Machine Learning W3 Tutorial

COMP30027 | Sandy Luo

Overview

Probability

Joint, conditional, marginal probabilities, PDF...

Entropy

Calculation

Probability Lab

code!

01

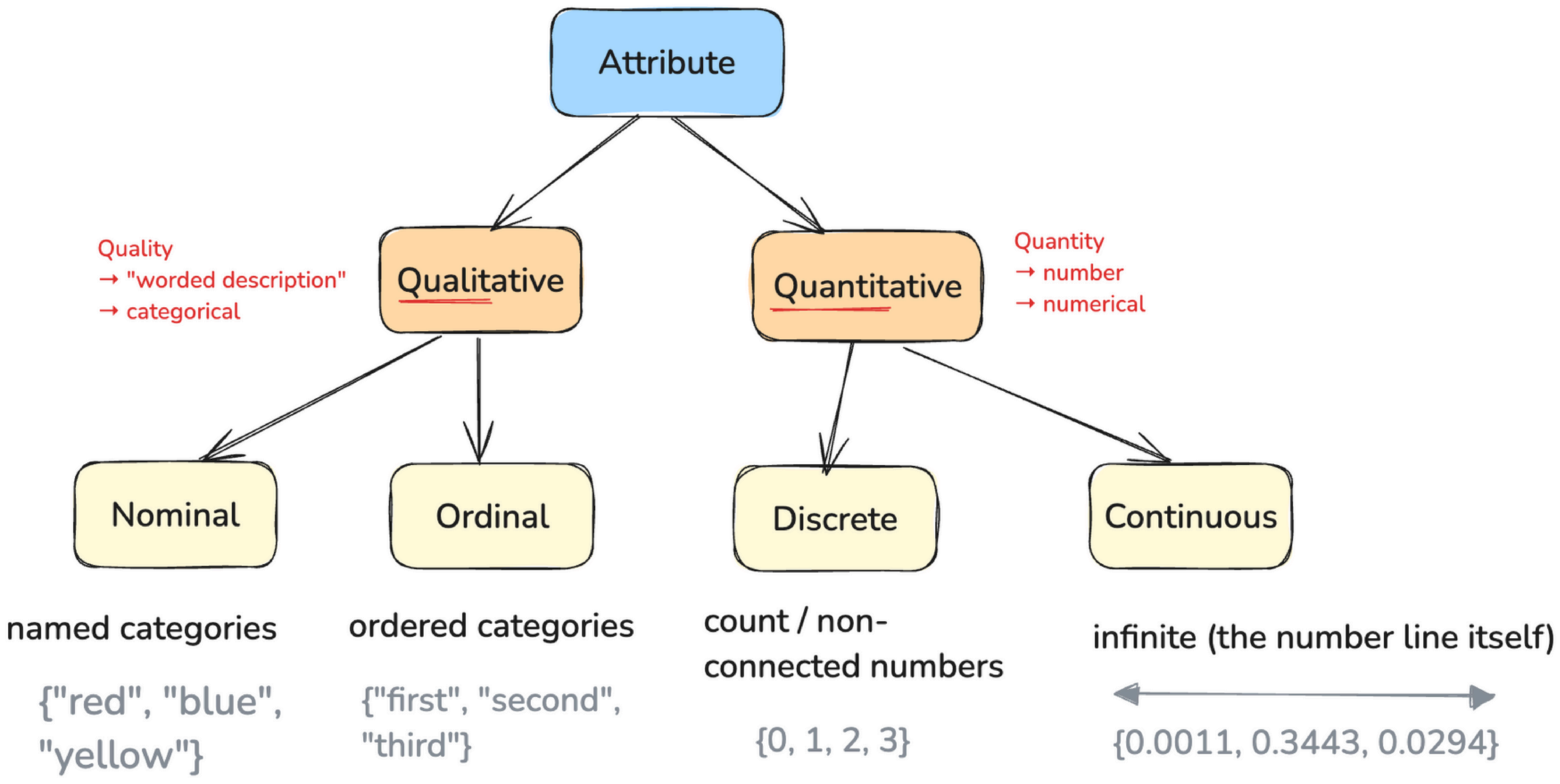
Notations

02

Probability Distributions

Probability

Attributes


Attribute
Quality
→ "worded description"
→ categorical
Qualitative
Quantitative
Quantity
→ number
→ numerical
Nominal
Ordinal
Discrete
Continuous
named categories
{"red", "blue", "yellow"}
ordered categories
{"first", "second", "third"}
count / non-connected numbers
{0, 1, 2, 3}
infinite (the number line itself)
{0.0011, 0.3443, 0.0294}

Notation

- $P(x)$ = probability of event x **Marginal probability**
- $P(x,y)$ = probability of both x and y occurring **Joint probability**
- $P(x|y)$ = probability x occurring, given y **Conditional probability**

Notation

- P(x) = probability of event x

P(x) = <u>Prior</u> probability
→ probability of x without other info.

