Lecture(10):Bayesian Inference

Bayesian models in cognitive science

- Vision
- Motor control
- Memory
- Language
- Inductive learning and reasoning….

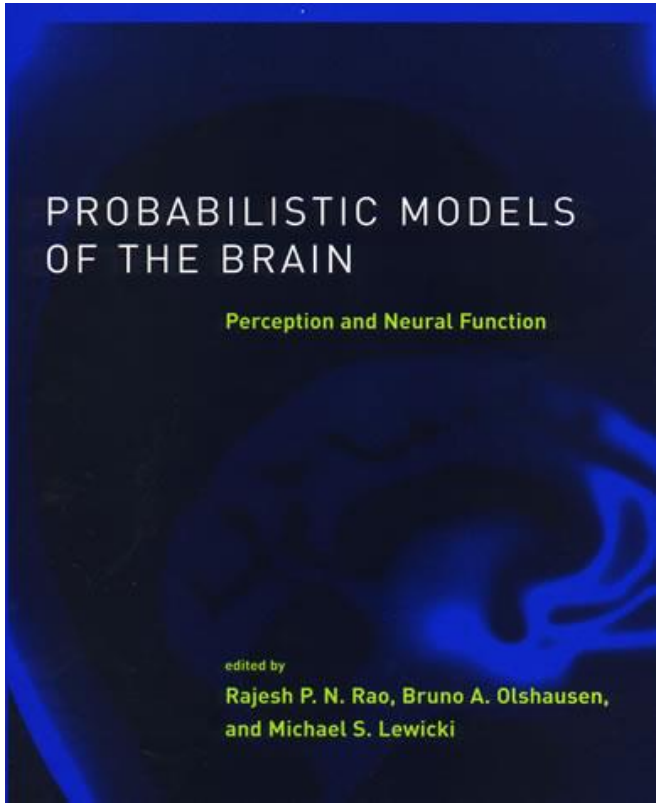
and ...

Applications in:

- Data mining
- Robotics
- Signal processing
- Bioinformatics
- Text analysis (inc. spam filters)
- and (increasingly) graphics!

Bayesian reasoning is ...

A theory of mind



Inductive reasoning

Input:

Cows can get Hick's disease.
Gorillas can get Hick's disease.

(premises)

All mammals can get Hick's disease.

(conclusion)

Task: Judge how likely conclusion is to be true, given that premises are true.

Inferring causal relations

Input:

	Took vitamin B23	Headache
Day 1	yes	no
Day 2	yes	yes
Day 3	no	yes
Day 4	yes	no
. . .	. . .	. . .

Does vitamin B23 cause headaches?

Task: Judge probability of a causal link given several joint observations.

Bayesian Learning

- ***Features of Bayesian learning methods:***
- Each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct.
- – This provides a more flexible approach to learning than algorithms that completely eliminate a hypothesis if it is found to be inconsistent with any single example.
- • Prior knowledge can be combined with observed data to determine the final probability of a hypothesis. In Bayesian learning, prior knowledge is provided by asserting
- – a prior probability for each candidate hypothesis, and
- – a probability distribution over observed data for each possible hypothesis.
- New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.
- • Even in cases where Bayesian methods prove computationally intractable, they can provide a standard of optimal decision making against which other practical methods can be measured

Difficulties with Bayesian Methods

• Require initial knowledge of many probabilities
– When these probabilities are not known in advance they are often estimated based on background knowledge, previously available data, and assumptions about the form of the underlying distributions.

• Significant computational cost is required to determine the Bayes optimal hypothesis in the general case (linear in the number of candidate hypotheses).

– In certain specialized situations, this computational cost can be significantly reduced.

Bayes Theorem

In machine learning, we try to determine the ***best hypothesis*** from some hypothesis space H, given the observed training data D.

• In Bayesian learning, the ***best hypothesis*** means the ***most probable*** hypothesis, given the data D plus any initial knowledge about the prior probabilities of the various hypotheses in H.

• Bayes theorem provides a way to calculate the probability of a hypothesis based on its prior probability, the probabilities of observing various data given the hypothesis, and the observed data itself.

Bayes Theorem

P(h) is ***prior probability of hypothesis h***
– P(h) to denote the initial probability that hypothesis h holds, before observing training data.

P(D) is ***prior probability of training data D***
– The probability of D given no knowledge about which hypothesis holds

P(h|D) is ***posterior probability of h given D***
– P(h|D) is called the ***posterior probability*** of ***h,*** because it reflects our confidence that ***h*** holds after we have seen the training data ***D.***
– The posterior probability P(h|D) reflects the influence of the training data ***D,*** in contrast to the prior probability P(h), which is independent of D.

