香港城市大學
City University of Hong Kong
專業 創新 胸懷全球
Professional·Creative
For The World

Overview

CS5483 Data Warehousing and Data Mining

Guess the value of y

0	0	0
0	1	0
1	0	0
1	1	1
0	0	y

Guess the value of y

0	0	0
0	1	0
1	0	0
1	1	1
0	0	y

First two columns have two 1's, so $y =$ ___.

Guess the value of y

	X_1	X_2	Y
1.	0	0	0
2.	0	1	0
3.	1	0	0
4.	1	1	1
	0	0	y

- X_i: Result of Test $i = 1,2$.
- Y: Diagnosis of certain disease.
- Closest to Row 1, so $y =$ __.

What are data?

- Dataset: a set of **facts**
  - A_____/features
  - Instances/samples/t_____
- _____ (CSV)
- _____(ARFF)
- Database:
  relational, object-oriented, spatial, text, multimedia,…
- Database Management Systems:
  MySQL, PostgreSQL, EnterpriseDB, MongoDB, MariaDB, Microsoft SQL Server, Oracle, Sybase, SAP HANA, MemSQL, SQLite, IBM DB2,…

X_1	X_2	Y
0,	0,	0
0,	1,	0
1,	0,	0
1,	1,	1

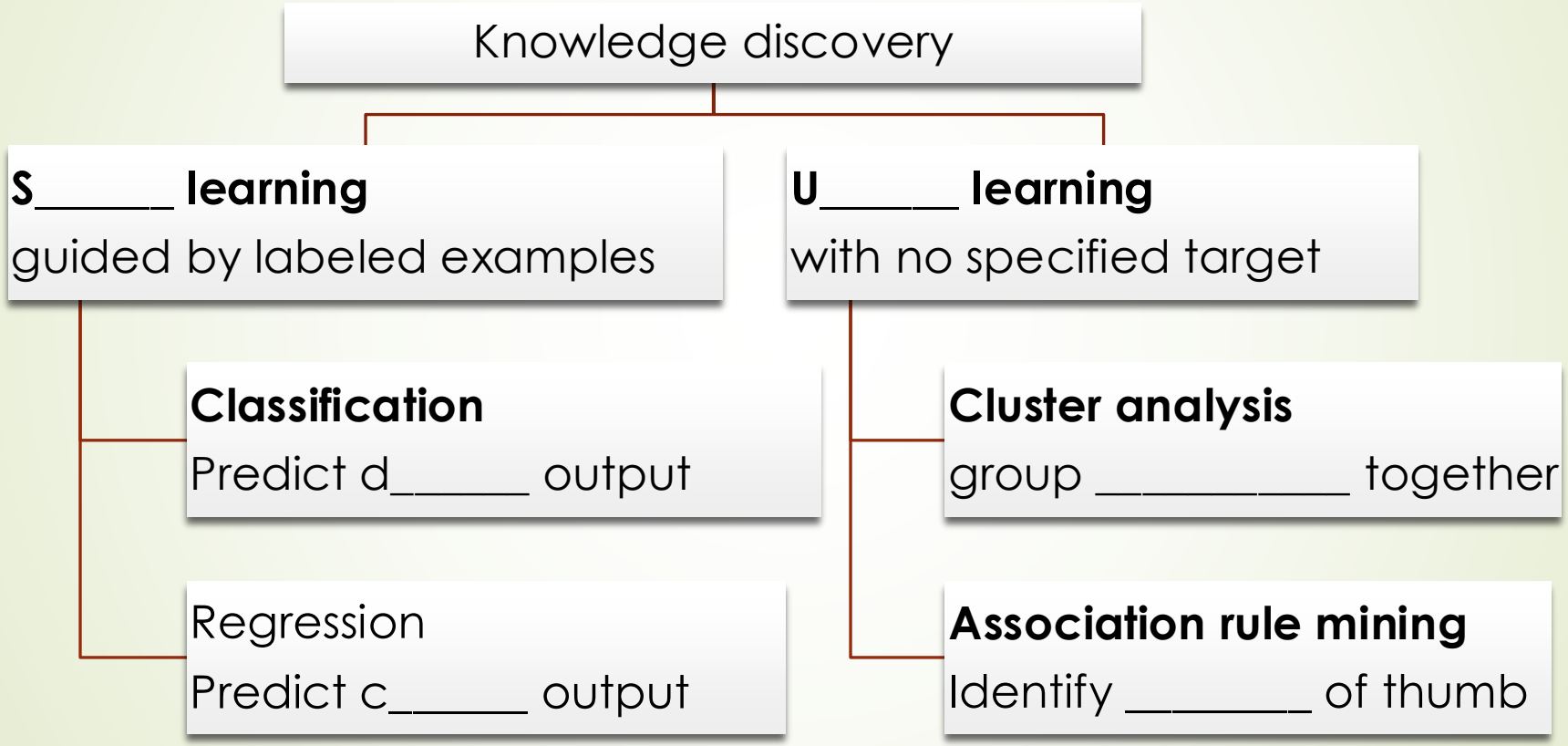
X1,	X2,	Y
0,	0,	0
0,	1,	0
1,	0,	0
1,	1,	1

```
@relation my_relation
% comments …
@attributes X1 {0,1}, X2 {0,1}, Y {0,1}
@data
0, 0, 0
0, 1, 0
1, 0, 0
1, 1, 1
```

