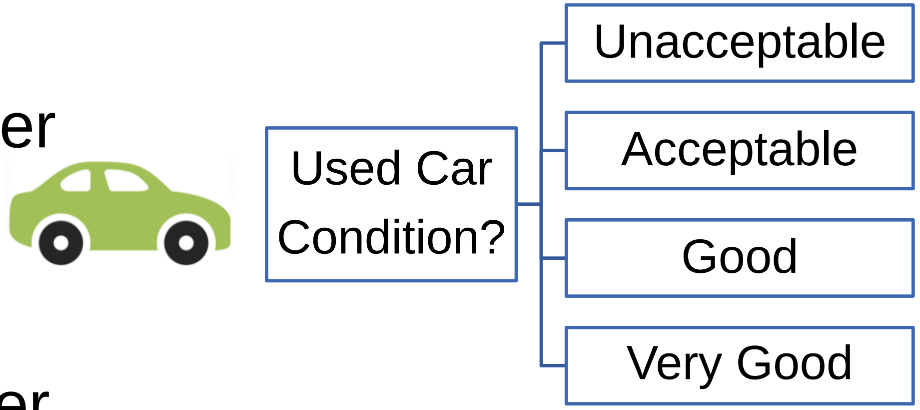
Week 3: Classification

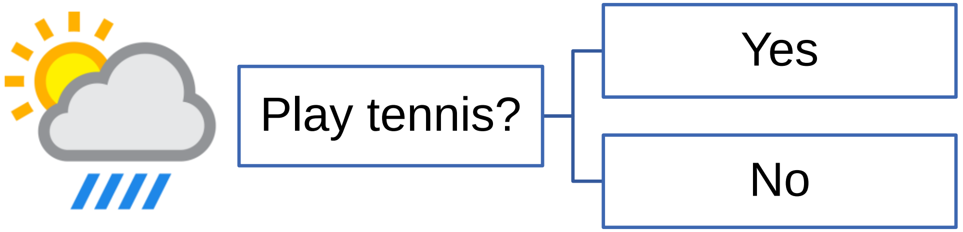
Outline

- Classification definition
- Problem framing
- Classification techniques
- Parametric and non-parametric methods
- Performance evaluation

Classification Task

- Assign **one** or **more class labels** to each instance of data
  - **Used car**: For each car, assign whether the car would be unacceptable, acceptable, good or very good
  - **Weather**: For each day, assign whether to play tennis or not (yes or no)





Data: Terminologies

- **Instance**, **example**, **record**: Each item of data with all of its info
  - Person, transaction, day, car, etc.
  - Represented as rows of a table
- **Attribute**, **field**, **feature**, or **variable**: Items of information collected about each instance
  - Income, title, date, condition, etc.
  - Type can be categorical or continuous (numeric)
  - Represented as the columns of a table

Data: Magazine Sales Example

- Each instance (row) is a magazine subscription
- Each field or attribute (column) is data about the subscription

attribute

instance

Client number	Client name	Date of birth	Income	Credit	Car owner	House owner	Address	Date of subscription	Magazine
23003	Johnson	13-04-76	18500	17800	no	no	1 Downing St.	15-04-94	car
23003	Johnson	13-04-76	18500	17800	no	no	1 Downing St.	21-06-93	music
23003	Johnson	13-04-76	18500	17800	no	no	1 Downing St.	30-05-92	comic
23009	Clinton	20-10-71	36000	26600	yes	no	2 Boulevard	NULL	comic
23013	King	15-08-70	NULL	NULL	NULL	NULL	3 High Road	30-02-95	sports
23003	Johnson	13-04-76	18500	17800	no	no	1 Downing St.	20-12-94	house

Labeled Data

- Where does labeled data come from?
  - Humans annotate a set of training examples (expensive)
  - Labeled data comes from the past where the results are known