P(D|h) is ***posterior probability of D given h***
– The probability of observing data ***D*** given some world in which hypothesis ***h*** holds.
– Generally, we write ***P(xly)*** to denote the probability of **event x** given **event y.**

Bayes Theorem

In ML problems, we are interested in the probability P(h|D) that h holds given the observed training data D.

• Bayes theorem provides a way to calculate the posterior probability P(h|D), from the prior probability P(h), together with P(D) and P(D|h).

Bayes Theorem: $$P(h \mid D) = \frac{P(D \mid h)\, P(h)}{P(D)}$$

P(h|D) increases with P(h) and P(D|h) according to Bayes theorem.

• P(h|D) decreases as P(D) increases, because the more probable it is that D will be observed independent of h, the less evidence D provides in support of h.

Bayes Theorem - Example

Sample Space for events A and B

A holds	T	T	F	F	T	F	T
B holds	T	F	T	F	T	F	F

$P(A) = 4/7 \quad P(B) = 3/7 \quad P(B|A) = 2/4 \quad P(A|B) = 2/3$

Is Bayes Theorem correct?

$P(B|A) = P(A|B)P(B) / P(A) = (2/3 * 3/7) / 4/7 = 2/4 \rightarrow$
CORRECT
$P(A|B) = P(B|A)P(A) / P(B) = (2/4 * 4/7) / 3/7 = 2/3 \rightarrow$
CORRECT

Maximum A Posteriori (MAP) Hypothesis, hMAP

• The learner considers some set of candidate hypotheses H and it is interested in finding the ***most probable hypothesis*** h → H given the observed data D

• Any such maximally probable hypothesis is called a ***maximum a posteriori (MAP) hypothesis hMAP***.

• We can determine the MAP hypotheses by using Bayes theorem to calculate the posterior probability of each candidate hypothesis.

$$
\begin{aligned}
h_{MAP} &\equiv \underset{h \in H}{\operatorname{argmax}} P(h|D) \\
&= \underset{h \in H}{\operatorname{argmax}} \frac{P(D|h)P(h)}{P(D)} \\
&= \underset{h \in H}{\operatorname{argmax}} P(D|h)P(h)
\end{aligned}
$$

Example - Does patient have cancer or not?

The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present.

• Furthermore, .008 of the entire population have cancer.

P(cancer) = .008 P(notcancer) = .992

P(+|cancer) = .98 P(-|cancer) = .02

P(+|notcancer) = .03 P(-|notcancer) = .97

• A patient takes a lab test and the result comes back positive.

P(+|cancer) P(cancer) = .98 * .008 = .0078

P(+|notcancer) P(notcancer) = .03 * .992 = .0298 →*hMAP is notcancer*

• Since P(cancer|+) + P(notcancer|+) must be 1

P(cancer|+) = .0078 / (.0078+.0298) = .21

P(notcancer|+) = .0298 / (.0078+.0298) = .79

Naïve Bayes Classifier

Classifiers

Where in the catalog should I place this product listing?

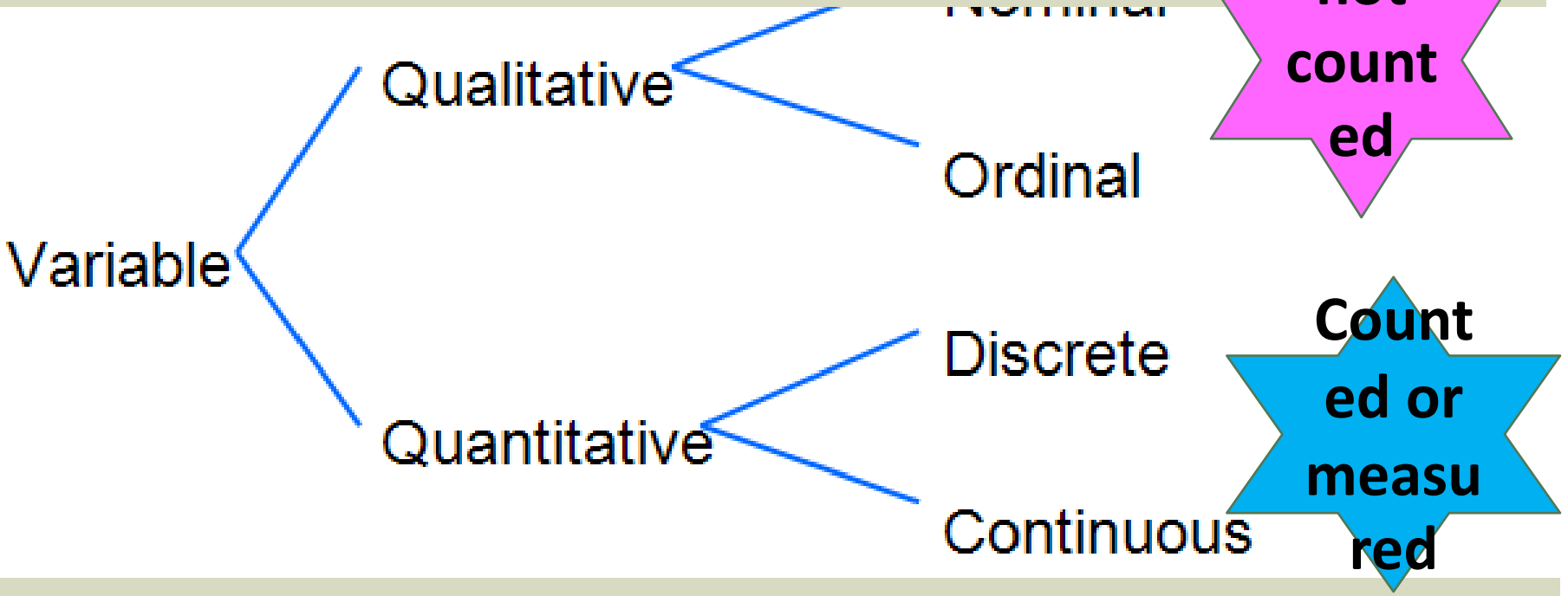
Is this email spam?

