Data Mining

CS57300
Purdue University

January 21, 2016

Probability and statistics basics

Probability basics

- Basic element: **Random variable**
  - Maps a set of possible values into a probability measure
  - *X* refers to random variable; *x* refers to a value of that random variable
- Types of random variables
  - Discrete RV has a finite or countably infinite set of possible values
    - e.g., Is there a storm warning? = <yes, no>
    - e.g., Number of pens bought <0,1,2,3,…>
  - Continuous RV can take any value within an interval
    - e.g., Temperature, Location

Probability basics

- **Sample space (S)**
  - Set of all possible outcomes of an experiment
- **Event**
  - Any subset of *outcomes* contained in the sample space S
  - When events ***A*** and ***B*** have no outcomes in common they are said to be *mutually exclusive*

Random variable(s)	**Sample space**	**Example event**
Two coin tosses	HH, HT, TH, TT	At least one H
Select one card	2♥, 2♦, ..., A♣ (52)	Face card of any suit

Axioms of probability

- For a sample space S with possible events $\mathbf{A_S}$:
  A function that associates real values with each event A is called a ***probability function*** if the following properties are satisfied:

  1. $0 \leq P(A) \leq 1$ for every A

  2. $P(S) = 1$

  3. $P(A_1 \vee A_2 \ ... \ \vee A_{n \in S}) = P(A_1) + P(A_2) + ... + P(A_n)$

     if $A_1, A_2, ..., A_n$ *are pairwise mutually exclusive events*

Interpreting probabilities

- Meaning of probability is focus of debate and controversy
- Two main views: Frequentist and Bayesian

Frequentist view

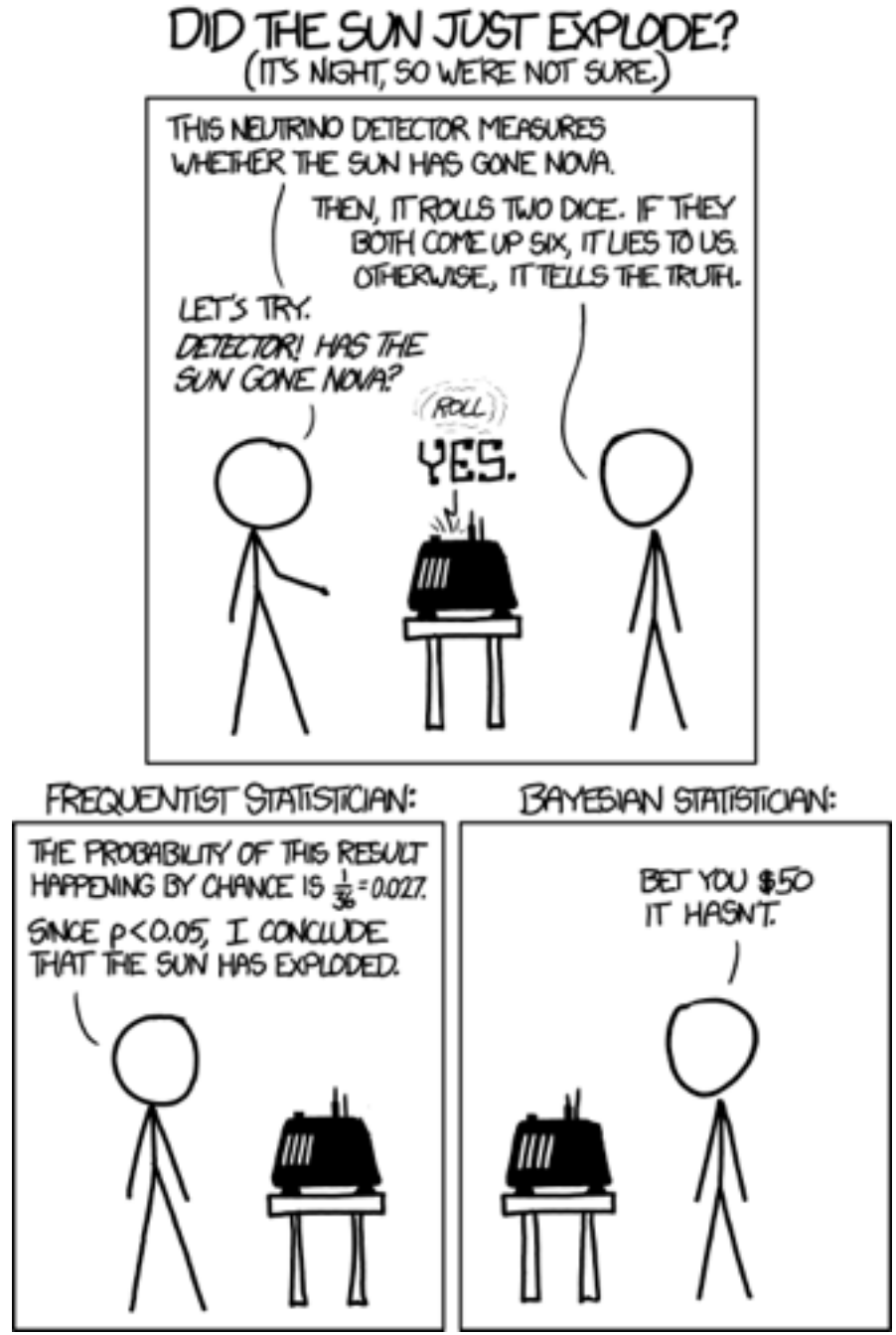
- Dominant perspective for last century
- Probability is an **objective** concept
    - Defined as the frequency of an event occurring under repeated trials in “same” situation
    - E.g., number of heads in repeated coin tosses
- Restricts application of probability to repeatable events

