UIC
BUSINESS

Business Data Mining

IDS 472 (Spring 2024)

Instructor: Wenxin Zhou

Lectures and office hours

UIC BUSINESS

- Web Page: https://uic.blackboard.com
- We meet on Tuesdays & Thursdays (from 11:00 am to 12:15 pm)
  - Lectures will be delivered at Behavioral Science Building 145
  - If classroom equipment allows, lectures will be recorded by Echo360 and can be accessed on Blackboard
- Piazza: For questions on class material and assignments, extra credit
- My office hours: Thursdays, 1:00 PM – 3:00 PM
- TA: Sheida Hassani (shassa45@uic.edu)
  - Very knowledgeable about Python programming language

Course material

UIC BUSINESS

- Recommended Textbook:

  An Introduction to Statistical Learning: with Applications in Python

  by G. James, D. Witten, T. Hastie, R. Tibshirani & J. Taylor

  Data Mining Practical Machine Learning Tools and Techniques (4th Edition)

  by Ian H. Witten, Eibe Frank, Mark A. Hall & Christopher J. Pal

- E-versions of these books will be provided on Blackboard

- Python: please install it on your machines (Windows of Mac OS). An instruction (Python_Setup.pdf) has been provided on Blackboard.

Grading components

UIC BUSINESS

- Homework: 50%
  - Homework assignments can be completed by group, two-three students per group
  - Submit answers on Gradescope as PDF
  - See the syllabus for late submission policy
- Mid-term: 20%
  - Thursday of Week 8 (Feb 29), in-class
- Final exam (Date TBA): 30%
- Piazza Extra Credit: 5%
  - Students start with zero points (0%). Students will receive additional percentage points (up to an additional 5%) for providing helpful and timely answers to other students' questions on Piazza throughout the semester. Your participation on Piazza will be evaluated subjectively, but will rely upon measures of timeliness, helpfulness and insight reflected in your answers, and respectfulness. It will not be possible to provide updates on the participation grade until the end of the semester.

Success in Class

UIC BUSINESS

- Review course material (lecture slides, syllabus, homework)
- Start assignments **early**
- Utilize resources when you have questions
  - Piazza
  - Fellow students
  - Office hours
  - ChatGPT (coding questions)

Wrap-up

UIC BUSINESS

- If I go too fast, slow me down with questions!
- Homework 0 posted on Blackboard, due Tuesday
- Questions?

What is data mining?

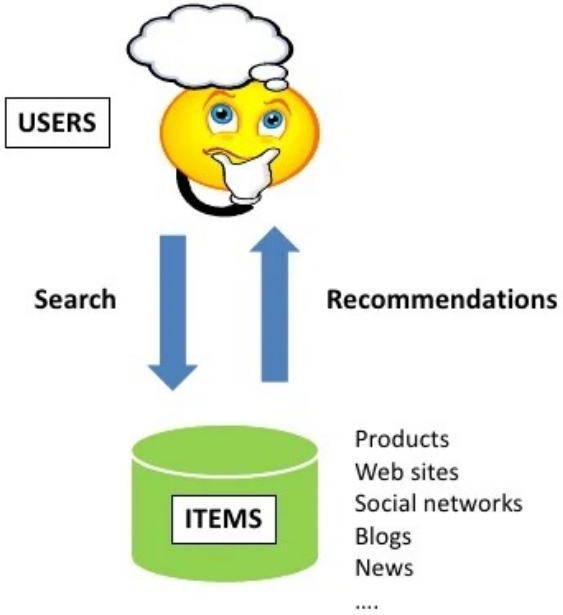
UIC BUSINESS

Exploration and analysis of large quantities of data to discover meaningful patterns

Examples

UIC BUSINESS

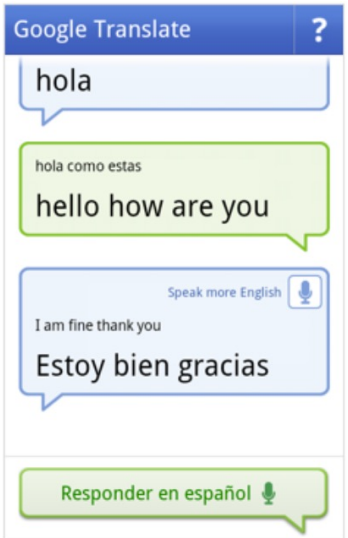
Recommendations

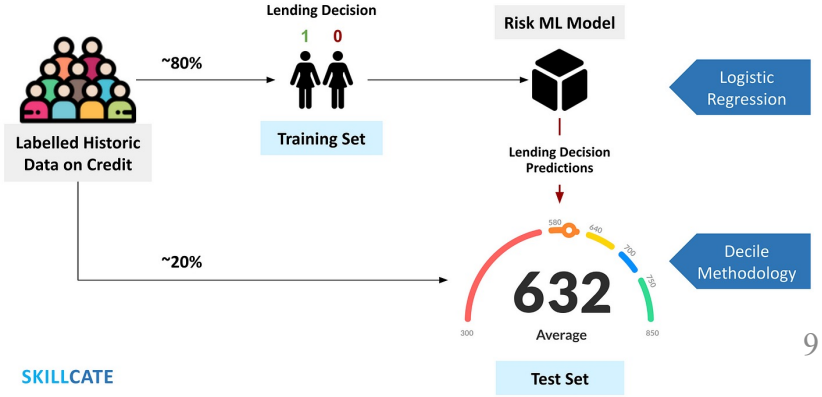


CETAS INSTANT INTELLIGENCE

yelp

PANDORA ebay amazon Walmart Save money. Live better. BEST BUY NETFLIX LinkedIn ORBITZ





SKILLCATE

General applications of data mining

UIC BUSINESS

- Marketing: campaign response, cross-selling, churning.
- Finance: financial outcome, bankruptcy prediction, fraud detection, credit scoring, credit risk analysis, loan evaluation.
- Operations: identification of cause of failure, order prediction.
- Medical: accurate diagnosis, pharmaceutical discovery.
- Customer relationship management
  - Customers most likely to respond to solicitations, customers likely to generate to high purchase revenues, customers at risk of leaving a subscription service (churn), customer lifetime value, …

Web mining, text mining, Image recognition, …

Challenge

UIC
BUSINESS

- Large volumes of data are not the same as knowledge
- Goal of data mining: extract knowledge from big data

