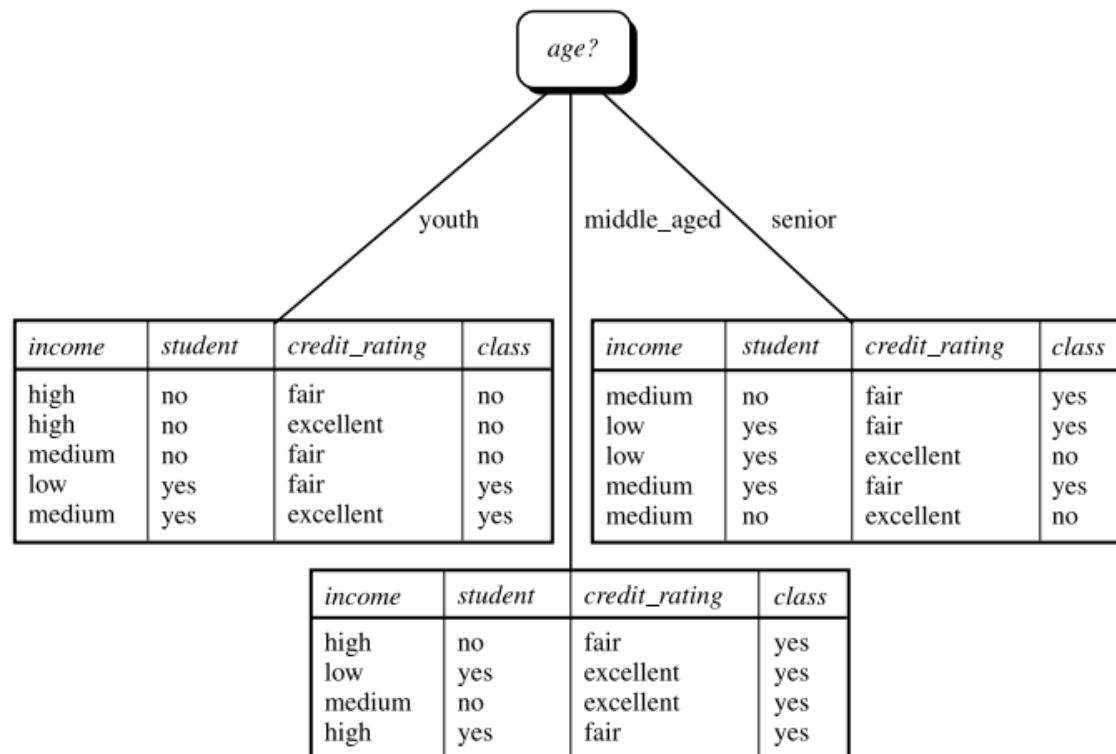


Figure 8.5 The attribute *age* has the highest information gain and therefore becomes the splitting attribute at the root node of the decision tree. Branches are grown for each outcome of *age*. The tuples are shown partitioned accordingly.

FOR EXAMPLE

$$Info(D) = -\frac{9}{14} \log_2 \left(\frac{9}{14} \right) - \frac{5}{14} \log_2 \left(\frac{5}{14} \right) = 0.940 \text{ bits.}$$

$$\begin{aligned} Info_{age}(D) &= \frac{5}{14} \times \left(-\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \right) \\ &\quad + \frac{4}{14} \times \left(-\frac{4}{4} \log_2 \frac{4}{4} \right) \\ &\quad + \frac{5}{14} \times \left(-\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) \\ &= 0.694 \text{ bits.} \end{aligned}$$

$$Gain(age) = Info(D) - Info_{age}(D) = 0.940 - 0.694 = 0.246 \text{ bits.}$$

$$Gain(income) = 0.029 \text{ bits,} \quad Gain(student) = 0.151 \text{ bits} \quad Gain(credit_rating) = 0.048 \text{ bits}$$

BAYES CLASSIFICATION METHODS (1/2)

- Bayesian classifiers are statistical classifiers. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}.$$

To predict the class label of X , $P(X|C_i)P(C_i)$ is evaluated for each class C_i . The classifier predicts that the class label of tuple X is the class C_i if and only if

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j) \quad \text{for } 1 \leq j \leq m, j \neq i. \quad (8.15)$$

BAYES CLASSIFICATION METHODS (2/2)