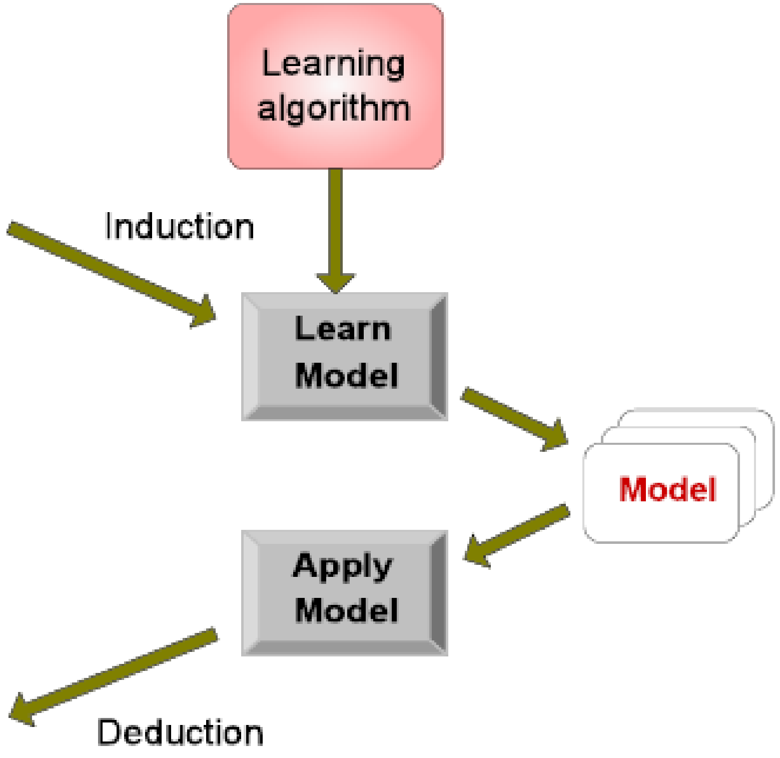
Classification: Definition

- Given a collection of records (training set)
  - Each record contains a set of attributes, one of the attributes is the class label
- Find a model for the class attribute as a function of the values of other attributes
- **Goal:** Previously unseen records can then be assigned class as accurately as possible
  - Performance is given as the accuracy of the labeling
  - A test set is used to determine the accuracy of the model

Classification: Definition

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

Problem Framing

- **Business application of the classification task**: Given a new customer, predict what types of magazines the person would like. If there are magazine types that the customers like but are not currently subscribed, design a marketing campaign to target that type of magazine subscription for those customers.

Problem Framing

- Frame the classification task
  - Given the attributes of a person, predict whether the person would like car magazines or not **(binary classification)**

class label

Age	Income	Credit	Car owner	House owner	Car magazine
20	18.5	17.8	0	0	1
25	36.0	26.6	1	0	0

Labeled Data

Problem Framing

- Frame the classification task
  - Given the attributes of a person, predict one type of magazine the person would like **(multiclass classification)**

class label

Age	Income	Credit	Car owner	House owner	Magazine type
20	18.5	17.8	0	0	car
25	36.0	26.6	1	0	comic

Labeled Data

Problem Framing

- Frame the classification task
  - Given the attributes of a person, for each type of magazine (i.e., car, music, comic, sports, house), predict whether the person would like that type **(multilabel classification)**

class labels

Labeled Data

Age	Income	Credit	Car owner	House owner	Car magazine	House magazine	Sports magazine	Music magazine	Comic magazine
20	18.5	17.8	0	0	1	1	0	1	1
25	36.0	26.6	1	0	0	0	0	0	1

Class Activity

Suppose you are building an application for a bank that will be used to help make lending decisions. The application makes a recommendation to approve or deny a loan based on a set of input data related to the loan application. Frame the classification task.

Class Activity

Suppose you are building an application for a bank that will be used to help make lending decisions. The application makes a recommendation to approve or deny a loan based on a set of input data related to the loan application. Frame the classification task.