Calculating probabilities: frequentist

- Repeated experiments
  - Let n be the number of times an experiment is performed
  - Let n(A) be the number of outcomes in which A occurs
  - Then as $n \to \infty$ $P(A) = n(A) / n$
- When the various outcomes of an experiment are equally likely, the task of computing probability reduces to counting
  - Let N be size of sample space (i.e., number of simple outcomes)
  - Let N(A) be the number of outcomes contained in A
  - Then: $P(A) = N(A) / N$

Bayesian view

- Increasing importance over last decade
    - Due to increase in computational power that facilitates previously intractable calculations
- Probability is a **subjective** concept
    - Defined as individual degree-of-belief that event will occur
    - E.g., belief that we will have another snow storm tomorrow
- Observed data helps us to update and inform our prior beliefs



http://xkcd.com/1132/

Calculating probabilities: Bayesian

- *Begin with prior belief estimates*: P(A)
    - E.g., On Sunday after the Seahawks won their conference championship, Vegas casinos believed there was a slight chance in favor of the Seahawks winning the Superbowl over the Patriots:
      P(Seahawks win)=0.525, P(Patriots win)=0.475
- *Update belief by conditioning on observed data*
  P(A|data) = P(data|A) P(A) / P(data)
    - But then Vegas observed that a heavy majority of the betters (80%) chose the Patriots so they updated their odds to increase their confidence in the Patriots and reduce their confidence in the Seahawks:
      P(Seahawks win|betting)=0.50, P(Patriots win|betting)=0.50
- Even when the same data is observed, if people have different priors, they can end up with different posterior probability estimates P(A|data)

Bayesian vs. frequentist

- Bayesian central tenet:
  - Explicitly model all forms of uncertainty
  - E.g., Parameters, model structure, predictions
- Frequentist often model same uncertainty but in less-principled manner, e.g.,:
  - Parameters set by cross-validation
  - Model structure averaged in ensembles
  - Smoothing of predicted probabilities
- Although interpretation of probability is different, underlying calculus is the same