Disadvantages of data mining

UIC BUSINESS

- Privacy issues
- Security issues
- Misuse of Information/inaccurate information

Supreme Court to Weigh Drug Data Mining Limits
(WASHINGTON(Dow Jones), Jan 2011)

Data Mining Overview

Data

UIC BUSINESS

- Data set: list of records each of which has values for some predefined fields
- Records = Tuples = Observations ⇔ rows
- Fields = Variables = Attributes/Features = Factors ⇔ columns

Variables

Records

	ID	Age	Sex	Weight	Height	Married	Migrantstatus
1	1	26	1	132	60	0	Nonmigrant
2	2	65	0	122	65	0	Migrant
3	3	15	1	184	67	0	Nonmigrant
4	4	7	1	145	59	0	Nonmigrant
5	5	80	0	100	64	0	Migrant
6	6	43	1	NA	NA	0	Nonmigrant
7	7	28	1	128	67	1	Nonmigrant
8	8	66	1	154	60	1	Nonmigrant
9	9	45	0	166	NA	0	Migrant
10	10	12	0	164	60	1	Migrant

Types of data measurements (for each field)

- Interval/Integer: continuous, can do full math on them

  For example: weight, height, price, account balance

- Nominal: discrete, more than two categories, no order.

  For example: marital status (single, married, divorced), color (red, blue, green, yellow)

- Binary (flag): nominal with only two values

  For example: yes/no, good/bad, pregnant/not pregnant

- Ordinal: discrete, ordered.

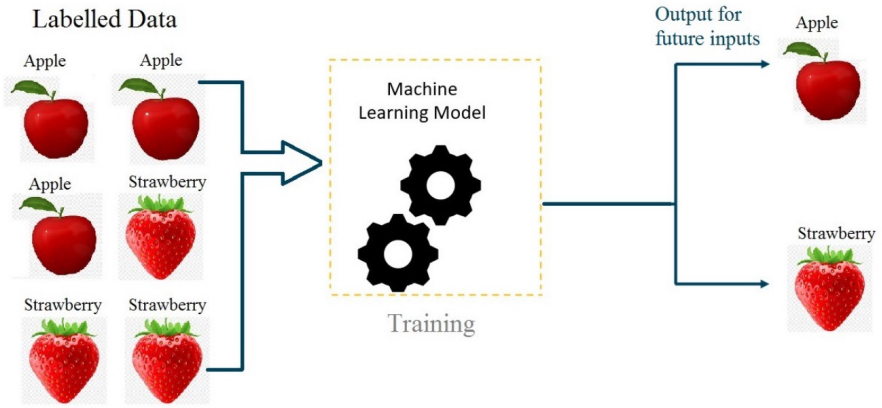
  For example: letter grade (A, B, C, D), level of satisfaction (highly dissatisfied to highly satisfied), size (small, medium, large)

Data mining activities

UIC BUSINESS

Supervised machine learning
(there is specific target attribute to predict)

- Classification
- Estimation (Regression)
- Prediction



Unsupervised machine learning
(there is no specific target)

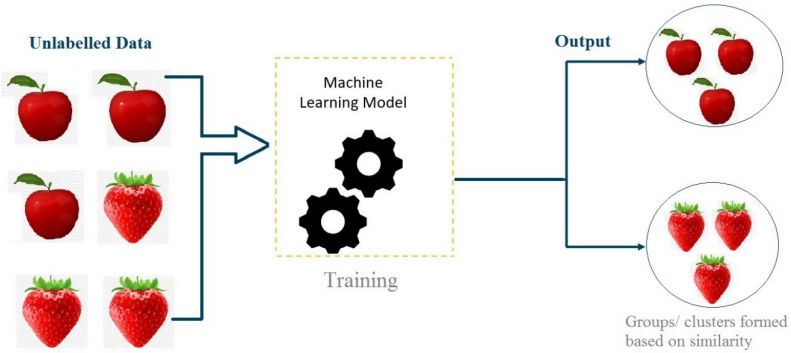
- Clustering
- Affinity grouping (Association rules)

Data mining activities

UIC BUSINESS

Classification

- Assign object to one of a number of predefined classes.
- Examples:
  - Classifying credit applications as low, medium, or high risk
  - Determine who is more likely to buy a particular product



Estimation

- Similar to classification, but with a continuous outcome
- Examples: Tax liability, house value

Data mining activities

UIC BUSINESS

Prediction

- Rules that explain how to predict a future value or classification, given characteristics
- Example:
  - stock price 1 month from now
  - foreclosures in next 6 months if loan terms remain unchanged

Data mining activities

UIC BUSINESS

Clustering: Segmenting a population into some non-predefined classes

- Examples:
  - Customer classes
  - Market segments
  - City-planning: Identifying groups of houses according to their house type, value, and geographical location

Affinity grouping: Determine what things go together

- Heart of "market-basket analysis"
- Examples:
  - People who buy a car seat also buy strollers
  - People who liked "The Office" often like "Portlandia"
  - "beer and diapers sell together on Friday evenings"

Example: Classifying email spam

UIC BUSINESS

Data – email messages with class labels {spam, legitimate}

& most commonly occurring words, punctuation marks (attributes)

	george	you	hp	free	!	edu	remove	our
Spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28	0.51
Email	1.27	1.27	0.90	0.07	0.11	0.29	0.01	0.18

Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between spam and email.

Learned rules:

IF (%george < 0.6) & (%you > 1.5) THEN Spam

IF (0.7 * %free – 1.3* our > 0) THEN Spam

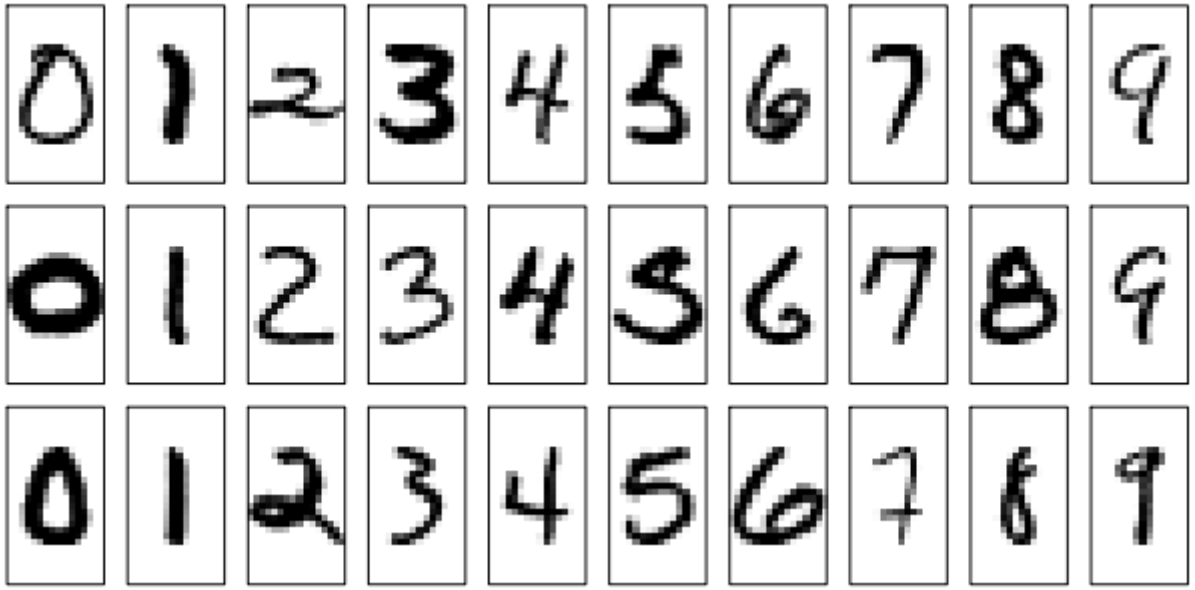
Example – Handwritten character recognition

UIC BUSINESS

Data – handwritten zip-codes on US postal mail

16 x 16 grayscale maps of pixel intensities (0-255)

- Classify as 0,1,…,9 or unknown
- If accurate enough, can be used for automated sorting of mail

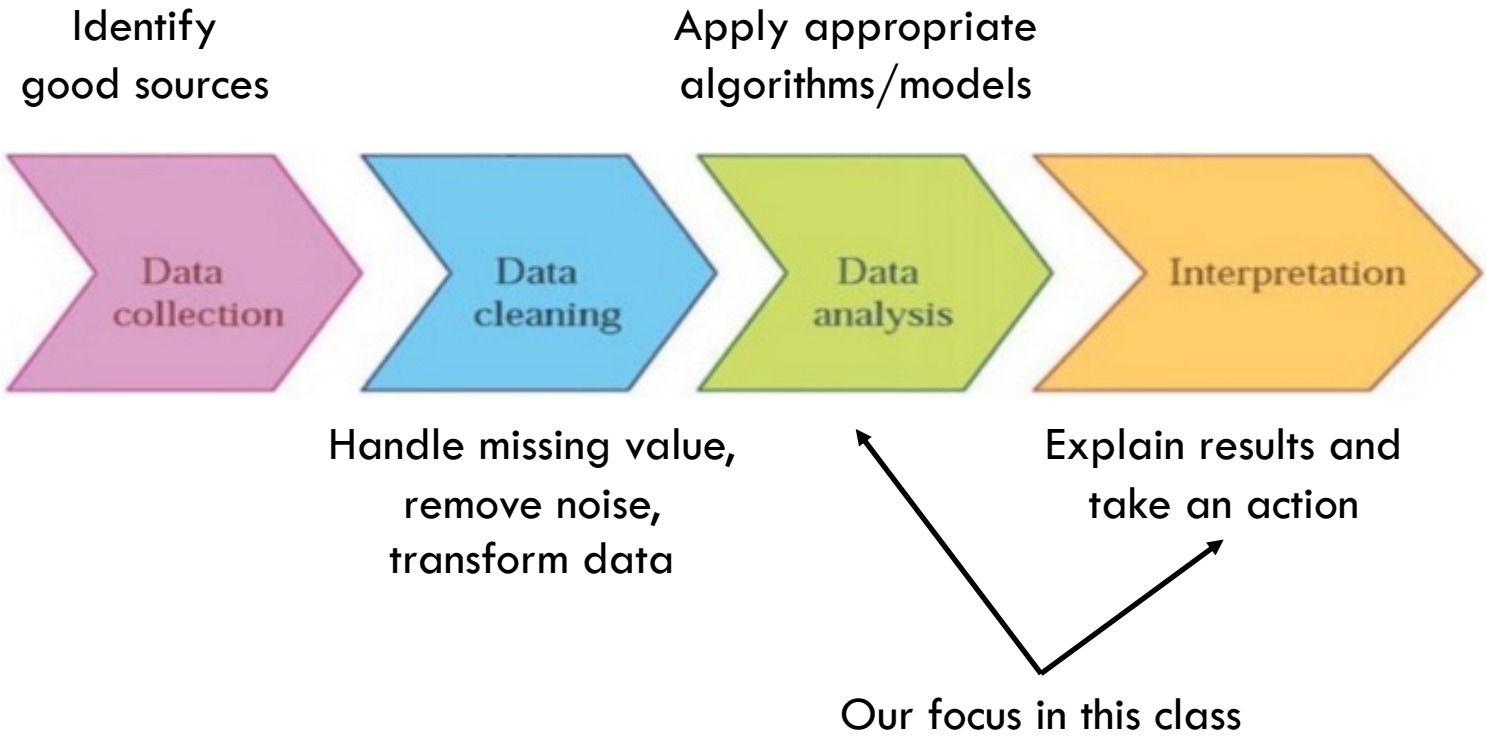


Examples of handwritten digits from US postal envelops

Review: What is data mining?

UIC BUSINESS

- Data Mining is the efficient discovery of previously unknown, valid, potentially useful, understandable patterns in large datasets



Tennis dataset

UIC BUSINESS

Attributes

Records

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
d1	sunny	hot	high	weak	no
d2	sunny	hot	high	strong	no
d3	cloudy	hot	high	weak	yes
d4	rain	mild	high	weak	yes
d5	rain	cool	normal	weak	yes
d6	rain	cool	normal	strong	no
d7	cloudy	cool	normal	strong	yes
d8	sunny	mild	high	weak	no
d9	sunny	cool	normal	weak	yes
d10	rain	mild	normal	weak	yes
d11	sunny	mild	normal	strong	yes
d12	cloudy	mild	high	strong	yes
d13	cloudy	hot	normal	weak	yes
d14	rain	mild	high	strong	no