$X = (\text{age} = \text{youth}, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit_rating} = \text{fair})$

- $P(\text{buys_computer} = \text{yes} | X) = ?$
- $P(\text{buys_computer} = \text{no} | X) = ?$

Table 8.1 Class-Labeled Training Tuples from the *AllElectronics* Customer Database

<i>RID</i>	<i>age</i>	<i>income</i>	<i>student</i>	<i>credit_rating</i>	<i>Class: buys_computer</i>
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

FOR INSTANCE

$$P(\text{buys_computer} = \text{yes}) = 9/14 = 0.643$$

$$P(\text{buys_computer} = \text{no}) = 5/14 = 0.357$$

To compute $P(\mathbf{X}|C_i)$, for $i = 1, 2$, we compute the following conditional probabilities:

$$P(\text{age} = \text{youth} | \text{buys_computer} = \text{yes}) = 2/9 = 0.222$$

$$P(\text{age} = \text{youth} | \text{buys_computer} = \text{no}) = 3/5 = 0.600$$

$$P(\text{income} = \text{medium} | \text{buys_computer} = \text{yes}) = 4/9 = 0.444$$

$$P(\text{income} = \text{medium} | \text{buys_computer} = \text{no}) = 2/5 = 0.400$$

$$P(\text{student} = \text{yes} | \text{buys_computer} = \text{yes}) = 6/9 = 0.667$$

$$P(\text{student} = \text{yes} | \text{buys_computer} = \text{no}) = 1/5 = 0.200$$

$$P(\text{credit_rating} = \text{fair} | \text{buys_computer} = \text{yes}) = 6/9 = 0.667$$

$$P(\text{credit_rating} = \text{fair} | \text{buys_computer} = \text{no}) = 2/5 = 0.400$$

Using these probabilities, we obtain

$$\begin{aligned} P(\mathbf{X} | \text{buys_computer} = \text{yes}) &= P(\text{age} = \text{youth} | \text{buys_computer} = \text{yes}) \\ &\quad \times P(\text{income} = \text{medium} | \text{buys_computer} = \text{yes}) \\ &\quad \times P(\text{student} = \text{yes} | \text{buys_computer} = \text{yes}) \\ &\quad \times P(\text{credit_rating} = \text{fair} | \text{buys_computer} = \text{yes}) \\ &= 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044. \end{aligned}$$

Similarly,

$$P(\mathbf{X} | \text{buys_computer} = \text{no}) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019.$$

To find the class, C_i , that maximizes $P(\mathbf{X}|C_i)P(C_i)$, we compute

$$P(\mathbf{X} | \text{buys_computer} = \text{yes})P(\text{buys_computer} = \text{yes}) = 0.044 \times 0.643 = 0.028$$

$$P(\mathbf{X} | \text{buys_computer} = \text{no})P(\text{buys_computer} = \text{no}) = 0.019 \times 0.357 = 0.007$$

Therefore, the naïve Bayesian classifier predicts $\text{buys_computer} = \text{yes}$ for tuple \mathbf{X} . ■

EVALUATION METRICS (1 / 2)

- **True positives (TP):** These refer to the positive tuples that were correctly labeled by the classifier. Let TP be the number of true positives.
- **True negatives (TN):** These are the negative tuples that were correctly labeled by the classifier. Let TN be the number of true negatives.
- **False positives (FP):** These are the negative tuples that were incorrectly labeled as positive. Let FP be the number of false positives.
- **False negatives (FN):** These are the positive tuples that were mislabeled as negative. Let FN be the number of false negatives.

EVALUATION METRICS (2/2)