Data mining?

- Automatic _____ (KDD).

What is knowledge discovery?



What is knowledge?

- To learn to ask (學問)
- Interesting pattern/structure
  - Pattern: regularity that repeats in a predictable manner
  - Interesting: valid, novel, useful, understandable by human, implementable by computer.

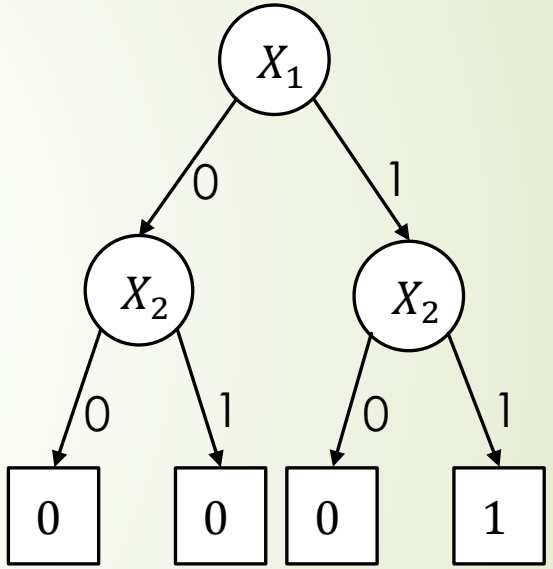
Guess the value of y for any (x_1, x_2)

	X_1	X_2	Y
1.	0	0	0
2.	0	1	0
3.	1	0	0
4.	1	1	1
	x_1	x_2	y

- Table lookup.

Guess the value of y for any (x_1, x_2)

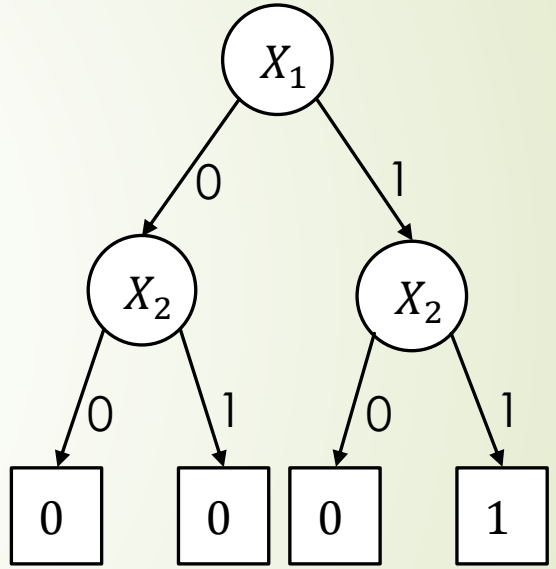
	X_1	X_2	Y
1.	0	0	0
2.	0	1	0
3.	1	0	0
4.	1	1	1
	x_1	x_2	y



- Table lookup.
- _____ tree.

