BUSINESS INTELLIGENCE

LECTURE 3

OUTLINE

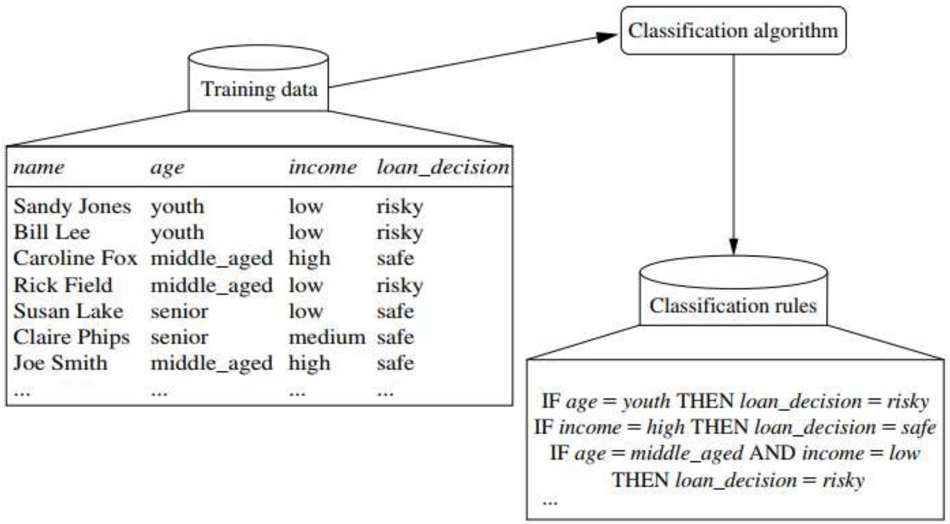
- Introduction
- Data mining overview
- Data pre-processing
- **Classification techniques**
- **Clustering techniques**
- **Association rules**
- References

CLASSIFICATION TECHNIQUES

CLASSIFICATION (1/2)

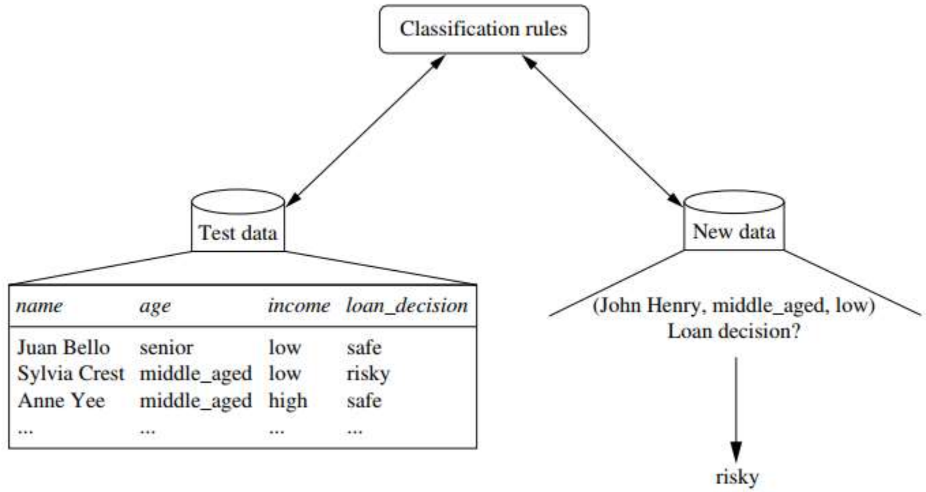
- Data classification is a two-step process:
  - A learning step where a classification model is constructed
  - A classification step where the model is used to predict class labels for given data
- E.g.,
  - Loan applicants are safe or risky
  - A customer profile to buy a computer
  - One of the treatment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

CLASSIFICATION (2/2)



name	*age*	*income*	*loan_decision*
Sandy Jones	youth	low	risky
Bill Lee	youth	low	risky
Caroline Fox	middle_aged	high	safe
Rick Field	middle_aged	low	risky
Susan Lake	senior	low	safe
Claire Phips	senior	medium	safe
Joe Smith	middle_aged	high	safe
...	...	...	...