- P(x,y) = probability of both x and y occurring

- P(x|y) = probability x occurring, given y

P(x|y) = <u>Posterier</u> probability
→ probability of x given y data

Q1

Approximately 1% of women aged between 40 and 50 have breast cancer. 80% of mammogram screening tests detect breast cancer when it is there. 90% of mammograms DO NOT show breast cancer when it's NOT there. Use this information to complete the following table with:

- the **joint probabilities** P(Cancer,Test) for each possible pair of cancer status and test result
- the **conditional probabilities** P(Test|Cancer) for each test result given cancer status

Cancer	Test	Joint prob.	Conditional prob.
Yes	Positive		80%
Yes	Negative		
No	Positive		
No	Negative		90%

Q1

P(c) = 0.01

Approximately 1% of women aged between 40 and 50 have breast cancer. 80% of mammogram screening tests detect breast cancer when it is there. 90% of mammograms DO NOT show breast cancer when it's NOT there. Use this information to complete the following table with:

- the **joint probabilities** P(Cancer,Test) for each possible pair of cancer status and test result
- the **conditional probabilities** P(Test|Cancer) for each test result given cancer status

Cancer	Test	Joint prob.	Conditional prob.
Yes	Positive	P(c=T,t=P) = P(c=T)P(t=P|c=T) = 0.01*0.8 = 0.008	80% P(t=P|c=T) = 0.8
Yes	Negative	P(c=T,t=N) = P(c=T)P(t=N|c=T) = 0.01*0.2 = 0.002	100% - 80% = 20%
No	Positive	P(c=N,t=P) = P(c=N)P(t=P|c=N) = 0.99*0.1 = 0.099	100% - 90% = 10%
No	Negative	P(c=N,t=N) = P(c=N)P(t=N|c=N) = 0.99*0.9 = 0.891	90% P(t=N|c=N) = 0.9

...given cancer = True

...given cancer = False

P(x,y) = P(y)*P(x|y)

Q2

Given the table above, compute the **marginal probability** of a positive result in the mammogram screening test.

Q2

Given the table above, compute the **marginal probability** of a positive result in the mammogram screening test.

Marginal probability of positive result → $P(t=P)$

$= $ sum over all $P(c,t)$ where $t=P$

$= P(c=T,t=P) + P(c=N,t=P)$

$= 0.008 + 0.099$

$= 0.107$

Q3

Suppose a woman in this age group receives a positive test result. Compute the **conditional probability** P(Cancer == Yes|Test == Postive).

$$\underbrace{\boxed{P(A|B)}}_{\text{posterior}} = \underbrace{\boxed{P(A)}}_{\text{prior}} \times \frac{\boxed{P(B|A)}\ \text{likelihood}}{\boxed{P(B)}\ \text{marginal}}$$

Q3

Suppose a woman in this age group receives a positive test result. Compute the **conditional probability** P(Cancer == Yes|Test == Postive).

- Bayes rule: $P(y|x) = P(x|y)*P(y) / P(x)$

- $P(c=T|t=P) = P(t=P|c=T) * P(c=T) / P(t=P)$
  $= 0.8 * 0.01 / 0.107$
  $= 0.075$

Entropy

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

- Measures amount of uncertainty in dataset
- More uncertain = high, less uncertain = low

Q6

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Compute the entropy of a random letter generator which can generate any of the 26 English letters (a-z), each with equal probability.

Q6

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Compute the entropy of a random letter generator which can generate any of the 26 English letters (a-z), each with equal probability.

- Probability of generating any character = 1/26
- H(X)= - 26 * 1/26 * log_2(1/26)
  - = - log_2(1/26)
  - = log_2(26)
  - ~= 4.7

Q7

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Compute the entropy of the actual probability distribution of letters in English text, using the empirical probability distribution computed earlier.

Q7

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Compute the entropy of the actual probability distribution of letters in English text, using the empirical probability distribution computed earlier.

```
entropy = -np.sum([p * np.log2(p) for p in letter_probabilities if p > 0])
```

Q7

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

You should get a lower value in Q7 than Q6. Why?

- When is entropy (measure of uncertainty) at it's highest?
- When is entropy at it's lowest?

- High entropy = high uncertainty = option probabilities similar → highest = uniform distribution
- Low entropy = low uncertainty = options very skewed / deterministic