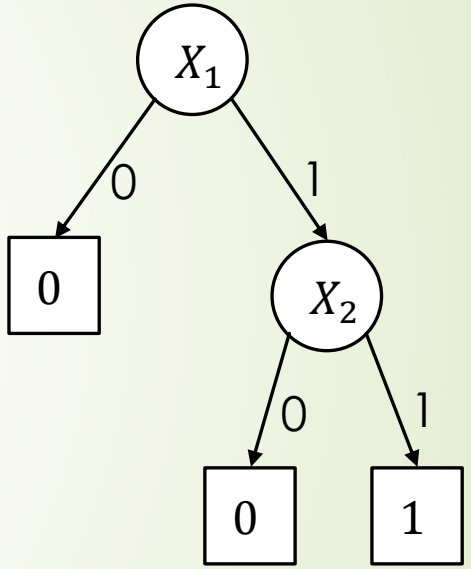
Guess the value of y for any (x_1, x_2)

	X_1	X_2	Y
1.	0	0	0
2.	0	1	0
3.	1	0	0
4.	1	1	1
	x_1	x_2	y



Guess the value of y for any (x_1, x_2)

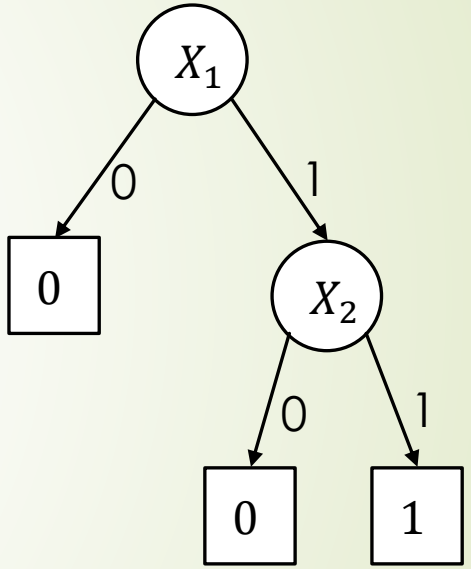
	X_1	X_2	Y
1.	0	0	0
2.	0	1	0
3.	1	0	0
4.	1	1	1
	x_1	x_2	y



- _____ to a smaller tree.
- Simplicity preferred.
  - Improve _____ and _____.

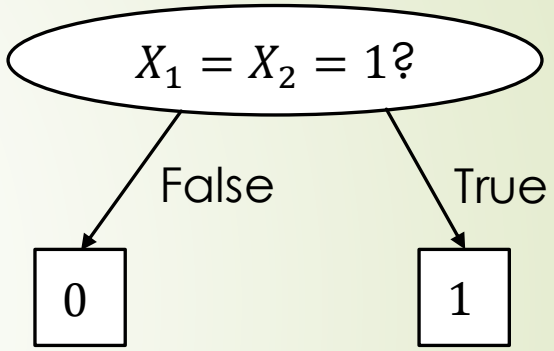
Guess the value of y for any (x_1, x_2)

	X_1	X_2	Y
1.	0	0	0
2.	0	1	0
3.	1	0	0
4.	1	1	1
	x_1	x_2	y



Guess the value of y for any (x_1, x_2)

	X_1	X_2	Y
1.	0	0	0
2.	0	1	0
3.	1	0	0
4.	1	1	1
	x_1	x_2	y



if $X_1 = 1$ and $X_2 = 1$, then $Y =$ ___, else $Y =$ ____.