(a)



name	*age*	*income*	*loan_decision*
Juan Bello	senior	low	safe
Sylvia Crest	middle_aged	low	risky
Anne Yee	middle_aged	high	safe
...	...	...	...

(b)

[2]

DECISION TREES (1/3)

- ID3
- C4.5
- CART

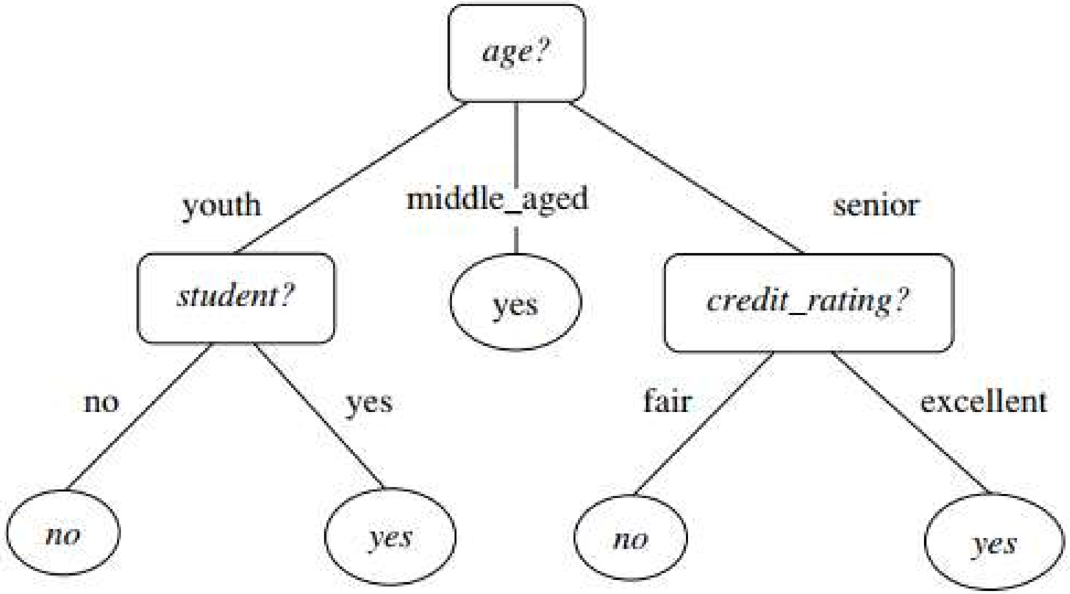


Figure 8.2 A decision tree for the concept *buys_computer*, indicating whether an *AllElectronics* customer is likely to purchase a computer. Each internal (nonleaf) node represents a test on an attribute. Each leaf node represents a class (either *buys_computer* = *yes* or *buys_computer* = *no*).

[2]

DECISION TREES (2/3)

- E.g.,

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

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

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

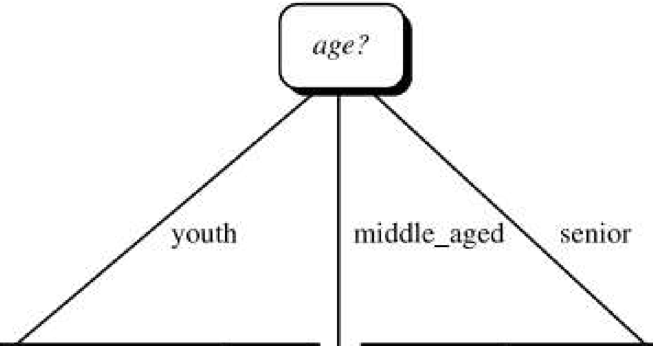
Table 8.1 Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
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

[2]

DECISION TREES (3/3)

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



youth

income	*student*	*credit_rating*	*class*
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

middle_aged

income	*student*	*credit_rating*	*class*
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

senior

income	*student*	*credit_rating*	*class*
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

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.