Probability distribution

- **Probability distribution** *(i.e., probability mass function or probability density function)* specifies the probability of observing every possible value of a random variable
- Discrete
  - Denotes probability that *X* will take on a particular value:

$$P(X = x)$$

- Continuous
  - Probability of any particular point is 0, have to consider probability within an interval:

$$P(a < X < b) = \int_a^b p(x)dx$$

Joint probability

- **Joint probability distribution** for a set of random variables gives the probability of every atomic event on those random variables

  E.g., P(Weather, Warning) = a 4 × 2 matrix of values:

	sunny	rainy	cloudy	snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

- Every question about events can be answered by the joint distribution

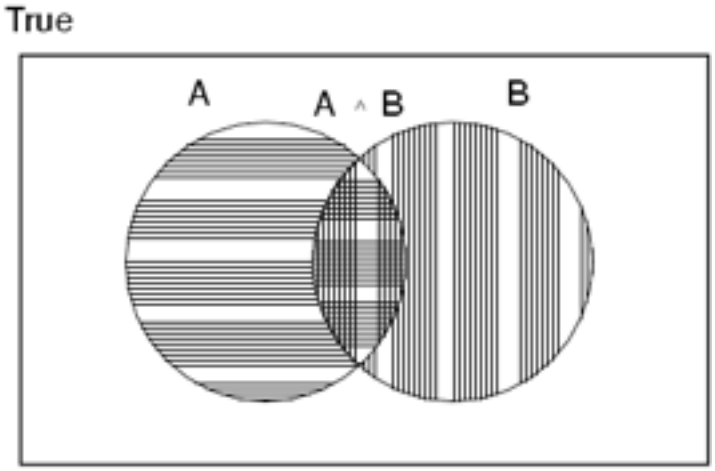
Conditional probability

- **Conditional** (or posterior) probability:
  - e.g., P(warning | snow) = 0.4
  - Complete conditional distributions:
    P(warning | snow) =
    {P(warning = Y | snow = T), P(warning = N | snow = T)},
    {P(warning = Y | snow = F), P(warning = N | snow = F)}
- If we know more, then we can update the probability by conditioning on more evidence
  - e.g., if Windy is also given then P(warning | snow, windy) = 0.5

Conditional probability

- Definition of conditional probability:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} \qquad \text{if } P(B) > 0$$



- **Product rule** gives an alternative formulation:

$$P(A \wedge B) = P(A|B)P(B)$$
$$= P(B|A)P(A)$$

- **Chain rule** is derived by successive application of product rule:

$$P(X_1, \ldots, X_n) = P(X_n|X_1, \ldots, X_{n-1})P(X_1, \ldots, X_{n-1})$$
$$= P(X_n|X_1, \ldots, X_{n-1})P(X_{n-1}|X_1, \ldots, X_{n-2})P(X_1, \ldots, X_{n-2})$$
$$= \ldots$$
$$= \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1})$$

Marginal probability

- **Marginal** (or unconditional) probability corresponds to belief that event will occur regardless of conditioning events
- Marginalization: $$P(A) = \sum_{b \in B} P(A, b)$$
$$= \sum_{b \in B} P(A|b)P(b)$$
- Example: What is P(cloudy)?

	sunny	rainy	cloudy	snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

Independence

- Two variables A and B are independent if knowing B tells you nothing about A and vice versa:

$$P(A|B) = P(A) \quad \text{or} \quad P(B|A) = P(B) \quad \text{or} \quad P(A, B) = P(A)\,P(B)$$

- Two variables *A* and *B* are **conditionally** independent given *Z*
  iff for all values of *A, B, Z:*

$$P(A, B \mid Z) = P(A \mid Z)\,P(B \mid Z)$$

- *Note: independence does not imply conditional independence or vice versa*

Example 1

- **Conditional independence does not imply independence**
- Gender and lung cancer are not independent
  $P(C \mid G) \neq P(C)$
- Gender and lung cancer are conditionally independent given smoking
  $P(C \mid G, S) = P(C \mid S)$
- Why? Because gender indicates likelihood of smoking, and smoking causes cancer