Will the customer buy the product?

- Classification
  - Assign labels to objects.
  - Usually supervised - training dataset of preclassified observations
- Commonly used classifiers
  - Naïve Bayes
  - Decision Trees
  - Logistic Regression

Types of Variables

Qualitative : a broad category for any variable that can't be counted (i.e. has no numerical value). Nominal and ordinal variables fall under this umbrella term.



Quantitative : A broad category that includes any variable that **can be counted**, or **has a numerical value** associated with it. Examples of variables that fall into this category include discrete variables and Continuous variables.

Types of Variables

"named", i.e. classified into one or more qualitative categories or description

(data that are counted)

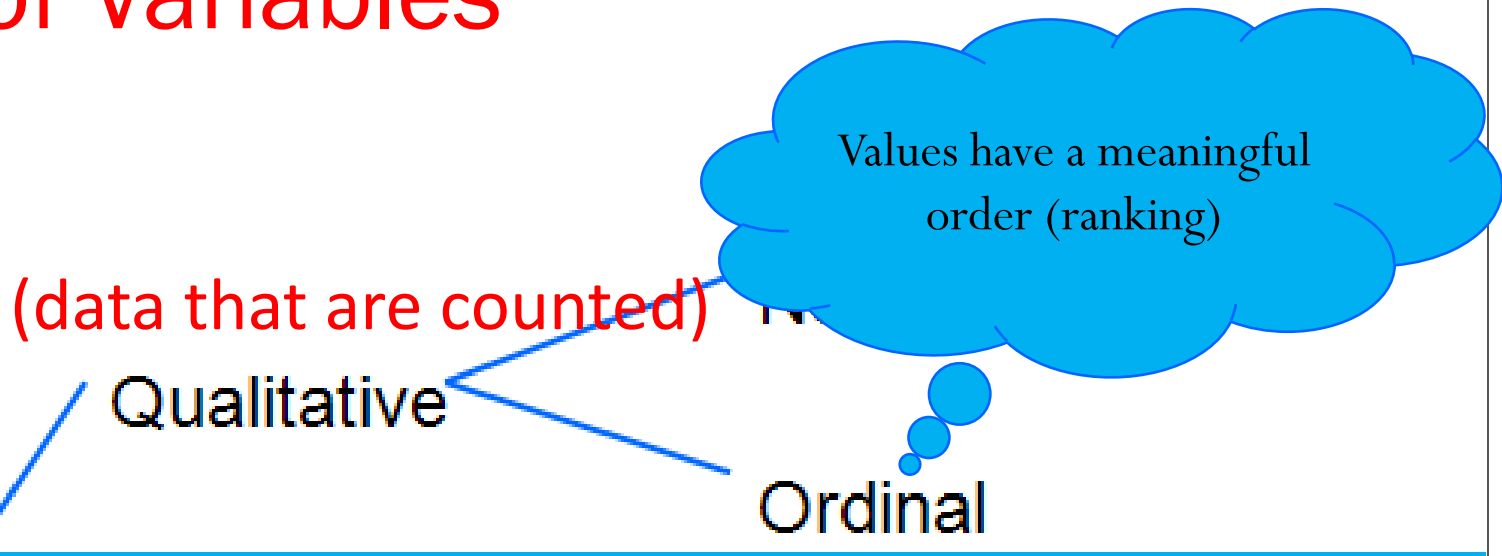
Nominal

Qualitative

In medicine, nominal variables are often used to describe the patient. Examples of nominal variables might include:

- Gender (male, female)
- Eye color (blue, brown, green, hazel)
- Surgical outcome (dead, alive)
- Blood type (A, B, AB, O)

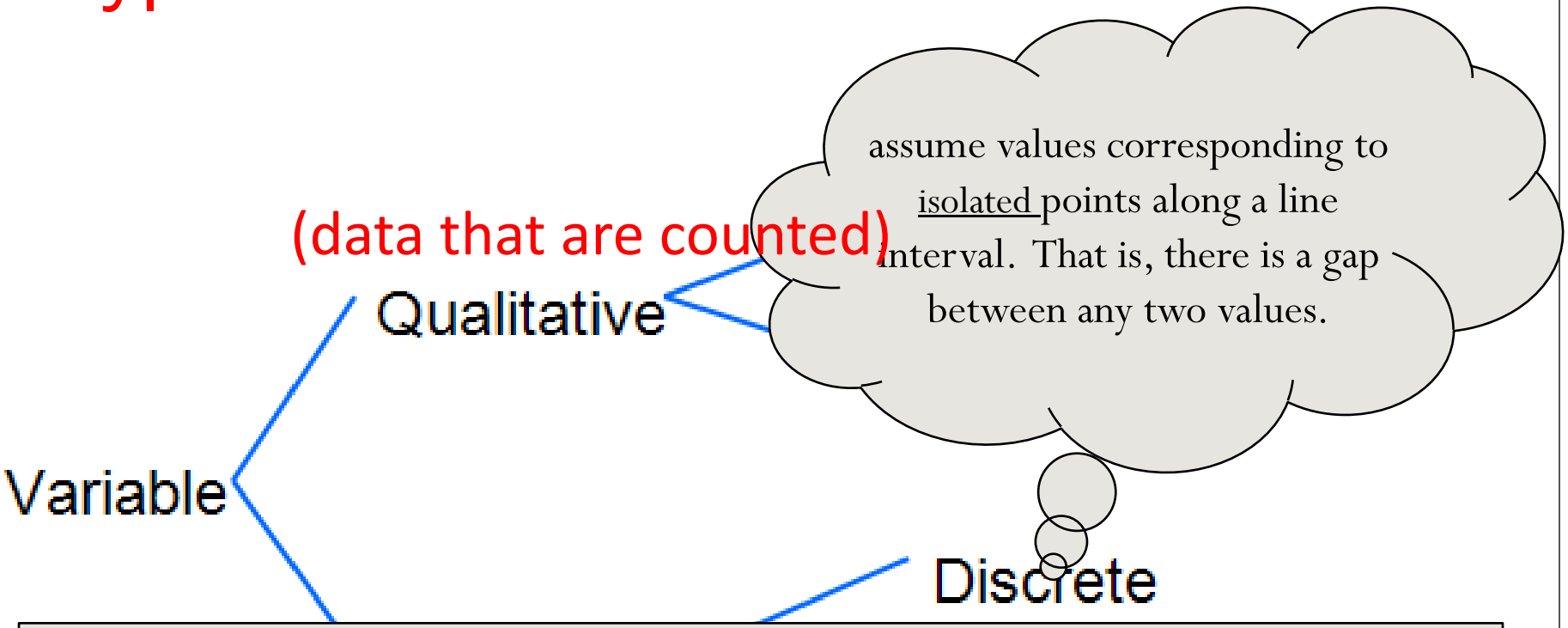
Types of Variables



In medicine, ordinal variables often describe the patient's characteristics, attitude, behavior, or status. Examples of ordinal variables might include:

- Stage of cancer (stage I, II, III, IV)
- Education level (elementary, secondary, college)
- Pain level (mild, moderate, severe)
- Satisfaction level (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied)
- Agreement level (strongly disagree, disagree, neutral, agree, strongly agree)

Types of Variables



Discrete Variables that have constant, equal distances between values, but the zero point is arbitrary.