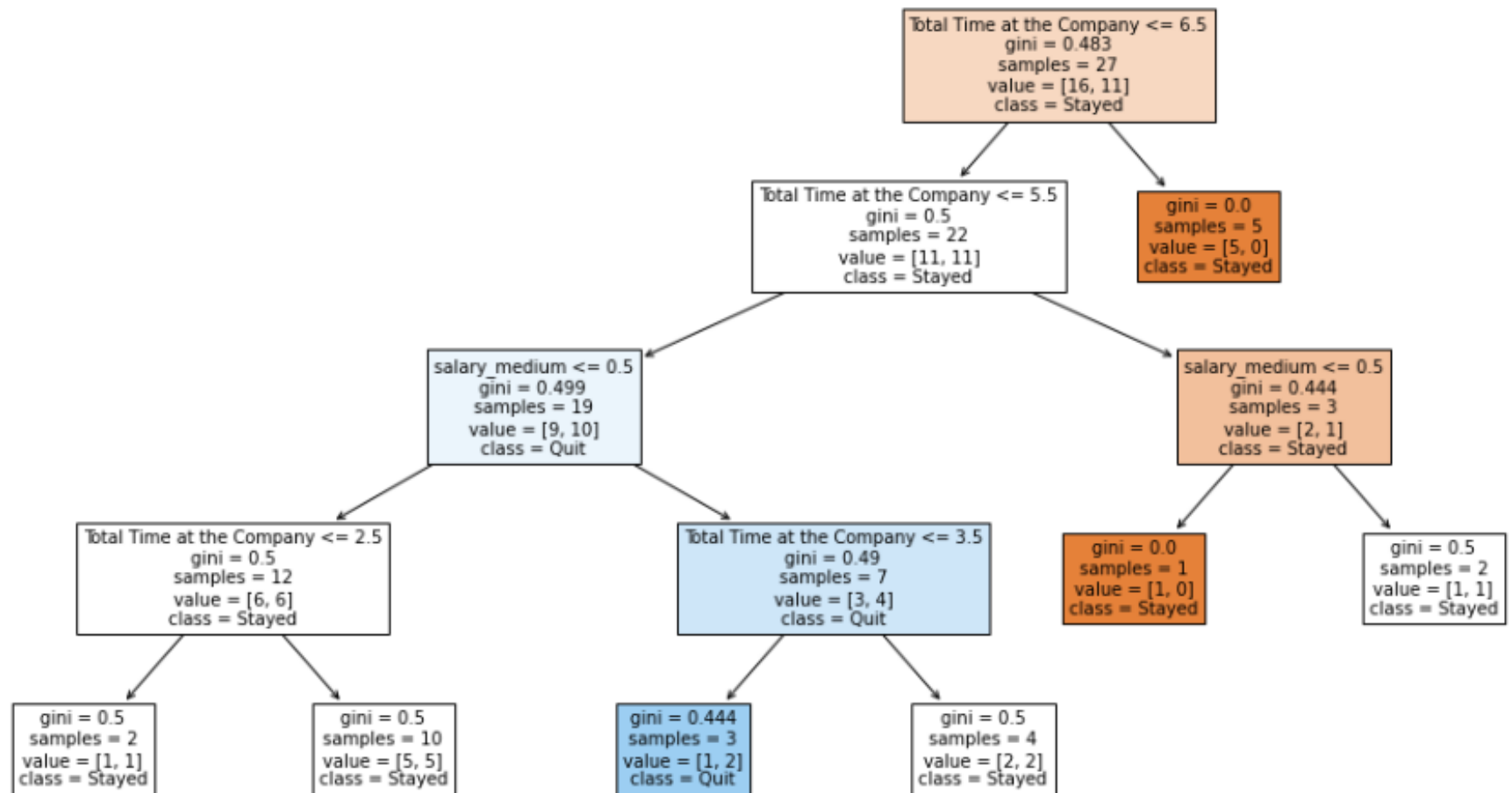
■ Confusion matrix

Actual class	Predicted class		Total
	<i>yes</i>	<i>no</i>	
	<i>yes</i>	<i>no</i>	
<i>yes</i>	<i>TP</i>	<i>FN</i>	<i>P</i>
<i>no</i>	<i>FP</i>	<i>TN</i>	<i>N</i>
Total	<i>P'</i>	<i>N'</i>	<i>P + N</i>

Measure	Formula
accuracy, recognition rate	$\frac{TP + TN}{P + N}$
error rate, misclassification rate	$\frac{FP + FN}{P + N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP + FP}$
F , F_1 , F -score, harmonic mean of precision and recall	$\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$
F_β , where β is a non-negative real number	$\frac{(1 + \beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}$

Figure 8.13 Evaluation measures. Note that some measures are known by more than one name. TP , TN , FP , P , N refer to the number of true positive, true negative, false positive, positive, and negative samples, respectively (see text).