Example 2

- **Independence does not imply conditional independence**
- Sprinkler-on and raining are independent
  $P(S \mid R) = P(S)$
- Sprinkler-on and raining are not conditionally independent given grass is wet
  $P(S \mid R, W) \neq P(S \mid R)$
- Why? Because once we know the grass is wet, if it's not raining, then the explanation for the grass being wet has to be the sprinkler

Expectation

- Denotes the expected value or mean value of a random variable X

- Discrete

$$E[X] = \sum_x x \cdot p(x)$$

- Continuous

$$E[X] = \int_x x \cdot p(x) dx$$

- Expectation of a function

$$E[h(X)] = \sum_x h(x) \cdot p(x)$$

$$E[aX + b] = a \cdot E[X] + b$$

- Linearity of expectation

$$E[X + Y] = E[X] + E[Y]$$

Variance

- Denotes the squared deviation of X from its mean

- Variance

$$Var(X) = E[(x - E[X])^2]$$
$$= E[X^2] - (E[X])^2$$

- Standard deviation

$$\sigma = \sqrt{Var(X)}$$

- Variance of a function

$$Var(aX + b) = a^2 \cdot Var(X)$$

$$Var(h(X)) = \sum_x (h(x) - E[h(x)])^2 \cdot p(x)$$

Common distributions

- Bernoulli
- Binomial
- Multinomial
- Poisson
- Normal

Bernoulli

- Binary variable (0/1) that takes the value of 1 with probability p
  - E.g., Outcome of a fair coin toss is Bernoulli with *p=0.5*

$$P(x) = p^x(1-p)^{1-x}$$

$$E[X] = 1(p) + 0(1-p) = p$$

$$\begin{aligned} Var(X) &= E[X]^2 - (E[X])^2 \\ &= 1^2(p) + 0^2(1-p) - p^2 \\ &= p(1-p) \end{aligned}$$

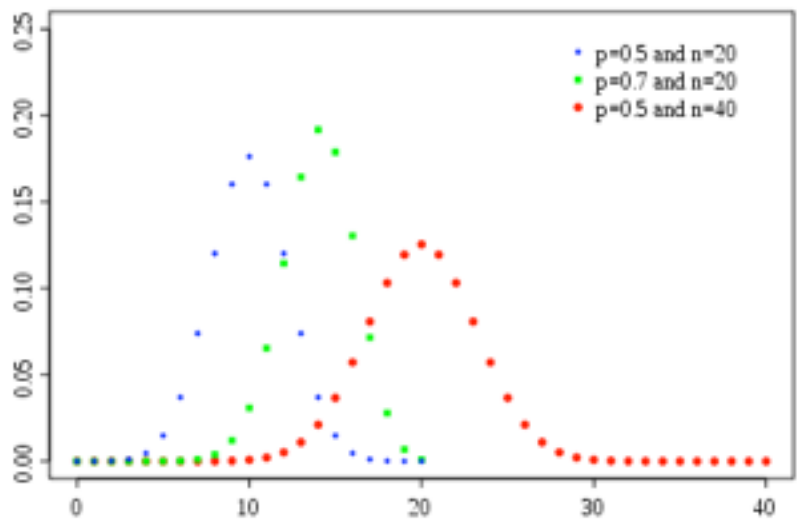
Binomial

- Describes the number of successful outcomes in *n* independent Bernoulli(p) trials
  - E.g., Number of heads in a sequence of 10 tosses of a fair coin is Binomial with *n=10* and *p=0.5*