[2]

FOR EXAMPLE

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

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

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

$Gain(income) = 0.029$ bits, $Gain(student) = 0.151$ bits $Gain(credit\_rating) = 0.048$ bits

[2]

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.

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

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

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

[2]

BAYES CLASSIFICATION METHODS (2/2)

$$X = (age = youth,\ income = medium,\ student = yes,\ credit\_rating = fair)$$

- P(buys_computer=yes|X) = ?
- P(buys_computer=no|X) = ?

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

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
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

[2]

FOR INSTANCE

$$P(buys\_computer = yes) = 9/14 = 0.643$$
$$P(buys\_computer = no) = 5/14 = 0.357$$

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

$$\begin{aligned}
P(age = youth \mid buys_computer = yes) &= 2/9 = 0.222\\
P(age = youth \mid buys_computer = no) &= 3/5 = 0.600\\
P(income = medium \mid buys_computer = yes) &= 4/9 = 0.444\\
P(income = medium \mid buys_computer = no) &= 2/5 = 0.400\\
P(student = yes \mid buys_computer = yes) &= 6/9 = 0.667
\end{aligned}$$

$$\begin{aligned}
P(student = yes \mid buys_computer = no) &= 1/5 = 0.200\\
P(credit_rating = fair \mid buys_computer = yes) &= 6/9 = 0.667\\
P(credit_rating = fair \mid buys_computer = no) &= 2/5 = 0.400
\end{aligned}$$

Using these probabilities, we obtain

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

Similarly,

$$P(\boldsymbol{X}|buys\_computer = no) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019.$$

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

$$P(\boldsymbol{X}|buys\_computer = yes)P(buys\_computer = yes) = 0.044 \times 0.643 = 0.028$$
$$P(\boldsymbol{X}|buys\_computer = no)P(buys\_computer = no) = 0.019 \times 0.357 = 0.007$$

Therefore, the naïve Bayesian classifier predicts $buys\_computer = yes$ for tuple $\boldsymbol{X}$. ■

[2]

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.

[2]

EVALUATION METRICS (2/2)

