Classification

Basic concepts

Huiping Cao

Examples of Classification Task

- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil
- Categorizing news stories as finance, weather, entertainment, sports, etc

Binary class, multi-class

Definition

- Given a collection of records
  - Each record contains a set of attributes
  - One of the attributes is the class
- Find a model/function f:
  - each attribute set $\rightarrow$ class label
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - Training set: build the model
  - Test set: validate the models

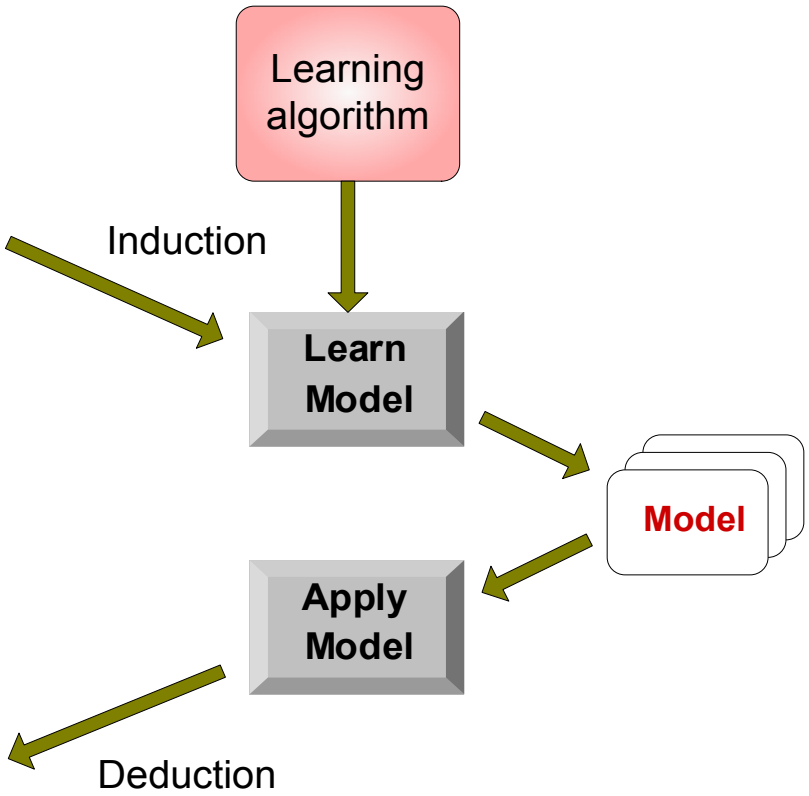
Illustrating Classification Task

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



Evaluation

- Confusion matrix

		Predicted Class	
		Class=1	Class=0
Actual Class	Class=1	f_{11}	f_{10}
	Class=0	f_{01}	f_{00}

- Performance metric

$$Accuracy = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

$$Error\ rate = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

Desirable classifier: high accuracy, low error rate

Properties

- A classification model
  - Predictive model
  - Descriptive model

Classical Classification Techniques

- Decision Tree Based Methods
- Nearest Neighbor (NN) Classifiers
- Bayesian Classifiers
- Support Vector Machine (SVM) Classifiers
- Logistic Regression classifier
- Neural Network classifier
- Ensemble Methods
- Class Imbalance Problem