Some rudimentary data mining models

UIC BUSINESS

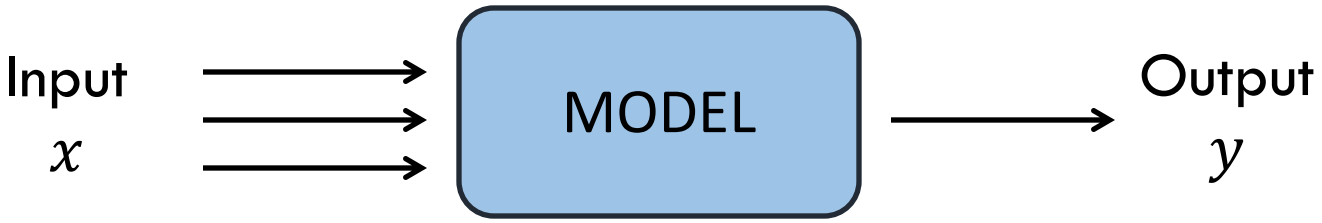
- Model I
  - if temperature = cool then PlayTennis = yes
  - if temperature = hot then PlayTennis = no
  - if temperature = mild then PlayTennis = yes

- Model II
  - if temperature = mild or cool then PlayTennis = yes
  - if temperature = hot and outlook = sunny then PlayTennis = no
  - if temperature = hot and outlook = cloudy then PlayTennis = yes

Which model is better?

Models

UIC BUSINESS



What is a "model"?

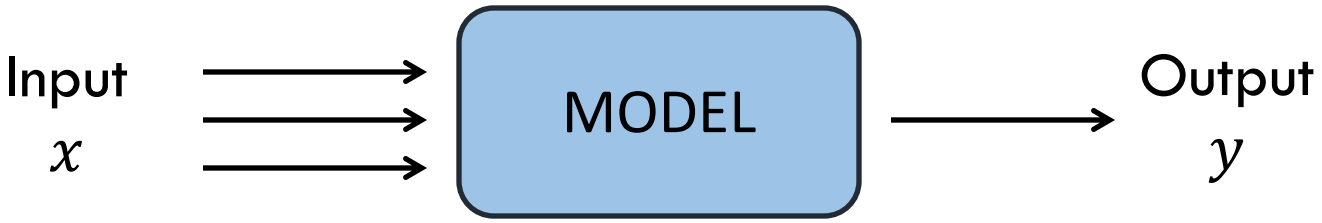
- Linear: $y = w_1x_1 + w_2x_2 + \ ... + w_kx_k$
- Conditional:

  If [(age < 35) and ($30K< income< $70K)] OR
  [(residence=urban) and (age > 50) and (savings > 50K)] OR
  [(income > 50K) and (married = true)] THEN donor = yes

- Decision trees, regression, nearest-neighbor, neural networks, naïve Bayes, …

Models

UIC BUSINESS



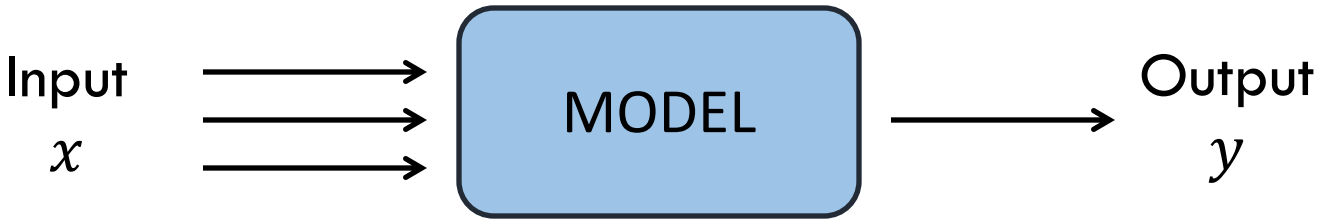
- Developed from data
- Output can be a prediction
  - Customer is a donor, transaction is fraudulent, … (binary)
  - Purchase revenue, insurance claim loss amount, … (real-valued)

Output variable: dependent variable (or outcome variable, target variable)

Input variables: independent variables

Models

UIC BUSINESS



- Developed from data on the past.
- Assumption: past is relevant for future
- How well will a model perform on new 'unseen' data

Data partition: training and testing data

UIC BUSINESS

A good data mining learns patterns from known instances that generalizes to unknown instances.

To check the performance of a model on unseen instances we partition data into "training set" and "testing set".

- Training: use to build models
- Test: use to measure performance on unseen data