- Confusion matrix

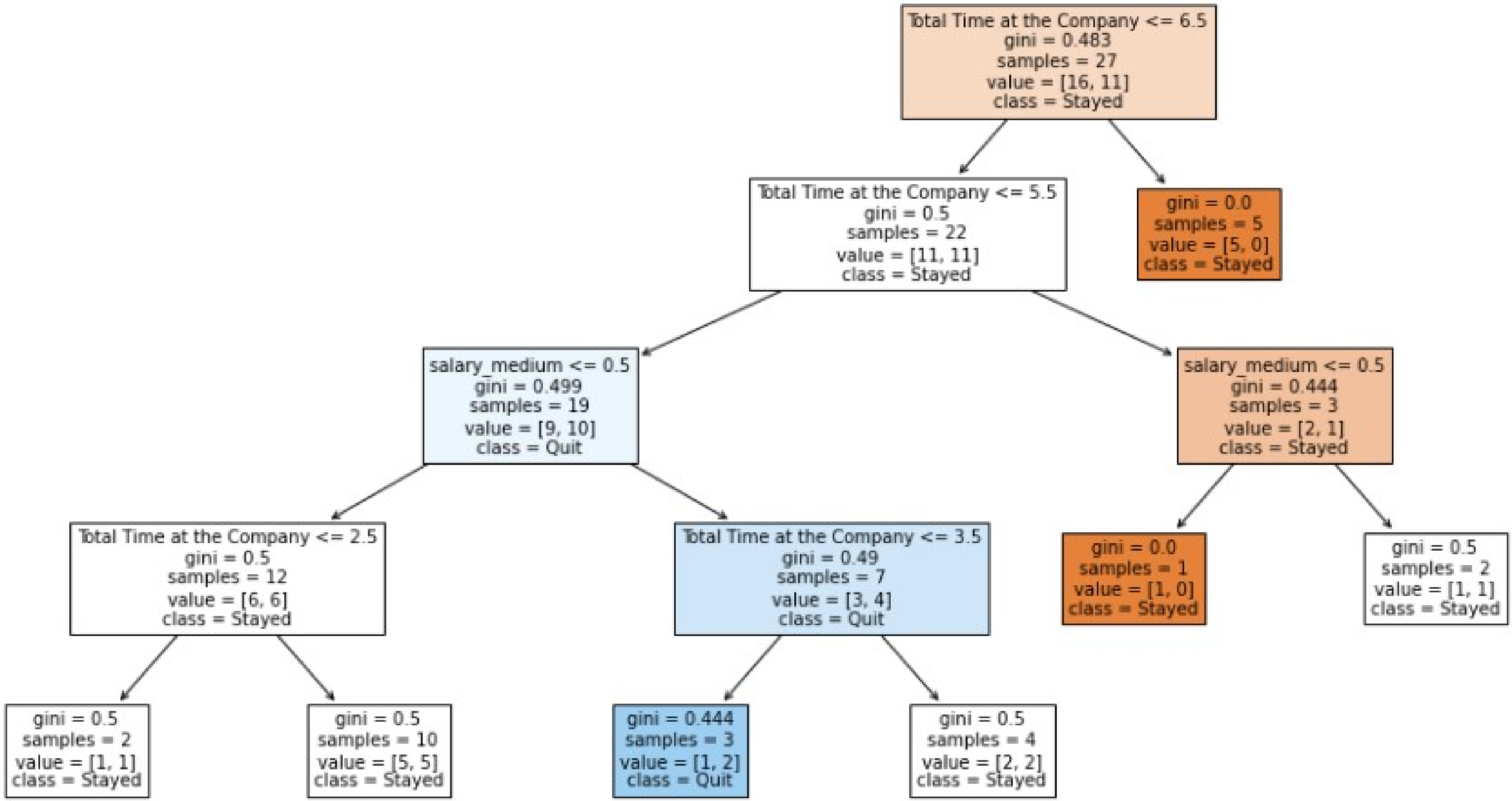
Actual class \ Predicted class	*yes*	*no*	Total
yes	*TP*	*FN*	*P*
no	*FP*	*TN*	*N*
Total	*P'*	*N'*	*P + N*

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	$\frac{2 \times precision \times recall}{precision + recall}$
F_β, where β is a non-negative real number	$\frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$

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).

[2]