COMPANY CHURN DEMO



CLUSTERING TECHNIQUES

CLUSTERING

- Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters.
- E.g.,
 - Customer segmentation
 - Handwritten character recognition
 - Web search results
 - Outlier analysis

PARTITIONING METHODS

- K-means
- K-modes
- K-medoids

K-MEANS EXAMPLE

- $S = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$
- $K = 2$
- $C1 = \{2, 3, 4, 10, 11, 12\}$
- $C2 = \{20, 25, 30\}$

- Randomly take 2 means as centroids
 - $m1 = 4$
 - $m2 = 12$

HIERARCHICAL METHODS

- Agglomerative
- Divisive

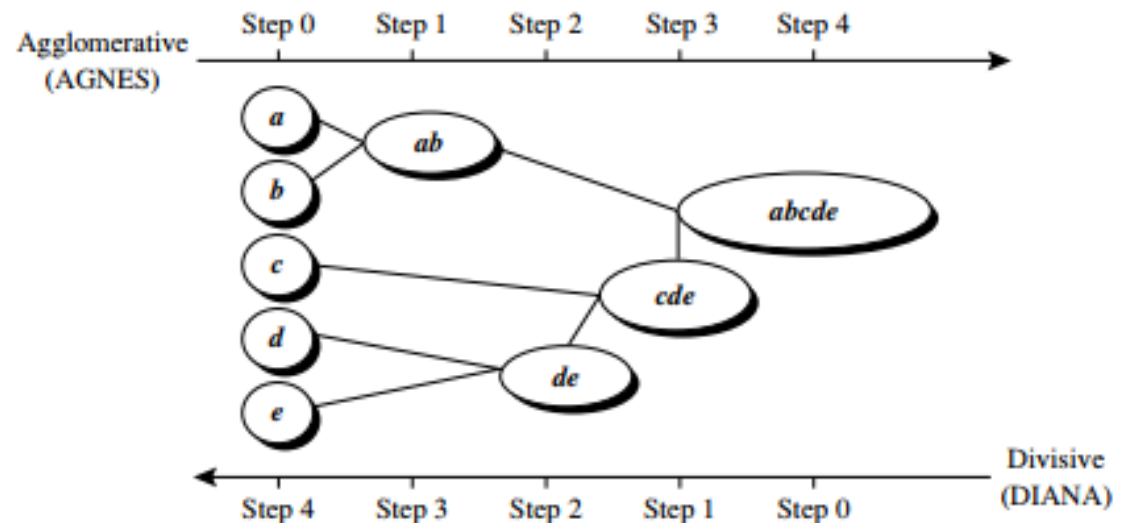
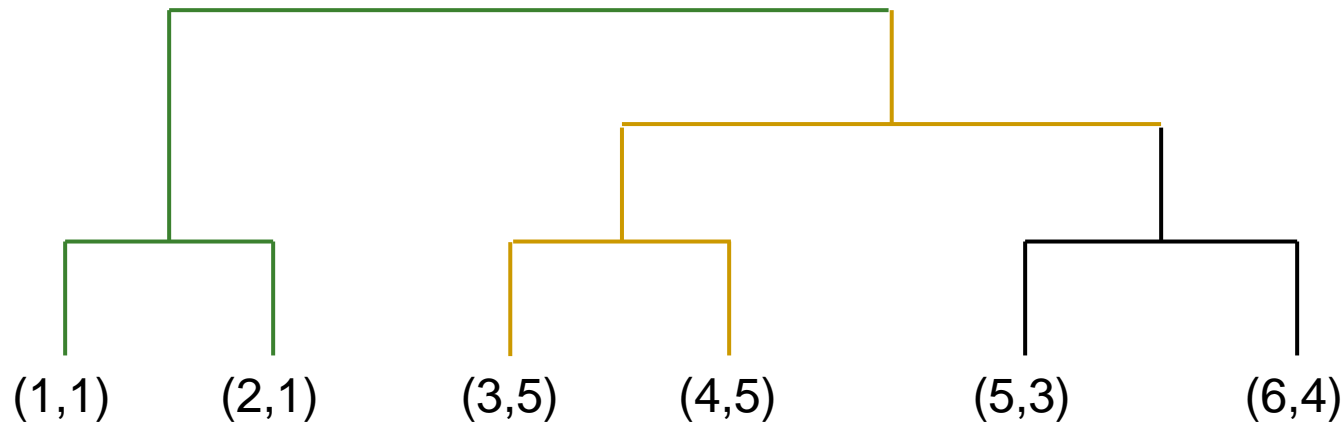


Figure 10.6 Agglomerative and divisive hierarchical clustering on data objects $\{a, b, c, d, e\}$.