Rule of thumb: 70% training, 30% testing

- When evaluating a model crucial pitfall to watch for is overfitting

Illustration of over-fitting

UIC BUSINESS

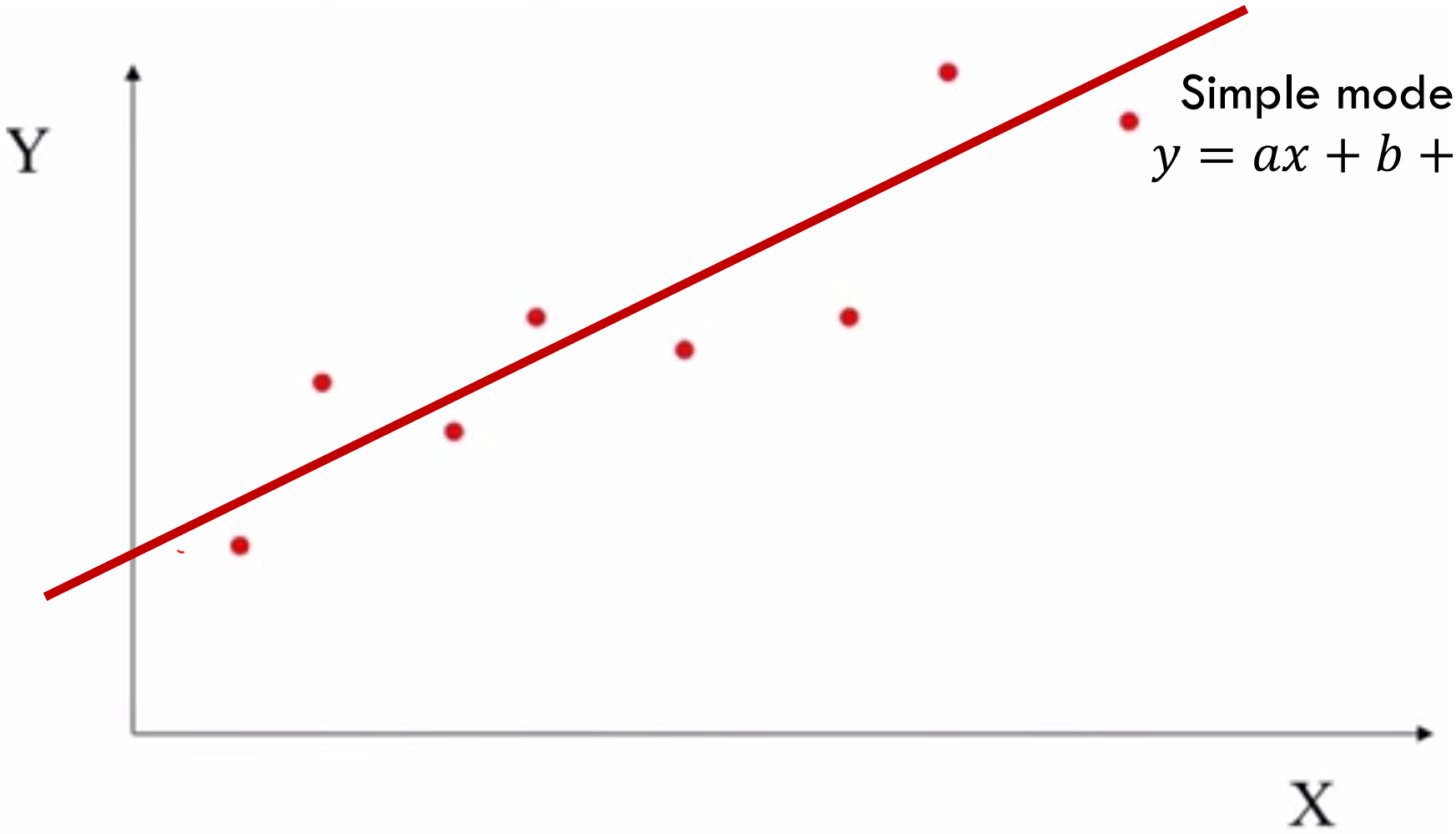
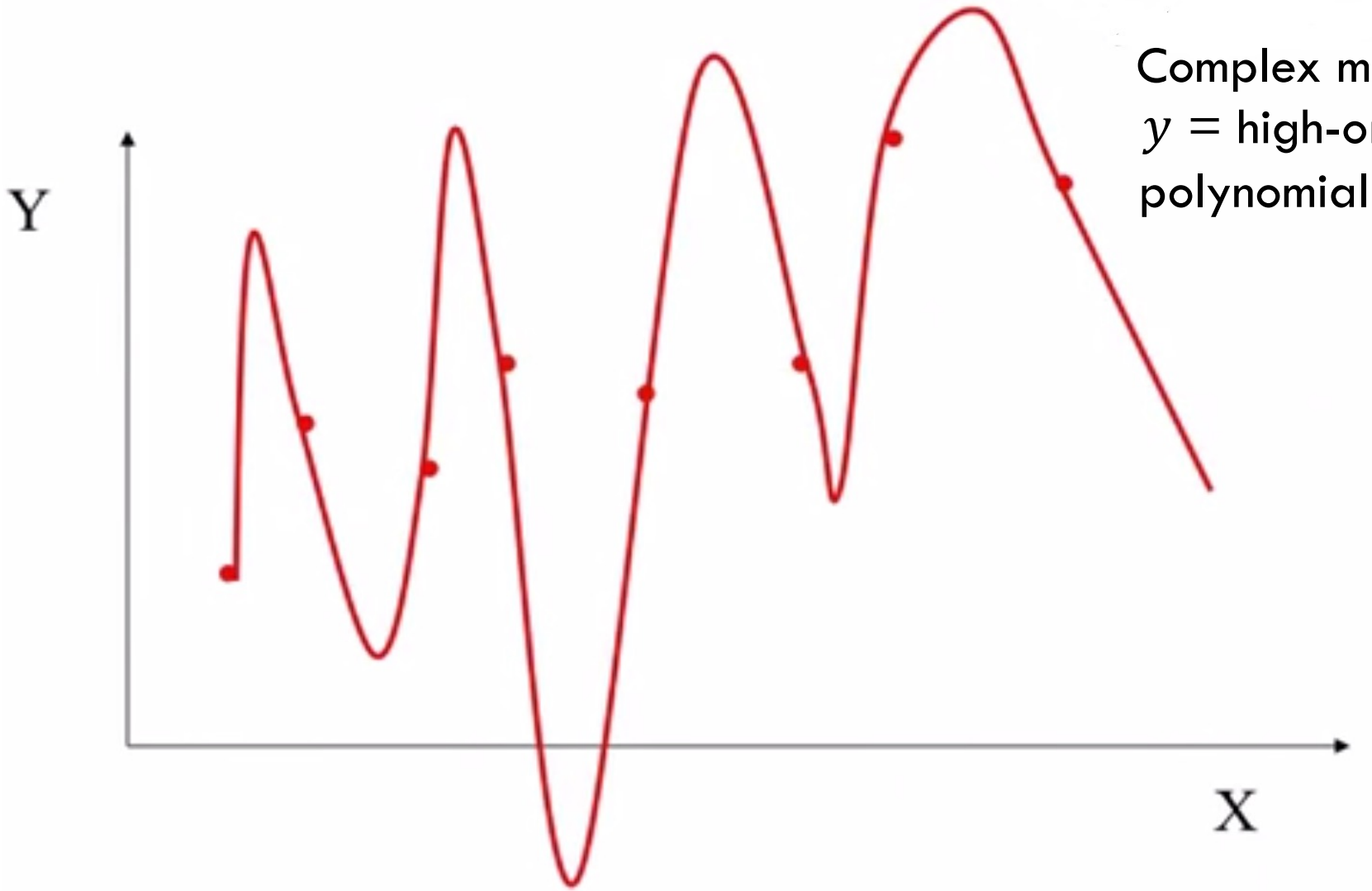