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

HIERARCHICHAL METHODS

- Agglomerative
- Divisive

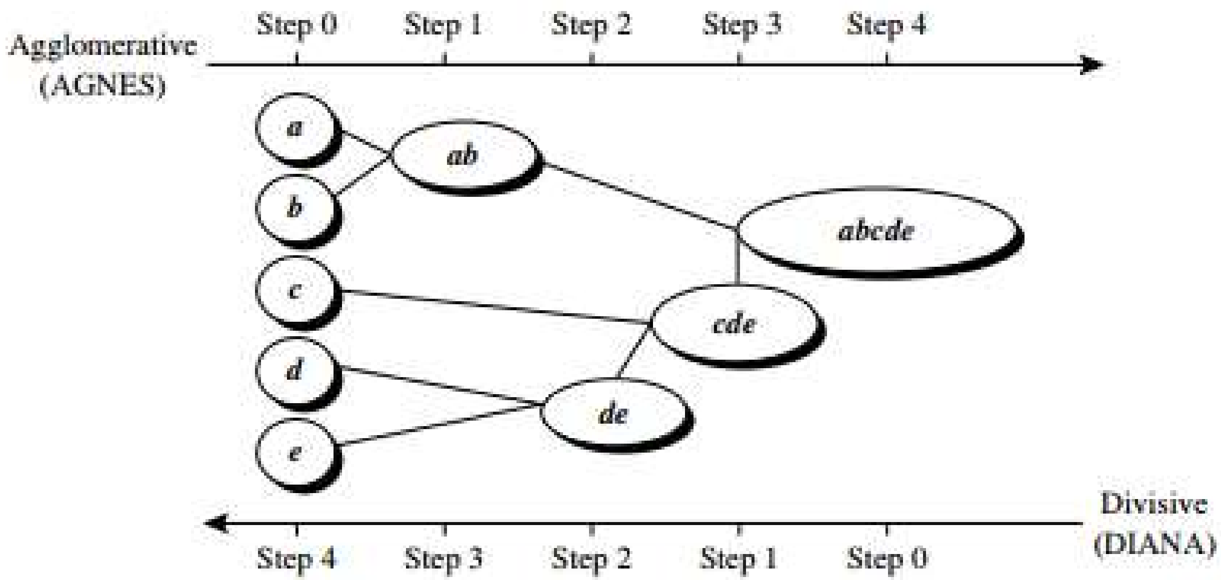
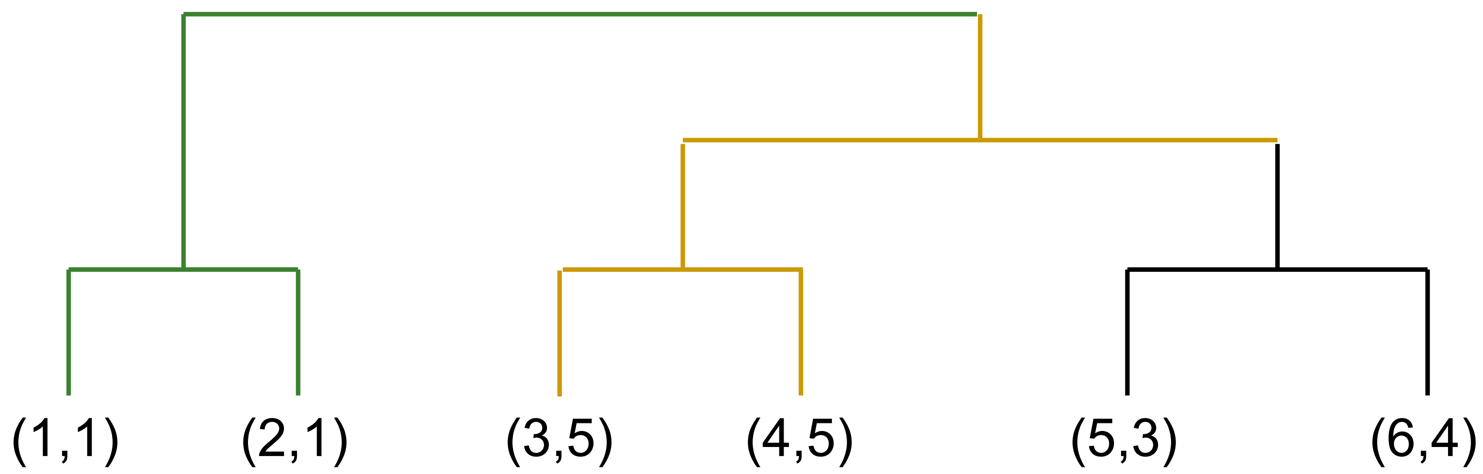


Figure 10.6 Agglomerative and divisive hierarchical clustering on data objects $\{a, b, c, d, e\}$.