AGGLOMERATIVE HIERARCHICAL METHOD EXAMPLE

■ $S = \{(1,1), (2,1), (3,5), (4,5), (5,3), (6,4)\}$



DENSITY-BASED METHODS

- We can model clusters as dense regions in the data space, separated by sparse regions, which can discover clusters of nonspherical shape.
- DBSCAN
 - Radius about each point (eps)
 - The minimum number of data points that should be around that point within that radius (MinPts)
 - E.g., (1.5, 2.5) with $\text{eps} = 0.3$, then the circle around the point with radius = 0.3, will contain only one other point inside it (1.2, 2.5)
- OPTICS
- DENCLUE

DBSCAN EXAMPLE

- $\text{eps} = 0.6$ and $\text{MinPts} = 4$
- The first data point (1,2)
- Cluster 1
 - (3,4), (2.5,4), (3,5), (2.8,4.5), (2.5,4.5)
- Cluster 2
 - (1,2), (1.5,2.5), (1.2,2.5), (1,3), (1,2.5)
- Outliers
 - (1,5), (5,6), (4,3)
- Example

x	y	d from (1,2)
1	2	0
3	4	2.8
2.5	4	2.5
1.5	2.5	0.7
3	5	3.6
2.8	4.5	3.08
2.5	4.5	2.9
1.2	2.5	0.53
1	3	1
1	5	3
1	2.5	0.5
5	6	5.6
4	3	3.1

GRID-BASED METHODS

- A grid-based clustering method takes a space-driven approach by partitioning the embedding space into cells independent of the distribution of the input objects.
- STING
- CLIQUE

EVALUATION METRICS

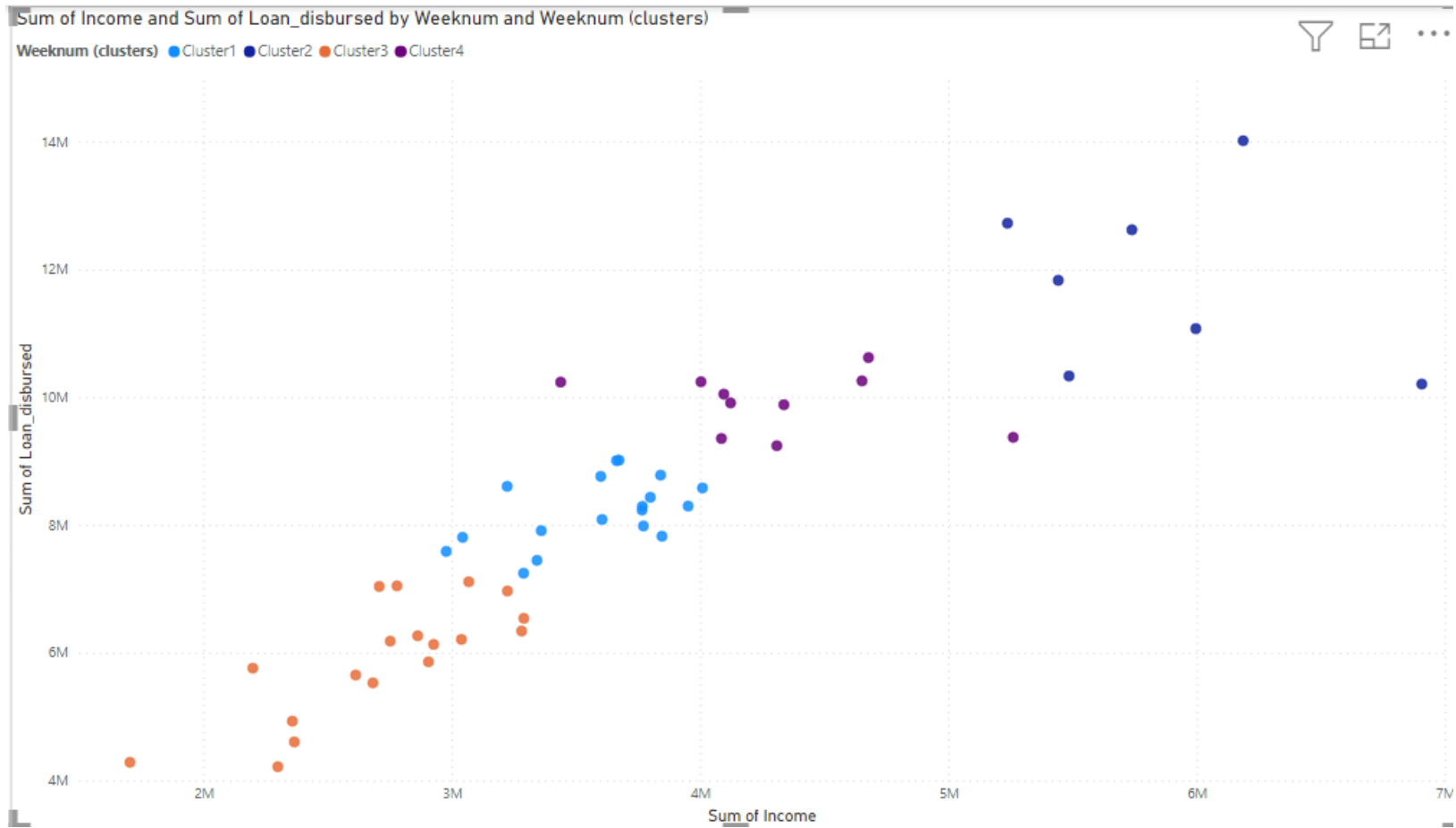
- Assessing clustering tendency so that nonrandom structure exists.
 - Hopkins statistic
- Determining the number of clusters in a data set.
 - The elbow method
- Measuring clustering quality.
 - Extrinsic methods
 - Intrinsic methods