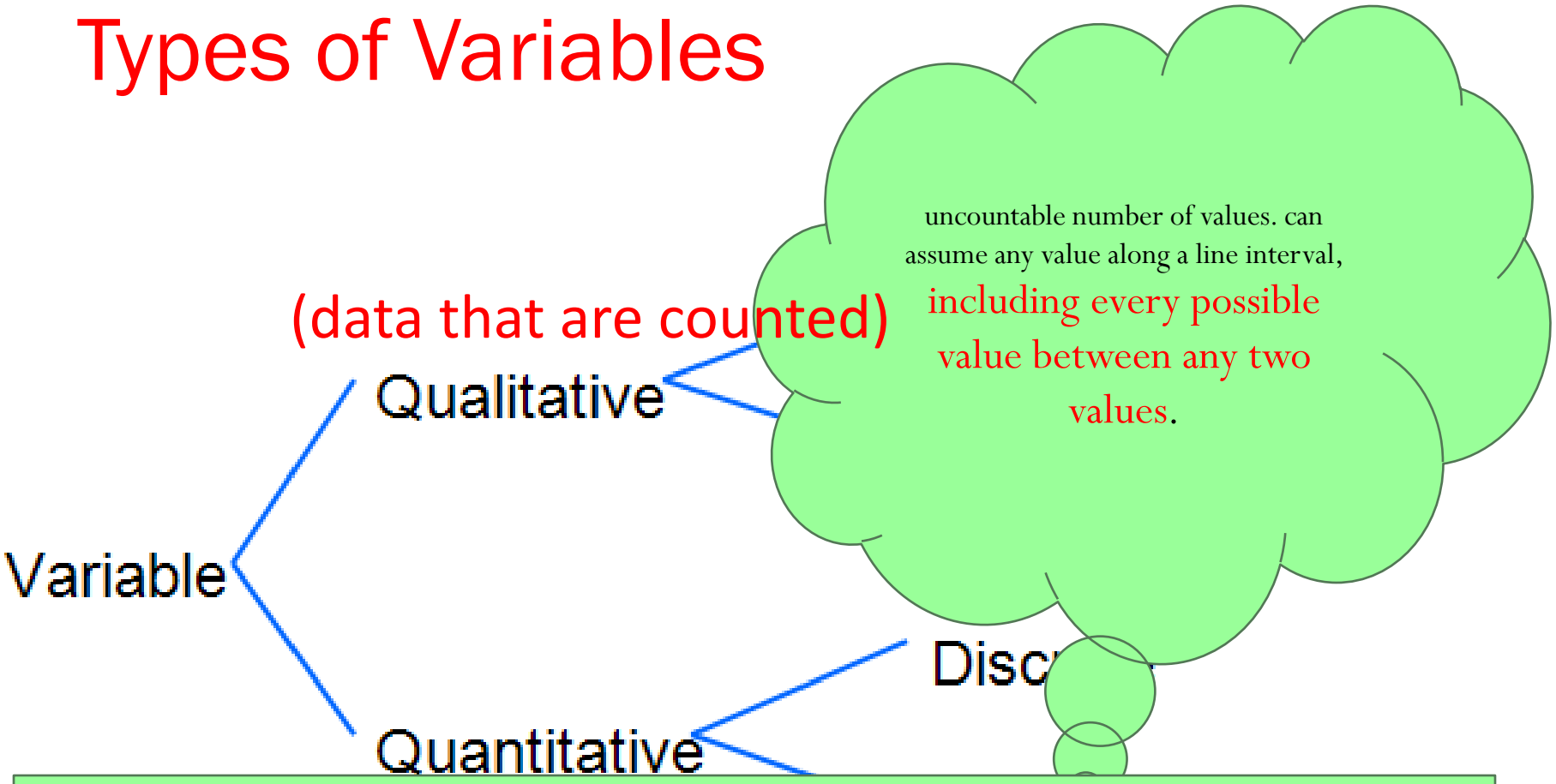
Examples of interval variables:

Intelligence (IQ test score of 100, 110, 120, etc.)

Pain level (1-10 scale)

Body length in infant

Types of Variables



Practically, **real values** can only be measured and represented using a finite number of digits