Illustration of over-fitting

UIC BUSINESS



Complex model
y = high-order polynomial in x

Illustration of over-fitting

UIC BUSINESS

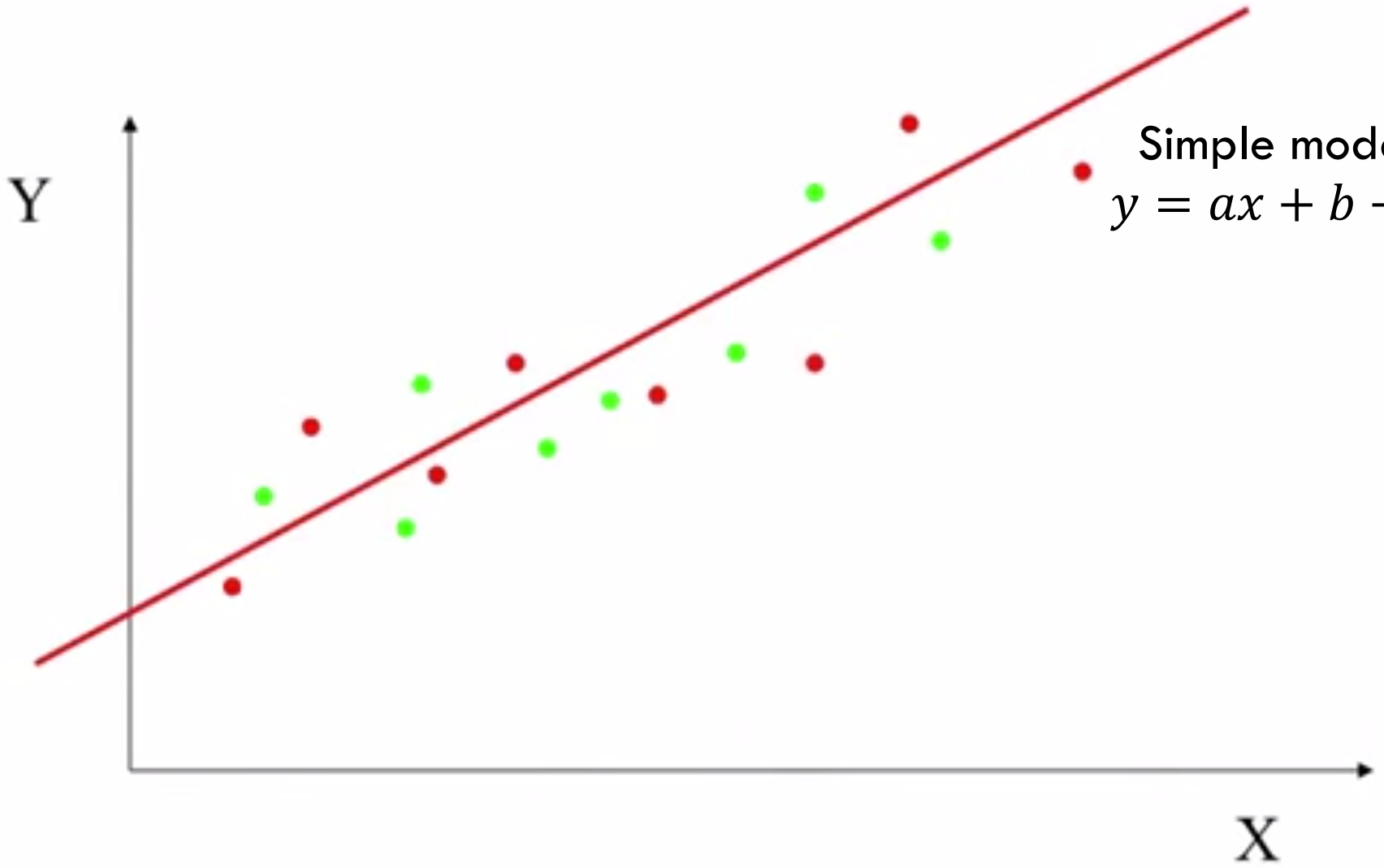
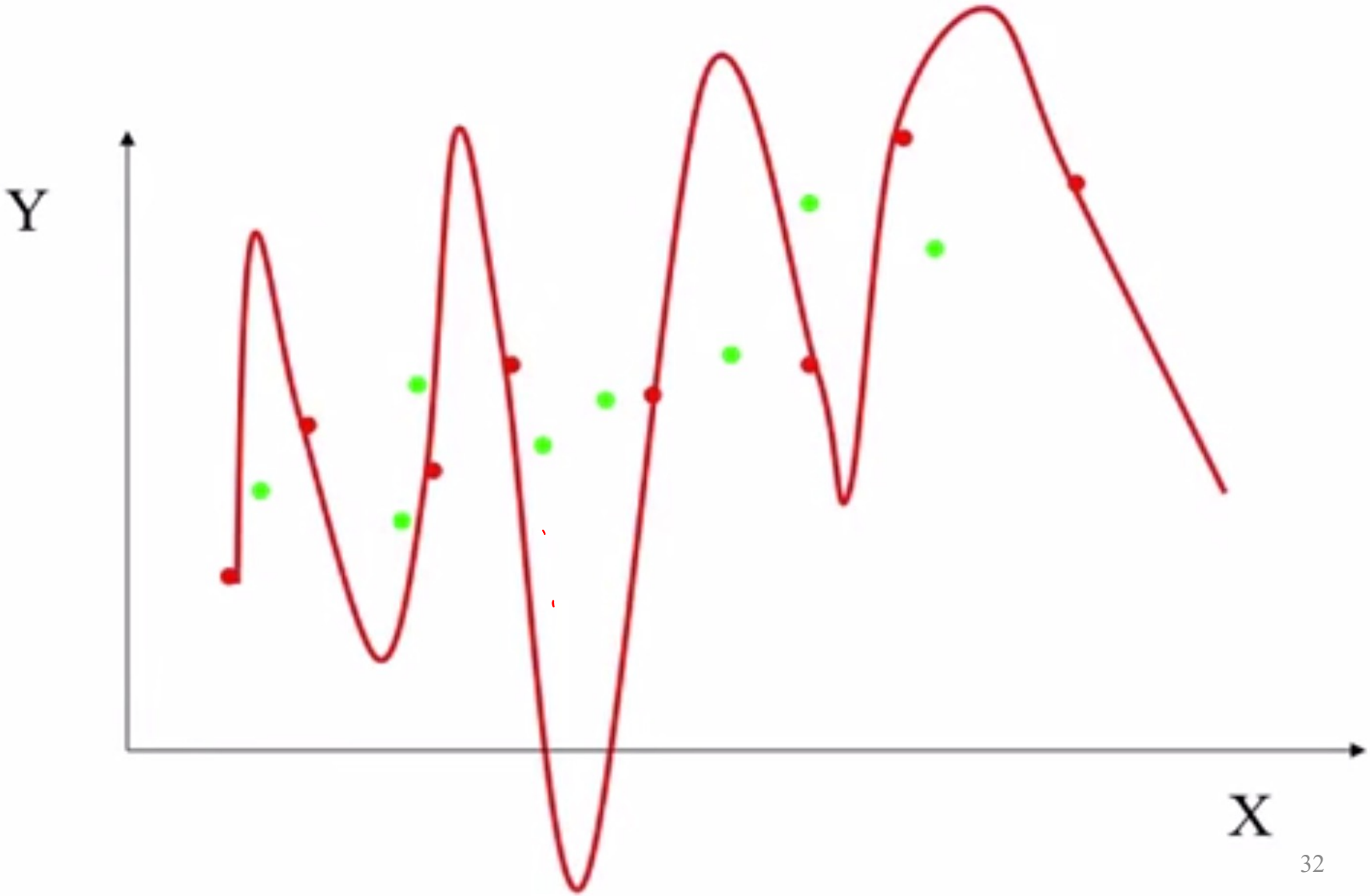


