Comp 90049 Intro to ML

Pei-Yun Sun (Stats, Data Science, CS)

Email: pssun@unimelb.edu.au

Concept: - what we're trying to predict / understand
- Output of the system
- label / classes (supervised learning)

Instance: single exemplar from data (consist of attribute values)

Attribute: single measurement of some aspect of an instance (features)

(i) Building a system that guesses what the weather (temperature, precipitation, etc.) will be like tomorrow

Concept: Weather: quantity e.g. temperature, precipitation, humidity, UV index ...

Instance: A day

Attribute: data from previous days

S: Regression (numeric)

↑ e.g. temperature

Generalisation: might work better in some cities than others

better if included geographic info

new features → weather patterns interact with

(ii) Predicting products that a customer would be interested in buying, based on other purchases that customer has previously made

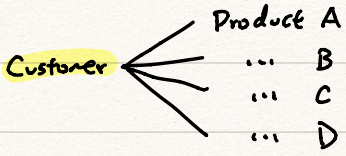
(a) Customer - Product pairs

Concept: Interested / Not interested

Instance: Pair

S: classification (I/NI)

(b)



Concept: Products

Instance: Customer

Attributes: name, age, shopping log, gender, ...

US: clustering / association rule mining (attr of prod → buy / not)

(groups of customers)

Generalisation: - Customer model in one country might not generalise to other countries

- if it learns everyday shopping patterns

⇒ may not work for outlier situations e.g. holiday purchasing

(iii) Skin cancer screening test

Concept: Cancer / Not cancer (binary)

Instance: patient

Attribute: result of blood test, images from skin, reports, observed syndromes, ...

S: (binary) Classification

Generalisation:

training data often have biases, e.g. skin cancer risk often increase with ages

⇒ correlated in traing set

- Good: the model could learn age $\xrightarrow{\text{predicts}}$ skin cancer
- Bad: if too dependent on age ⇒ may not work well on young patients if there were very few instances of younger patients

(iv) Automatically identifying the author of a given piece of literature

Concept : Author

(a) A single literature & fixed set of authors

S : Classification

(b) Open-domain : potentially anybody could have written it

US : Clustering

(c) Plagiarism detection

Outlier detection / semi-supervised learning.

Classification not effective, like (b)

instance : literature
attributes : words, grammatical structure, language, year of publish language ...

Generalisation :

- Classification case : the model could reliably classify new examples from each author, but need to train a new model for a new set of authors
- Plagiarism / outlier : might be able to generalise to new authors (learn general rules to decide "what's outlier")

(v) Finding the best burrito in the United States of America

Concept : Best restaurant (a) / product (b)

Instance : restaurant (a) / product (b)

Attribute : (a) restaurant : ranking of restaurant, customers compliments (positive feedback)

(b) product : ingredients, sauces, spices

S (classification)

Generalisation : depend on attributes

"best burrito" model might also be able to pick the "best pizza" but not "best coffee" (people care about other factors e.g. noise level, atmosphere ...)

5. What kinds of assumptions might a machine learning model make when tackling these problems?

① Concept is actually related to the attributes (obvious!)

- We only include attributes we think are likely to predict the concept
- e.g. you won't use attributes like "patient's favorite song" as an attribute for skin cancer detection

BUT song —(good predictor)→ age —(risk factor)→ skin cancer

↑ might be a good predictor

② Each model makes assumptions about the ways the attributes can relate to concepts

e.g.

(i) treat all attributes as independent predictors ✓

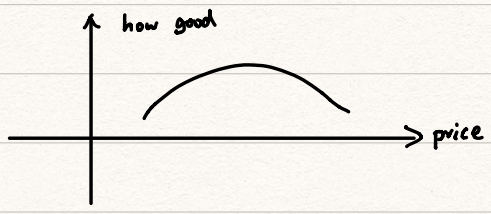
Allow predictors to interact ⇒ could lead to an overly complex model if there are many attributes to start with

(ii) Numerical attributes:

- Generally expect linear / monotonic relationships between attributes & concepts

good simplifying assumption but limits what the model can learn

e.g. "best burrito" (U-shape)



Supervised : - instances labelled with classes (training data)

- Classify / predict instances in test data (no labels)

Un-S : Not based on labelled traiiy data (ignore)

↑ find hidden patterns / groups