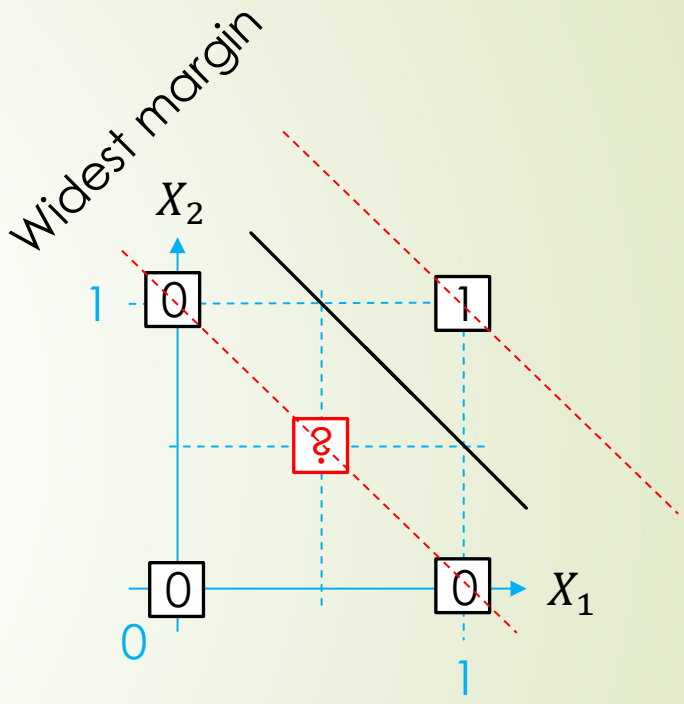
Guessing with numeric values

	X_1	X_2	Y
1.	0	0	0
2.	0	1	0
3.	1	0	0
4.	1	1	1
	0.5	0.5	y

- $Y = X_1 \cdot X_2$, so $y =$ _____.
- What about $Y = X_1^2 \cdot X_2^2$?
- What if $Y \in \{0,1\}$?

Guessing with numeric values

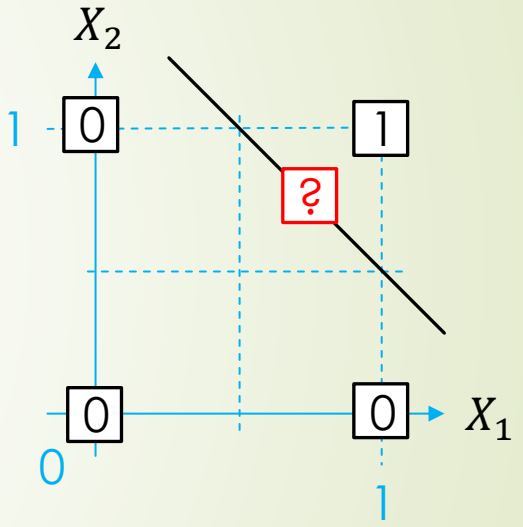
	X_1	X_2	Y
1.	0	0	0
2.	0	1	0
3.	1	0	0
4.	1	1	1
	0.5	0.5	y



- What if $Y \in \{0,1\}$?
- Visualize data by a scatter plot.
- Draw a decision boundary. So $y =$ ____. Optimality?
- Minimize the risk of error.

Guessing with numeric values

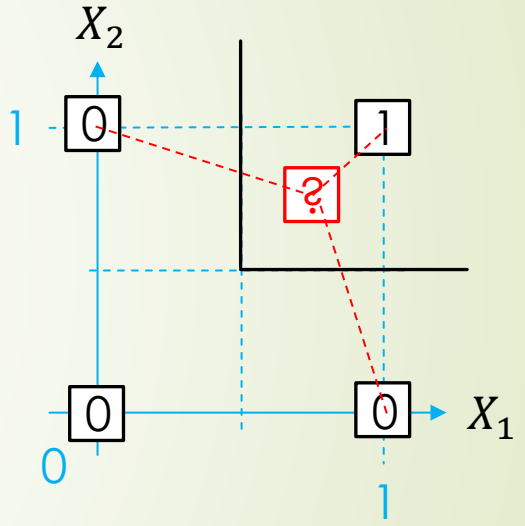
	X_1	X_2	Y
1.	0	0	0
2.	0	1	0
3.	1	0	0
4.	1	1	1
	0.75	0.75	y



- What about points on the boundary?