[2]

AGGLOMERATIVE HIERARCHICHAL 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 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

- eps = 0.6 and 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

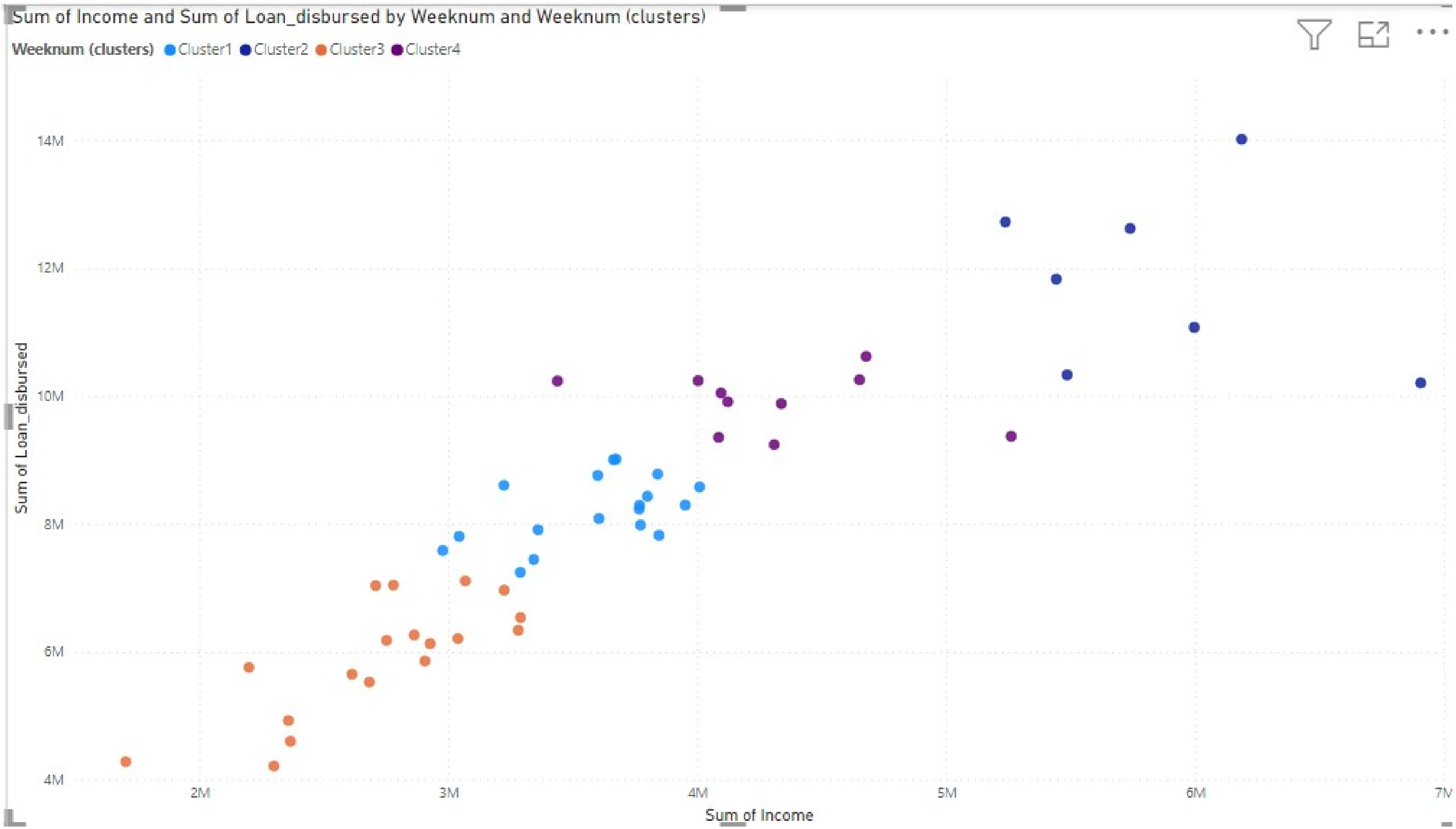
GRID-BASED METHODS