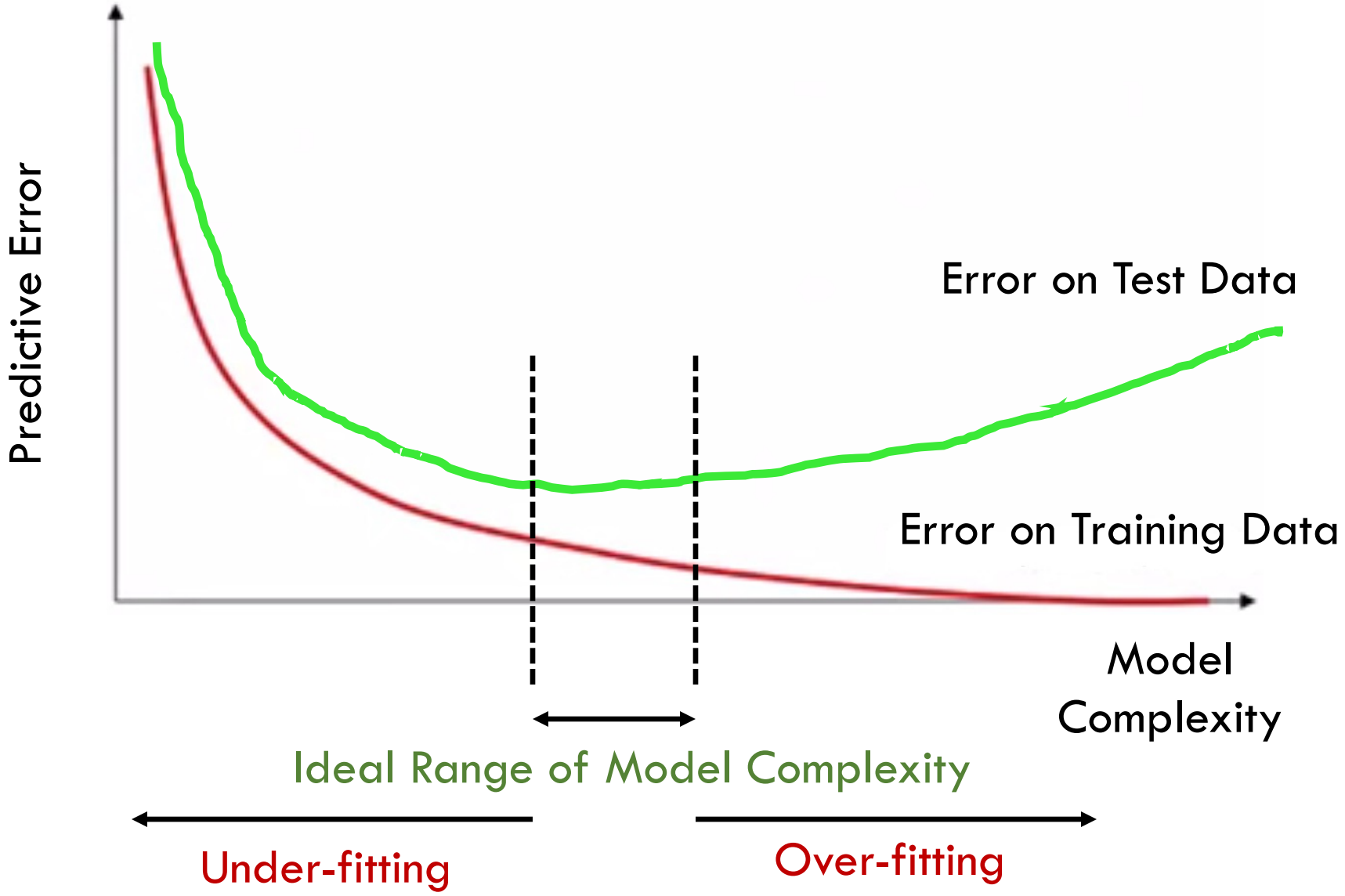
Illustration of over-fitting

UIC BUSINESS



How over-fitting affects prediction

UIC BUSINESS



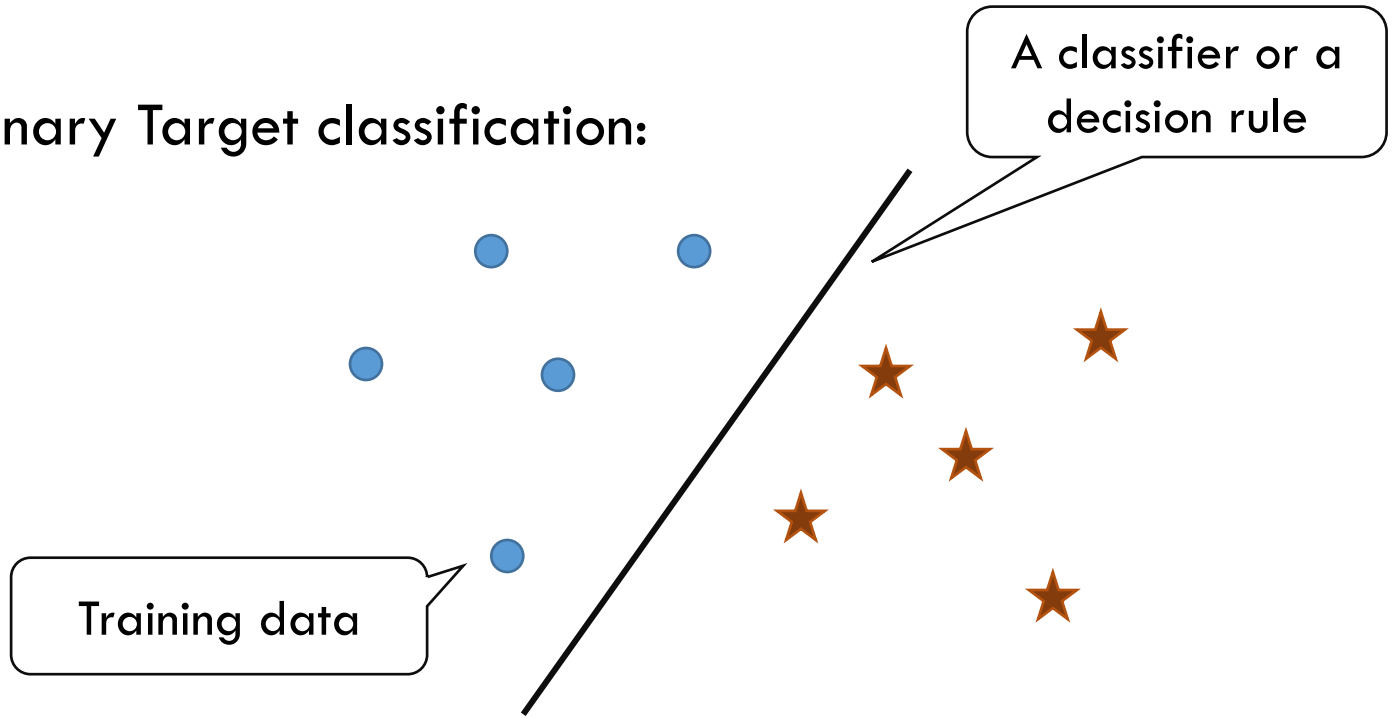
1-Rule Method (decision-stump)

Binary classification

UIC BUSINESS

- Assume we have "training" data: observations with known values for some attributes and a binary target.
- Goal: Construct a classifier that will predict the target value for new observations based on their attributes.

Binary Target classification:

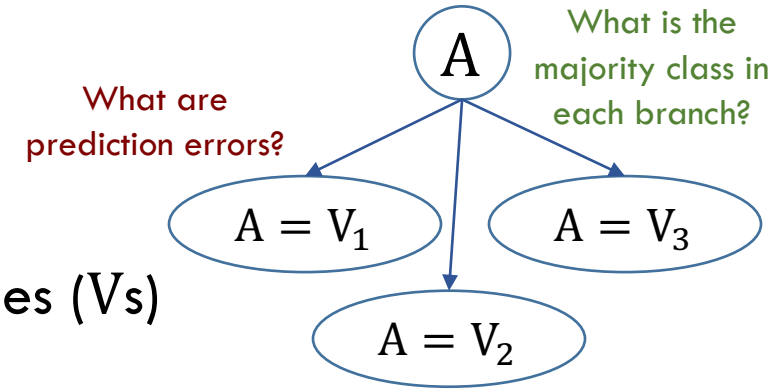


Inferring rudimentary rules: 1R or 1-rule

UIC BUSINESS

Simple method to find classification rules.

For each attribute A:



- Branch according to the attribute's values (Vs)
- Count how often each class appears in each branch
- Classification rule for each branch: choose the class (C) that occurs most often in the training data (most frequent class)
- Make rules like "if A = V then C = c"
- Calculate the error rate of each rule

Choose attribute whose set of rules yields best results (minimum error)

Example (tennis dataset)

UIC BUSINESS

- Shall I play tennis?

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
d1	sunny	hot	high	weak	no
d2	sunny	hot	high	strong	no
d8	sunny	mild	high	weak	no
d9	sunny	cool	normal	weak	yes
d11	sunny	mild	normal	strong	yes
d3	cloudy	hot	high	weak	yes
d7	cloudy	cool	normal	strong	yes
d12	cloudy	mild	high	strong	yes
d13	cloudy	hot	normal	weak	yes
d4	rain	mild	high	weak	yes
d5	rain	cool	normal	weak	yes
d6	rain	cool	normal	strong	no
d10	rain	mild	normal	weak	yes
d14	rain	mild	high	strong	no

Outlook

9 yes / 5 no

Sunny

Day	Outlook	Humid	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

2 yes/3 no

NO

Error: 2/5

Cloudy