$$WCSS = \sum_{C_k}^{C_n} \left(\sum_{d_i \in C_k}^{d_m} distance(d_i, C_k)^2 \right)$$

Where,

C is the cluster centroids and *d* is the data point in each Cluster.

BANK LOAN DISBURSAL CLUSTERING DEMO



ASSOCIATION RULES

ASSOCIATION RULES

- Frequent patterns and association rules are helpful for some scenario such as recommendation.
- Which patterns are interesting
 - support
 - confidence
 - lift
- Apriori algorithm
- FP-growth

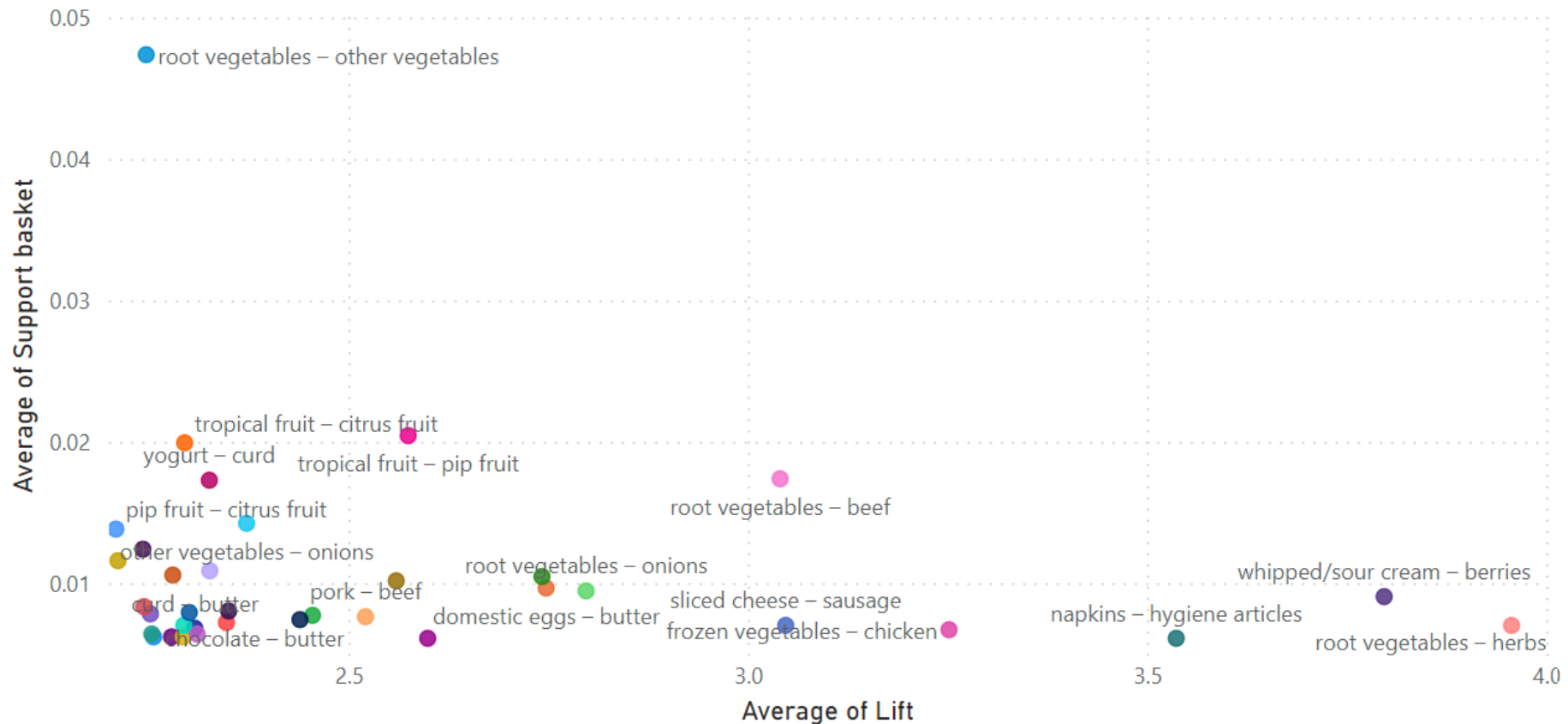
$$\text{Support} = \frac{\text{Number of transactions including one or multiple products}}{\text{Total number of transactions}}$$