Guessing with numeric values

	X_1	X_2	Y
1.	0	0	0
2.	0	1	0
3.	1	0	0
4.	1	1	1
	0.75	0.75	y



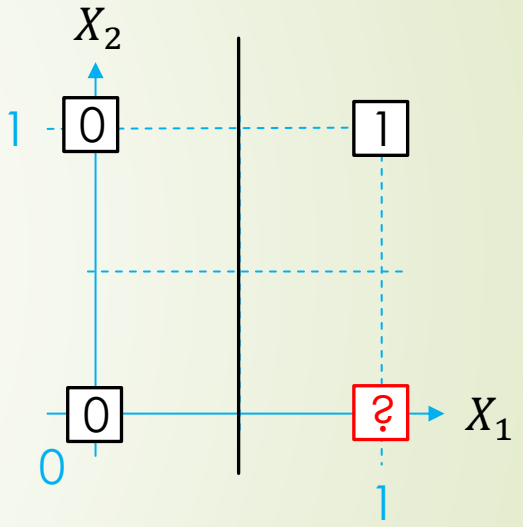
- What about points on the boundary?
- With a different boundary, $y =$ ___. Why?
- $(0.75, 0.75)$ is more similar to _____ than to others. (Nearest neighbour)

Know your data

- Attribute types
  - Nominal/categorical/qualitative, ordinal or not
  - Numeric/quantitative
- Visualization
  - Scatter plots, histogram, boxplots, quantile plots, quantile-quantile plots,…
  - Related statistics: mean, median, mode, variance, standard deviation, quartiles, interquartile range,…
- Similarity/proximity measures
  - Euclidean, Manhattan, Minkowski or supremum distances, Jaccard coefficient, term-frequency vectors, cosine measure, Tanimoto coefficient,…

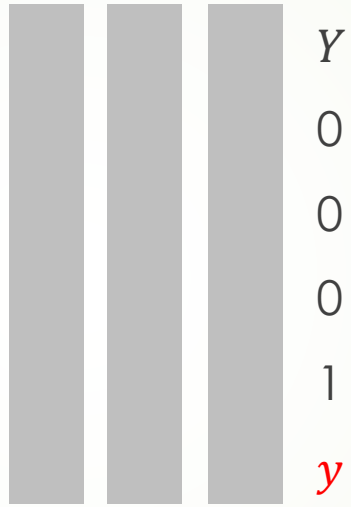
Missing values

	X_1	X_2	Y
1.	0	0	0
2.	0	1	0
3.	~~1~~	~~0~~	~~0~~
4.	1	1	1
	1	0	y



- _____ the instance.
- $y =$ ___

Redundant attributes



$Y = X_1 \cdot X_2$ X_3 r_____

$= X_2 \cdot (1 - X_3)$ X_1 r_____

$= X_1 \cdot (1 - X_3)$ X_2 r_____

Redundancy helps

	X_1	X_2	X_3	Y
1.	0	0	0	0
2.	0	1	1	0
3.	▮	0	1	0
4.	1	1	0	1
	1	0	1	y

- $Y = X_2 \cdot (1 - X_3)$, so y = __ by keeping the row with missing value
- Is it better than removing the instance?

Irrelevant attributes can be removed

	X_1		Y
1.	0		0
2.	0		0
3.	1		1
4.	1		1
	x_1		y

- $Y = X_1$
- X_2 i_____ of Y (even given X_1)

Knowledge Discovery from Databases

1. **P**_____
   set the **goal** (what to learn)
2. **P**_____
   - Data **cleaning/integration**: handle **noise/errors/missing values**
   - Data **selection/reduction/transformation/discretization**: create **target data set** with relevant samples/variables
3. **Data mining**
   - Apply **learning algorithm**(s) to compute desired patterns from processed data
4. **I**_____
   - **Iterate** if performance is unsatisfactory
   - **Deploy** if ready: report, incorporate, apply

Different data processing methods

- Missing value
  - Ignore, surrogate splits, impute (fill in) manually or with mean/median/mode,…
- Noise
  - Bin smoothing, regression, outlier analysis
- Attribute selection
  - Measures/test of redundancy/relevance
    - χ^2 correlation test, correlation coefficient, Mutual information
  - Methods
    - Correlation matrix with heatmap
    - Feature importance (from decision trees)
    - Forward selection vs backward elimination

References

- Han11 Chapter 1-3.