Day	Outlook	Humid	Wind
D3	Cloudy	High	Weak
D7	Cloudy	Normal	Strong
D12	Cloudy	High	Strong
D13	Cloudy	Normal	Weak

4 yes / 0 no

Yes

Error: 0/4

Rain

Day	Outlook	Humid	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

3 yes / 2 no

Yes

Error: 2/5

Total Error: $2/5 \times 5/14 + 0/4 \times 4/14 + 2/5 \times 5/14 = 4/14$

Example of 1R

UIC BUSINESS

Evaluating Attributes in the Tennis-play Data				
	Attribute	Rules	Errors	Total Errors
1	Outlook	Sunny → no	2/5	4/14
	Set of rules	Cloudy → yes	0/4	Minimum Error
		Rainy → yes	2/5	
2	Temperature	Hot → no*	2/4	5/14
		Mild → yes	2/6	
		Cool → yes	1/4	
3	Humidity	High → no	3/7	4/14
		Normal → yes	1/7	
4	Windy	Weak → yes	2/8	5/14
		Strong → no*	3/6	

* A random choice has been made between two equally likely outcomes

Support and confidence of a rule

UIC BUSINESS

- Support = proportion of records that satisfy the rule.
- Confidence = proportion of correct predictions among records where it applies.

In other words, given the rule "If A then B" we have

Support $= P(A)$

Confidence $= P(B \mid A) = \frac{P(A \cap B)}{P(A)}$

In our example, the support and confidence of the rule "If sunny and then no" are $5/14$ and $3/5$ respectively.

Example

UIC BUSINESS

- Suppose we want to predict whether the customers will increase spending if given a loyalty card.
- What is the target variable?
- Assume we have obtained the rule

  If "has college degree" and "single" then "likely to increase"

- Assume out of 1000 customers, 100 satisfy this rule's condition. Of these 100, 75 are likely to increase spending and 25 are not likely to increase spending.
- What is the support of this rule? 100 / 1000 = 0.1
- What is the confidence of this rule? 75 / 100 = 0.75