Continuous attributes are typically represented as floating-point variables

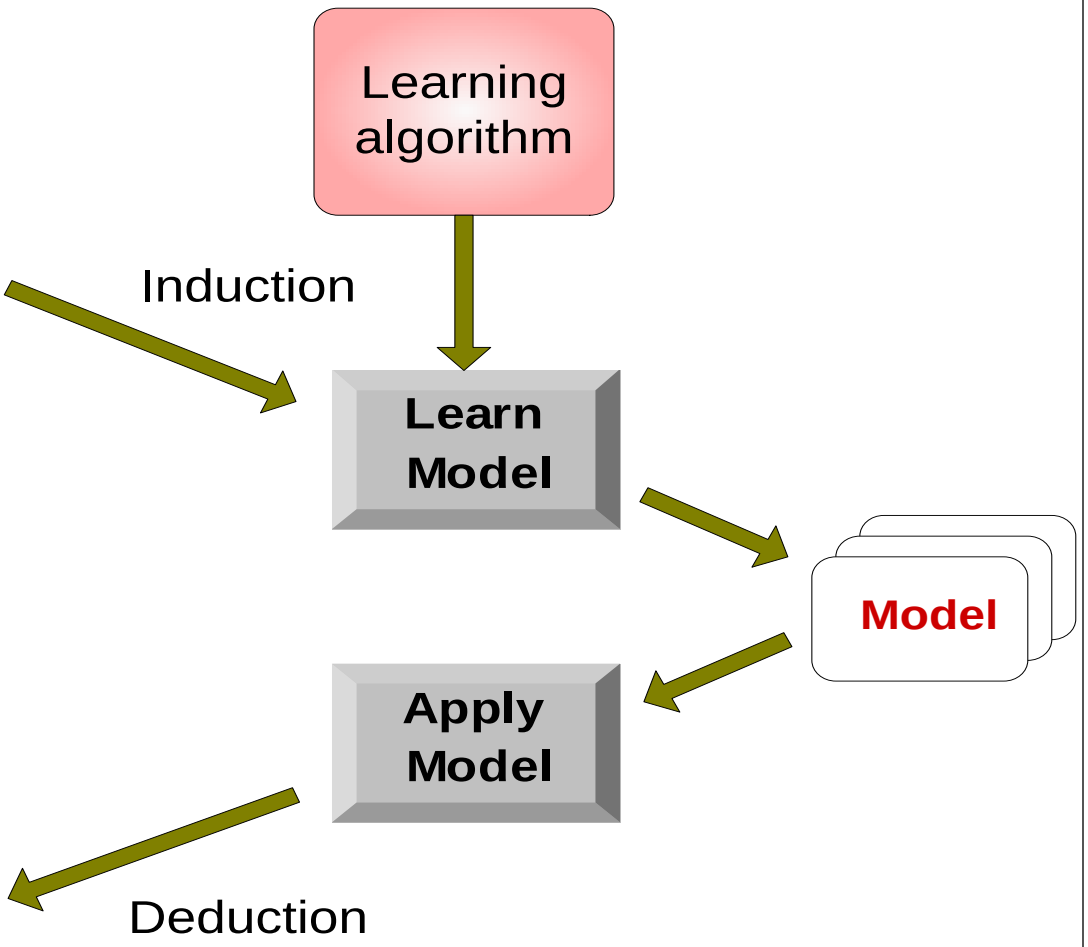
Illustrating Classification Task

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



Naïve Bayes classifier approach

If two people live in the same city, the probability that person **A** gets home in time for dinner, and the probability that person **B** gets home in time for dinner are independent; that is, we wouldn't expect one to have an affect on the other. But if a snow storm hits the city and introduces a probability **C** that traffic will be at a stand still, you would expect that the probability of both **A** getting home in time for dinner and **B** getting home in time for dinner, would change.

- Based on the observed object attributes
  - Naïvely assumed to be conditionally independent of
  - Class label probabilities are determined using Bayes' L
  - Determine the most probable class label for each objec
- Example:
  - Classify an object based on its attributes {shape, color, we
  - Given an object that is {spherical, yellow, < 60 grams}
    - P(tennis ball, given spherical, yellow, < 60 grams) = 0.32
    - P(apple, given spherical, yellow, < 60 grams) = 0.09
    - P(bowling ball, given spherical, yellow, < 60 grams) = 0.00000001
- **Input** variables are discrete, categorical
- **Output**:
  - Probability score for each possible class label
    - Proportional to the true probability
  - Assigned class label, based on the highest probability score

Build training dataset to predict customer purchase

- Predict if the customer will purchase the product based on their profile:
  - Age bins
  - Occupation
  - Income tier
- Note: Continuous variables are transformed into categorical variables.

Purchase_fla	Age_tiers	Occupation	Income_tiers_1000s
Yes	40 to 50	Professor	<80
Yes	30 to 40	Data Scientist	>200
Yes	50 to 60	Professor	>200
Yes	30 to 40	Professor	>200
Yes	>60	Doctor	>200
Yes	50 to 60	Professor	80 to 120
Yes	>60	Doctor	120 to 200
Yes	30 to 40	Professor	120 to 200
Yes	50 to 60	Professor	80 to 120
Yes	40 to 50	Electrician	120 to 200
Yes	40 to 50	Doctor	80 to 120
Yes	20 to 30	Data Scientist	120 to 200
Yes	50 to 60	Data Scientist	80 to 120
Yes	20 to 30	Professor	>200
No	30 to 40	Electrician	>200
No	30 to 40	Electrician	120 to 200
No	20 to 30	Electrician	>200
No	30 to 40	Professor	>200
No	40 to 50	Electrician	>200
No	>60	Professor	>200
No	30 to 40	Electrician	<80
No	50 to 60	Electrician	120 to 200
No	30 to 40	Electrician	120 to 200
No	>60	Doctor	80 to 120
No	20 to 30	Professor	<80
No	30 to 40	Electrician	>200
No	>60	Electrician	120 to 200
No	>60	Data Scientist	80 to 120
No	30 to 40	Professor	>200
No	40 to 50	Doctor	<80
No	30 to 40	Electrician	80 to 120
No	>60	Doctor	>200
No	30 to 40	Professor	120 to 200
No	30 to 40	Doctor	>200
No	20 to 30	Data Scientist	80 to 120