- 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

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

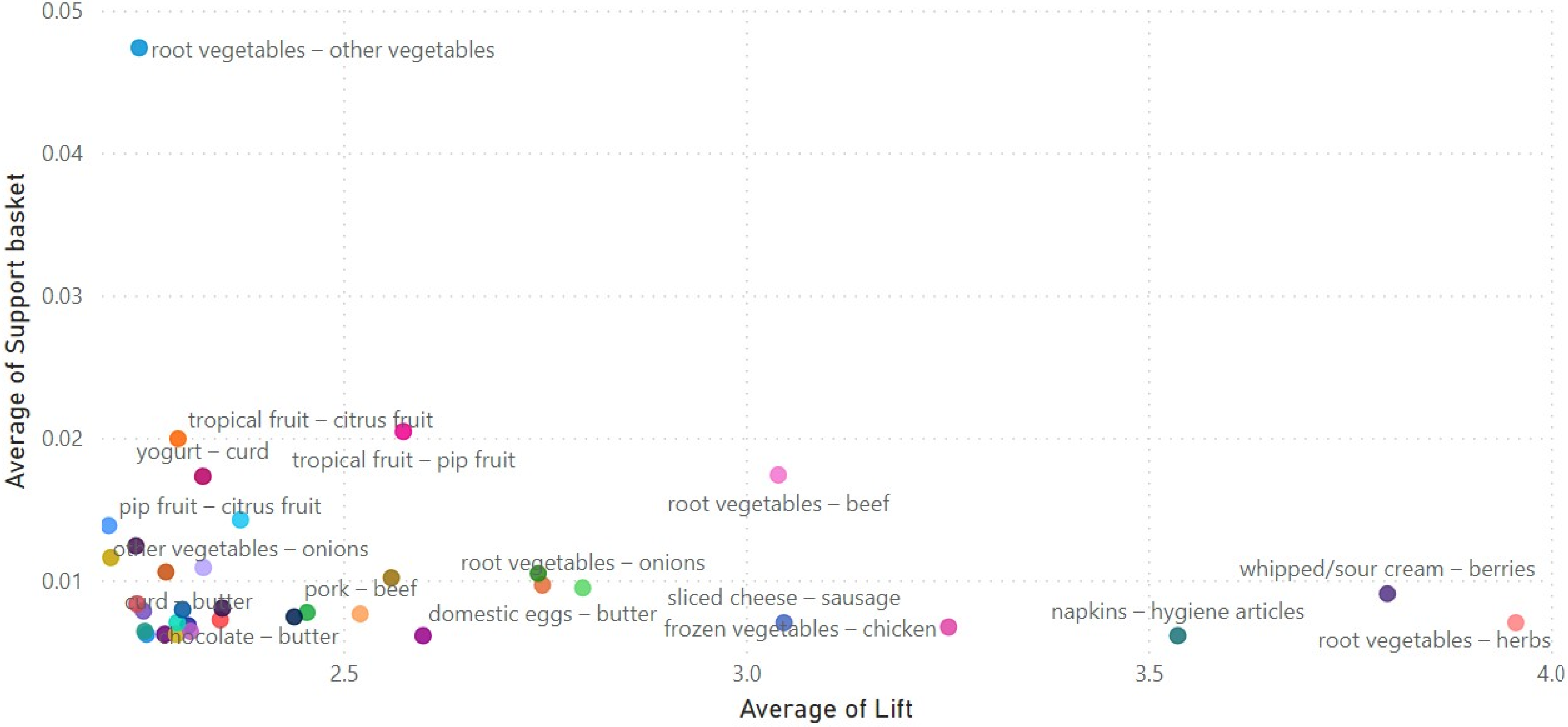
$$Support = \frac{Number\ of\ transactions\ including\ one\ or\ multiple\ products}{Total\ number\ of\ transactions}$$

