CLASSIFICATION

SYRACUSE UNIVERSITY
School of Information Studies

HOW TO MODEL A CLASSIFICATION PROBLEM

Bank loan approval:

What is the decision to make?

What is the unit of analysis?

What attributes are helpful for classification?

WHAT IS THE DECISION TO MAKE?

What is the decision to make?

"Approve" or "deny" a loan application

The decision is saved in the target attribute

WHAT IS THE UNIT OF ANALYSIS?

The unit of analysis means an example in your data set. A classification decision will be made for each example.

For bank loan classification, an individual application is an example, which will be either approved or denied.

An individual person may not be good unit of analysis, because one person may submit multiple applications over time, and each deserves a decision.

WHAT ATTRIBUTES ARE HELPFUL FOR CLASSIFICATION?

What attributes are useful for classification?

Potentially useful attributes:

E.g., applicant's age, job title, income, credit score, amount requested

Some might be more useful than others.

Classification algorithms can rank the attributes by their contribution to classification.

SAMPLE DATA FOR BANK LOAN CLASSIFICATION

Application	Job Title	Income	Credit Score	Decision
1	teacher	50K	700	approve
2	manager	60K	300	deny

Each row is an example.
Each column is an attribute.
The last attribute is the decision to make, the target attribute.

HOW TO TEACH A COMPUTER TO CLASSIFY?

Step 1: Collect training data.

E.g., a collection of past loan decisions made by financial experts

Step 2: Use a machine-learning algorithm to build a classifier based on relevant variables.

Step 3: Apply the classifier to new data.

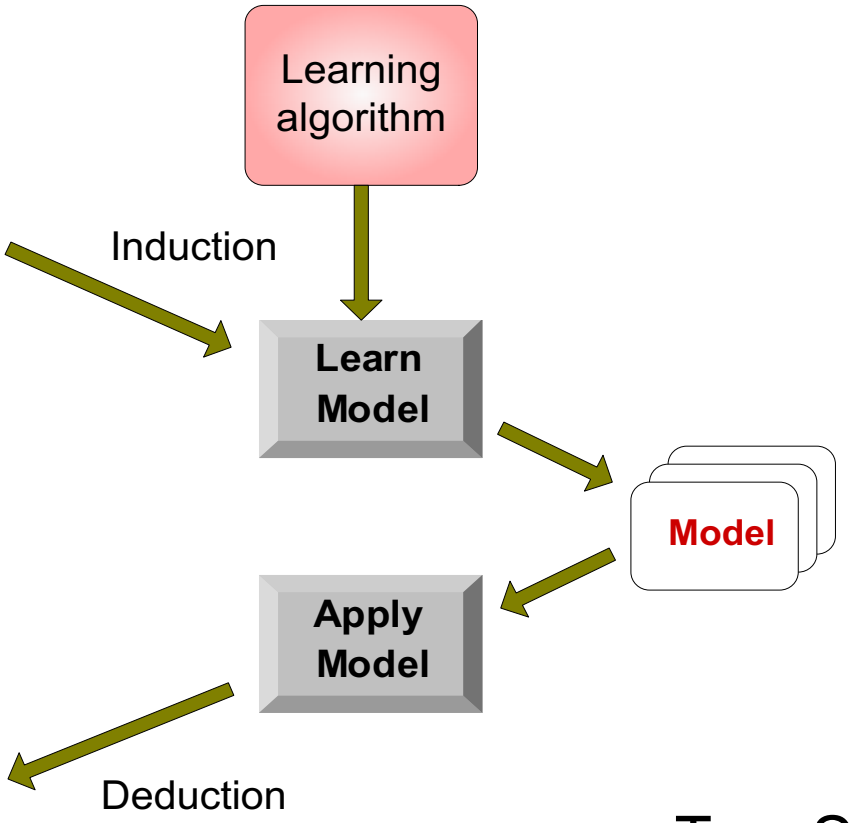
ILLUSTRATING CLASSIFICATION TASK

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Two Steps

ARE WE DONE?

No. Prediction models need maintenance.

What if an approved loan defected?

- Add defective loans to the “deny” pool and retrain the model

What if a denied application was approved by another bank and performed well?

- No good solution without data sharing

ANOMALY DETECTION

Detect significant deviations from normal behavior

Applications:

- Credit card fraud detection
- Network intrusion detection

Can be modeled as classification problem

- Classify each transaction as fraud or not