Conditional probability

The probability of event C occurring given event A has occurred

Denoted as P(C | A)

Example:

A fair 6-sided die is thrown

Let A = {an even number is rolled}

If C = {a 3 is rolled}, then P(C | A) = 0

If C = {a 4 is rolled}, then P(C | A) = 1/3

Knowing that A occurred, provides information about the probability of C

Formal definition:

$$P(C \mid A) = \frac{P(A \cap C)}{P(A)} \quad for\ P(A) > 0$$

where $P(A \cap C)$ denotes probability of events A and C occurring

Derivation of Bayes' Law

By definition of conditional probability,

$$P(C \mid A) = \frac{P(A \cap C)}{P(A)} \quad (1)$$

Alternatively,

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)} \rightarrow P(A \cap C) = P(A \mid C)P(C) \quad (2)$$

Substituting back into the definition yields:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

Known as Bayes' Law

A conditional probability can be expressed as a function of another conditional probability

Application of Bayes' Law

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

Scenario

John flies frequently and likes to upgrade his seat to first class.

If John arrives at least two hours early, then he will get the upgrade 75 percent of the time.

Otherwise, he will get the upgrade 35 percent of the time.

John arrives at least two hours early only 40 percent of the time.

Suppose that John did not receive an upgrade on his most recent attempt.

What is the probability that he arrived late?

$$P(Late \mid No\ Upgrade) = \frac{P(No\ Upgrade \mid Late)P(Late)}{P(No\ Upgrade)}$$

$$= \frac{(1 - 0.35)(1 - 0.40)}{1 - (0.40 * 0.75 + 0.60 * 0.35)} = 0.80$$

Apply Naïve assumption and remove constant

For observed attributes $A = (a_1, a_2, \ldots a_m)$, compute

$$P(C_i \mid A) = \frac{P(a_1, a_2, \ldots, a_m \mid C_i) P(C_i)}{P(a_1, a_2, \ldots, a_m)} \quad i = 1, 2, \ldots, n$$

and assign the classifier C_i with the largest $P(C_i \mid A)$

Two simplifications to the calculations

Apply naïve assumption - each a_j is conditionally independent of each other, then

$$P(a_1, a_2, \ldots, a_m \mid C_i) = P(a_1 \mid C_i) P(a_2 \mid C_i) \cdots P(a_m \mid C_i) = \prod_{j=1}^{m} P(a_j \mid C_i)$$

Denominator $P(a_1, a_2, \ldots, a_m)$ is a constant and can be ignored

Building Naïve Bayesian classifier

Applying the two simplifications

$$P(C_i \mid a_1, a_2, ..., a_m) \propto \left(\prod_{j=1}^{m} P(a_j \mid C_i) \right) P(C_i) \qquad i = 1, 2, ..., n$$

To build a Naïve Bayesian Classifier, collect the following statistics from the training data:

$P(C_i)$ for all the class labels

$P(a_j \mid C_i)$ for all possible a_j and C_i

Assign the classifier label C_i that maximizes the value of

$$\left(\prod_{j=1}^{m} P(a_j \mid C_i) \right) P(C_i) \qquad i = 1, 2, ..., n$$

Naïve Bayesian classifiers for product purchase example

- Class labels: {Yes, No}
  - P(Yes) = 0.39
  - P(No) = 0.61
- Conditional Probabilities
  - P(Electrician | Yes) = 0.42
  - P(Electrician | No) = 0.27
  - P(Data Scientist | Yes) = 0.21
  - P(Data Scientist | No) = 0.27
  - … and so on

