Comp 90049   Intro to ML

✓ Pei-Yun Sun (Stats, Data Science, CS)
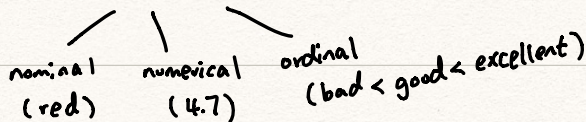
✓ Email: pssun @ unimelb.edu.au

① **Concept** : - what we're trying to learn (target)

- Output of system

- Labels / classes (Supervised Learning)

② **Instance** : Single exemplar from data (consist of attribute values)

③ **Attribute** : Single measurement of some aspect of an instance

nominal   numerical   ordinal
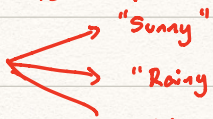(red)   (4.7)   (bad < good < excellent)

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

⎧ Concept : Weather (e.g. quanity: temperature)

⎨ Instance : A day

⎩ Attribute : data from previous days

Supervised : Regression (numeric : temperature)

Classification ⟵ → "Sunny"
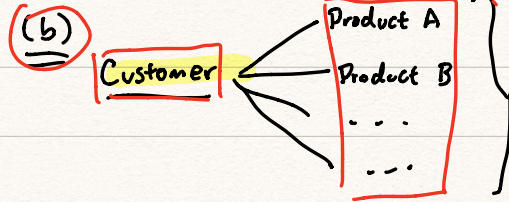   → "Rainy"   (nominal)
               . . .

(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

Supervised: Classification (binary)

(b)

Customer → Product A
Customer → Product B
      ⋯
      ⋯

Concept : Products

Instance : Customer

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

Unsupervised: Clustering / association rule mining

↑ find similar customers

item attributes ⟹ Customer "A" Buying

(iii)   Skin cancer screening test

Concept: Cancer / Not cancer (binary)

Instance: Patient

Attribute: Skin color, age, symptoms, place used to stay, ...

S: classification (binary)

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

Concept: Author

Output: one of the authors

(10)

(a) A single literature & fixed set of authors

S: classification

(b) Open-domain : potentially anybody (any author)

Unsupervised: Clustering

compare → "Author" (prediction)

(c) Plagiarism detection

Outlier detection / semi-supervised learning   ← later in the course

Instance : Literature

Attributes : Title, language, publish year, style, length, words statistics

↰ use both <u>labelled</u> & <u>unlabelled</u> data

(author known) + (don't know author)

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

(a)   (b)

Concept : best restaurant / burrito (product)

Instance : (a) restaurant   (b) burrito

Attribute : (a) restaurant : sale # of burrito , (comments) reviews , environment , price , location

(b) burrito : taste , size , sauces , spices
(ingredients)

① System :   Ranked   Score   System : Best : B1.
              1. B1   50   OR
              2. B2   30
              3. B3   20

Classification ("cat" / "dog") ⟍ labels
Regression (house price)   /

Supervised : — instances labelled with classes (training data)
(labels)
              — Classify / predict instances in test data

Unsupervised : — Not based on labelled data
(no labels)
              — Clustering , association rule, PCA , EM , ...

Supervised : Classification (nominal) → Restaurant "A"