$$\text{Confidence of product one} \rightarrow \text{Basket} = \frac{\text{Support of basket}}{\text{Support of product one}}$$

$$\text{Confidence of product two} \rightarrow \text{Basket} = \frac{\text{Support of basket}}{\text{Support of product two}}$$

$$\text{Lift} = \frac{\text{Support of basket}}{(\text{Support of product one} * \text{Support of product two})}$$

BASKET ANALYSIS DEMO



PREDICTION

LINEAR REGRESSION (1/2)

■ Simple linear regression

□ $Y = a + bX$

- X: independent variable
- Y: outcome variable
- a: Y-intercept
- b: slope of the line

$$a = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum x^2) - (\sum x)^2}$$

■ For instance

□ $\text{Weight} = 80 + 2(\text{Height})$

FOR INSTANCE

- Find a linear regression equation when given the dataset below:

x	y
2	3
4	7
6	5
8	10

LINEAR REGRESSION (2/2)

■ Multiple linear regression

□ $Y = a + b_1X_1 + b_2X_2 + \dots + b_nX_n + e$

■ $X_1 \dots X_n$: independent variables

■ Y : outcome variable

■ a : Y-intercept

■ b : slope of the line

■ e : residuals (error)

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)}$$

$$a = b_0 = Y - b_1 X_1 - b_2 X_2$$

■ For instance

□ $\text{BMI} = 18.0 + 1.5 (\text{diet score}) + 1.6 (\text{male}) + 4.2 (\text{age} > 20)$