Purchase_flag	Age_tiers	Occupation	Income_tiers_1000s
Yes	40 to 50	Professor	<80
Yes	30 to 40	Data Scientist	>200
Yes	50 to 60	Professor	>200
Yes	30 to 40	Professor	>200
Yes	>60	Doctor	>200
Yes	50 to 60	Professor	80 to 120
Yes	>60	Doctor	120 to 200
Yes	30 to 40	Professor	120 to 200
Yes	50 to 60	Professor	80 to 120
Yes	40 to 50	Electrician	120 to 200
Yes	40 to 50	Doctor	80 to 120
Yes	20 to 30	Data Scientist	120 to 200
Yes	50 to 60	Data Scientist	80 to 120
Yes	20 to 30	Professor	>200
No	30 to 40	Electrician	>200
No	30 to 40	Electrician	120 to 200
No	20 to 30	Electrician	>200
No	30 to 40	Professor	>200
No	40 to 50	Electrician	>200
No	>60	Professor	>200
No	30 to 40	Electrician	<80
No	50 to 60	Electrician	120 to 200
No	30 to 40	Electrician	120 to 200
No	>60	Doctor	80 to 120
No	20 to 30	Professor	<80
No	30 to 40	Electrician	>200
No	>60	Electrician	120 to 200
No	>60	Data Scientist	80 to 120
No	30 to 40	Professor	>200
No	40 to 50	Doctor	<80
No	30 to 40	Electrician	80 to 120
No	>60	Doctor	>200
No	30 to 40	Professor	120 to 200
No	30 to 40	Doctor	>200
No	20 to 30	Data Scientist	80 to 120

Naïve Bayesian classifier example, cont.

- Given applicant attributes of
  A= {Age 30–40,
  Occupation Electrician,
  Income 80–120}
- Since P(No | A) > (Yes | A), assign the label No, the customer will not purchase.
  P(Yes | A) ~ (0.21*0.42*0.28)*0.39 = 0.009
  P(No | A) ~ (0.36*0.22*0.40)*0.61 = 0.019

a_j	C_i	$P(a_j \mid C_i)$
30-40	Yes	0.21
30-40	No	0.36
Electrician	Yes	0.42
Electrician	No	0.22
80-120	Yes	0.28
80-120	No	0.40

Naïve Bayesian implementation considerations

- Numerical underflow
  - Resulting from multiplying several probabilities near zero
  - Preventable by computing the logarithm of the products
- Zero probabilities due to unobserved attribute/classifier pairs
  - Resulting from rare events
  - Handled by smoothing—adjusting each probability by a small amount
- Assign the classifier label, C_i, that maximizes the value of

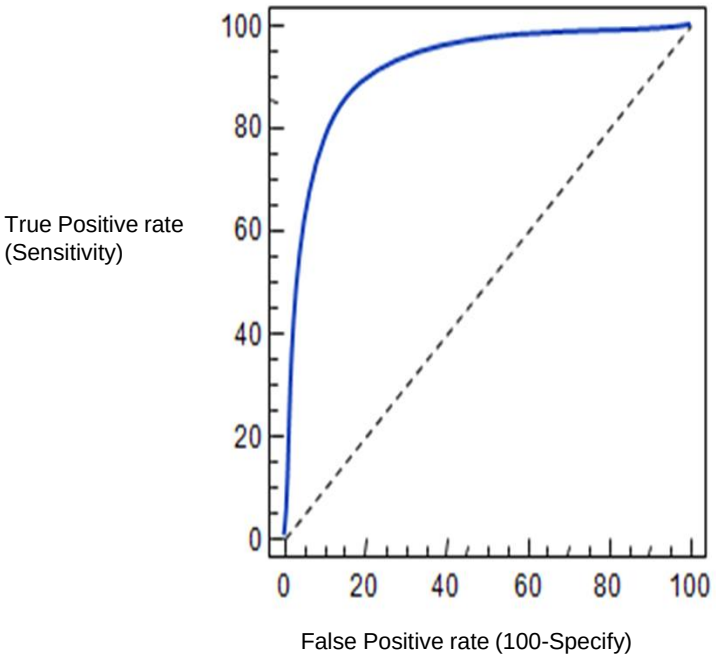
$$\left(\sum_{j=1}^{m} \log P'(a_j \mid C_i) \right) + \log P(C_i)$$

where i = 1,2,…,n and

P' denotes the adjusted probabilities

Diagnostics

- Hold-out data
  - How well does the model classify new instances?
- Cross-validation
- ROC curve/AUC
- Confusion Matrix



Naïve Bayesian classifier—reasons to choose (+) and cautions (-)

Reasons to choose (+)	Cautions (-)
Handles missing values quite well	Numeric variables must be discrete, categorized, Intervals
Robust to irrelevant variables	Sensitive to correlated variables Double-counting
Easy to implement	Not good for estimating probabilities Stick to class label or yes/no
Easy to score data	
Resistant to overfitting	
Computationally efficient Handles very high-dimensional problems Handles categorical variables with many levels	

Check your knowledge

1. Consider the following training dataset:
   - Apply the Naïve Bayesian Classifier to this dataset and compute the probability score for $P(y = 1 \mid X)$ for $X = (1,0,0)$
   - Show your work

X1	X2	X3	Y
1	1	1	0
1	1	0	0
0	0	0	0
0	1	0	1
1	0	1	1
0	1	1	1

Training Dataset

Thanks

for your attention