Given the attributes of a loan application, predict whether to approve or deny a loan **(binary classification)**

Classification: More Examples

- 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 or sports

Classification Techniques

- Learning algorithms to build a classification model
  - Instance-based learning (k-nearest neighbor)
  - Decision trees
  - Naïve Bayes and Bayesian belief networks
  - Support vector machines

Parametric ML Algorithms

- Algorithms that simplify the function to a known form (make strong assumptions about the data)
- Has a fixed number of parameters
- The algorithms involve two steps:
  1) Select a form for the function
  2) Learn the coefficients for the function from the training data

Parametric ML Algorithms

Benefits	Limitations
• **Simpler**: Easier to understand and interpret results • **Speed**: Very fast to learn from data • **Less Data**: Do not require as much training data and can work well even if the fit to the data is not perfect	• **Constrained**: By choosing a functional form these methods are highly constrained to the specified form • **Limited Complexity**: More suited to simpler problems • **Poor Fit**: In practice the methods are unlikely to match the underlying mapping function

Non-Parametric ML Algorithms

- Algorithms that do not make strong assumptions about the data (free to learn any functional form from training data)
- Has a flexible number of parameters
- Good when there is **a lot of data**, **no prior knowledge**, and we **do not want to worry too much about choosing just the right features**

Non-Parametric ML Algorithms

Benefits	Limitations
• **Flexibility**: Capable of fitting a large number of functional forms • **Power**: No assumptions (or weak assumptions) about the underlying function • **Performance**: Can result in higher performance models for prediction	• **More data**: Require a lot more training data to estimate the mapping function • **Slower**: A lot slower to train (have far more parameters to train) • **Overfitting**: More of a risk to overfit training data and harder to explain why specific predictions are made

Parametric versus Non-Parametric

Parametric ML Algorithms

- Linear regression
- Logistic regression
- Naïve Bayes
- Perceptron

Non-Parametric ML Algorithms

- K-nearest neighbor
- Decision tree

Performance Evaluation

- Classification model performance evaluation
  - **Accuracy**: Predicting class label correctly (measure predictive capability)
  - **Speed**
    - Time to construct model (training time)
    - Time to use the model (classification/prediction time)
  - **Robustness**: Handling noise and missing values
  - **Interpretability**: Understanding and insight provided by the model

Performance Evaluation

Labeled Data

id	x1	x2	class label
1	1	2	yes
2	3	5	no
3	3	7	yes
4	8	8	no
5	3	6	yes
6	8	7	no

Split
Train = 50%
Test = 50%

Train Set

id	x1	x2	class label
1	1	2	yes
2	3	5	no
3	3	7	yes

Trained model is applied to test set – class labels are removed first

Test Set

id	x1	x2	class label
4	8	8	no
5	3	6	yes
6	8	7	no

Compare actual labels to predicted labels

id	x1	x2	predicted label
4	8	8	no
5	3	6	no
6	8	7	yes

Performance Evaluation

- Confusion matrix
  - Compare actual labels to predicted labels

	Predicted Class	
Actual Class		Class = yes	Class = no
	Class = yes	**TP**	**FN**
	Class = no	**FP**	**TN**

True positive (TP) = Correctly identified
False positive (FP) = Incorrectly identified
True negative (TN) = Correctly rejected
False negative (FN) = Incorrectly rejected

Performance Evaluation

- Confusion matrix

N = 3	Predicted Class		Total
Actual Class		Class = yes	Class = no	
	Class = yes	**0 (TP)**	**1 (FN)**	1
	Class = no	**1 (FP)**	**1 (TN)**	2
Total		1	2	3

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Accuracy
= (0 + 1) / (0 + 1 + 1 + 1)
= 1 / 3
= 0.33

Summary

- **“No Free Lunch”** theorem
  - No one algorithm works best for every problem
  - Many factors at play (e.g., size and structure of data set)
- What should you do?
  - Try many different algorithms for your problem
  - Use a hold-out **test set** to evaluate performance and select the winner