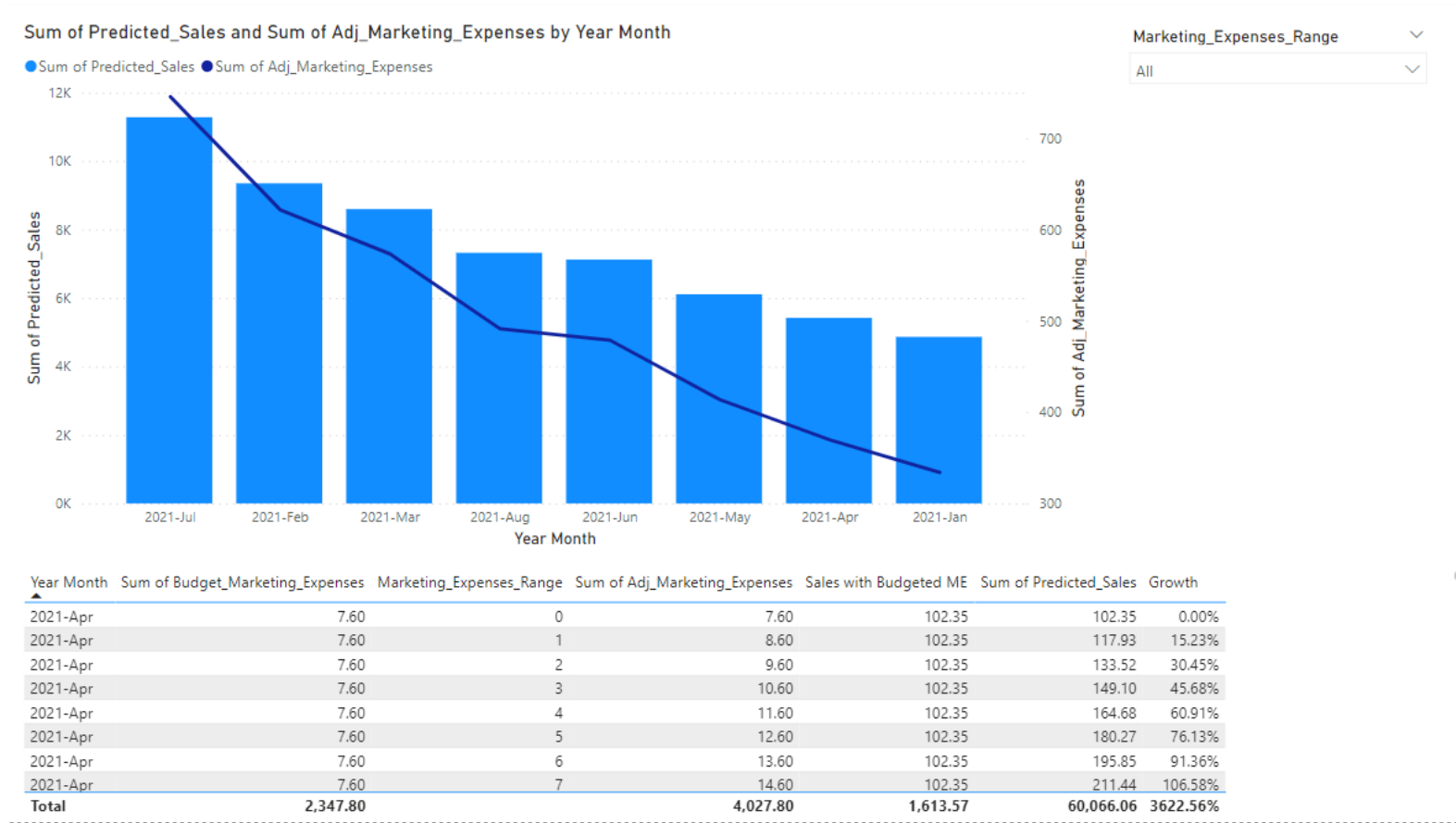
FOR INSTANCE

- Find a linear regression equation when given the dataset below:

Car	Price (thousand dollars)	Age (years)	Mileage (thousand miles)
1	29	1	18
2	25	2	25
3	21	2	50
4	18	3	68
5	15	4	75
6	15	5	65

$$\text{Price} = 32.46 - 1.54(\text{Age}) - 0.15(\text{Mileage})$$

SALES AND MARKETING EXPENSES DEMO



QUESTIONS AND ANSWERS



Picture from: <http://philadelphiaculpturegym.blogspot.com/2013/09/save-date-free-talk-and-q-on-affordable.html>

REFERENCES

- [1] Tobias Zwingmann, AI-Powered Business Intelligence, Kindle Edition, O'reilly Press, 2022
- [2] Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Techniques”, Third Edition, Morgan Kaufmann Publishers, 2012.
- [3] Jeen Su Lim, John Heinrichs, “Digital Business Intelligence Management with Big Data Analytics” Kindle Edition, O'reilly Press, 2021.
- [4] David L. Olson, Dursun Delen, “Advanced Data Mining Techniques”, Springer-Verlag, 2008.
- [5] Brian Larson, “Delivering Business Intelligence with Microsoft SQL Server 2016”, McGraw-Hill Education; 4 edition, 2016.
- [6] Oracle, “Data Mining Concepts”, 18c, E83730-03, 2018
- [7] Oracle, “Data Mining Application Developer’s Guide”, 2013.