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

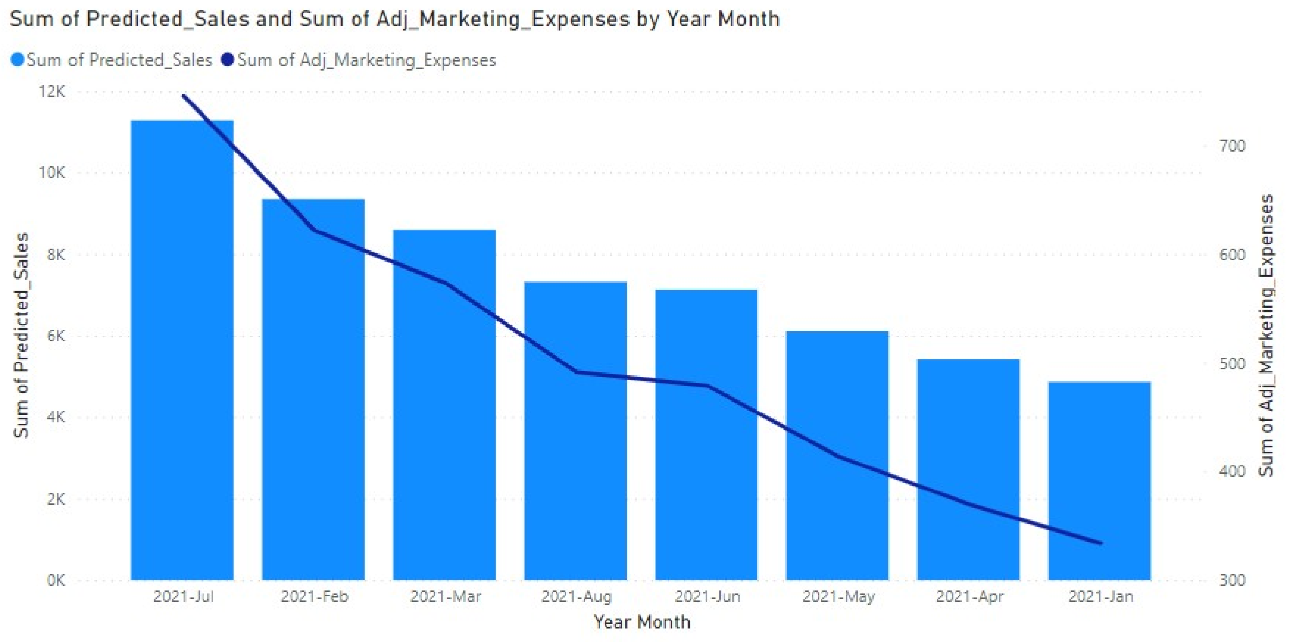
$$Confidence\ of\ product\ two \rightarrow Basket = \frac{Support\ of\ basket}{Support\ of\ product\ two}$$

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

BASKET ANALYSIS DEMO



SALES AND MARKETING EXPENSES DEMO



Marketing_Expenses_Range

All

Year Month	Sum of Budget_Marketing_Expenses	Marketing_Expenses_Range	Sum of Adj_Marketing_Expenses	Sales with Budgeted ME	Sum of Predicted_Sales	Growth
2021-Apr	7.60	0	7.60	102.35	102.35	0.00%
2021-Apr	7.60	1	8.60	102.35	117.93	15.23%
2021-Apr	7.60	2	9.60	102.35	133.52	30.45%
2021-Apr	7.60	3	10.60	102.35	149.10	45.68%
2021-Apr	7.60	4	11.60	102.35	164.68	60.91%
2021-Apr	7.60	5	12.60	102.35	180.27	76.13%
2021-Apr	7.60	6	13.60	102.35	195.85	91.36%
2021-Apr	7.60	7	14.60	102.35	211.44	106.58%
Total	**2,347.80**		**4,027.80**	**1,613.57**	**60,066.06**	**3622.56%**

QUESTIONS AND ANSWERS

Picture from: http://philadelphiasculpturegym.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.