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$E[X] = np$$

$$Var[X] = np(1-p)$$



Multinomial

- Generalization of binomial to *k* possible outcomes; outcome *i* has probability p_i of occurring
    - E.g., Number of {outs, singles, doubles, triples, homeruns} in a sequence of 10 times at bat is Multinomial
- Let X_i denote the number of times the *i*-th outcome occurs in *n* trials:

$$P(x_1, ...x_k) = \binom{n}{x_1, ...x_k} p_1^{x_1} p_2^{x_2} ...p_k^{x_k}$$

$$E[X_i] = np_i$$

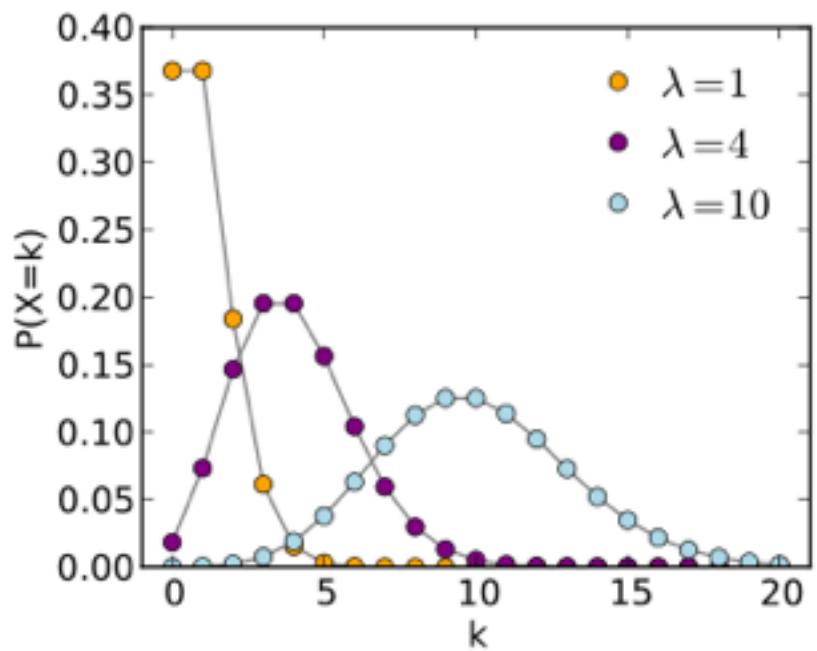
$$Var(X_i) = np_i(1 - p_i)$$

Poisson

- Describes the probability of a given number of events occurring in a fixed interval of time (or space), given an average arrival rate (λ) and independent events that occur randomly over time (or space)
  - E.g., Given an average of 4 power failures per winter, what is the probability that there will be more than 7 failures this winter?

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$\lambda = E[X] = Var[X]$$



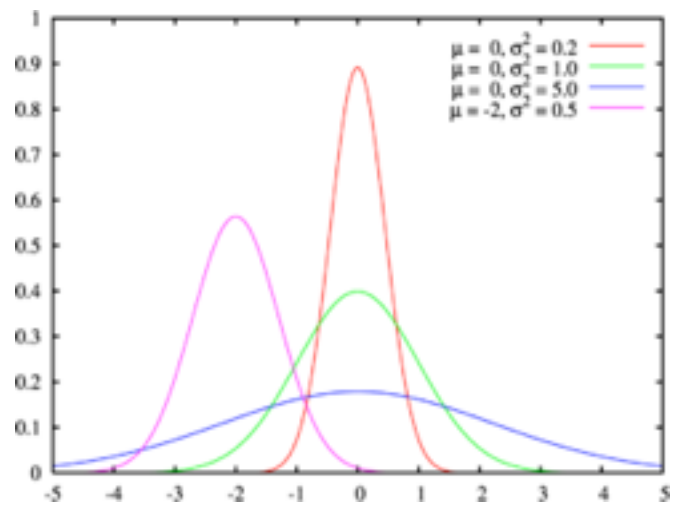
Normal (Gaussian)

- Important distribution gives well-known bell shape
- Central limit theorem:
  - Distribution of the mean of n samples becomes normally distributed as n ↑, regardless of the distribution of the underlying population

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$E[X] = \mu$$

$$Var(X) = \sigma^2$$



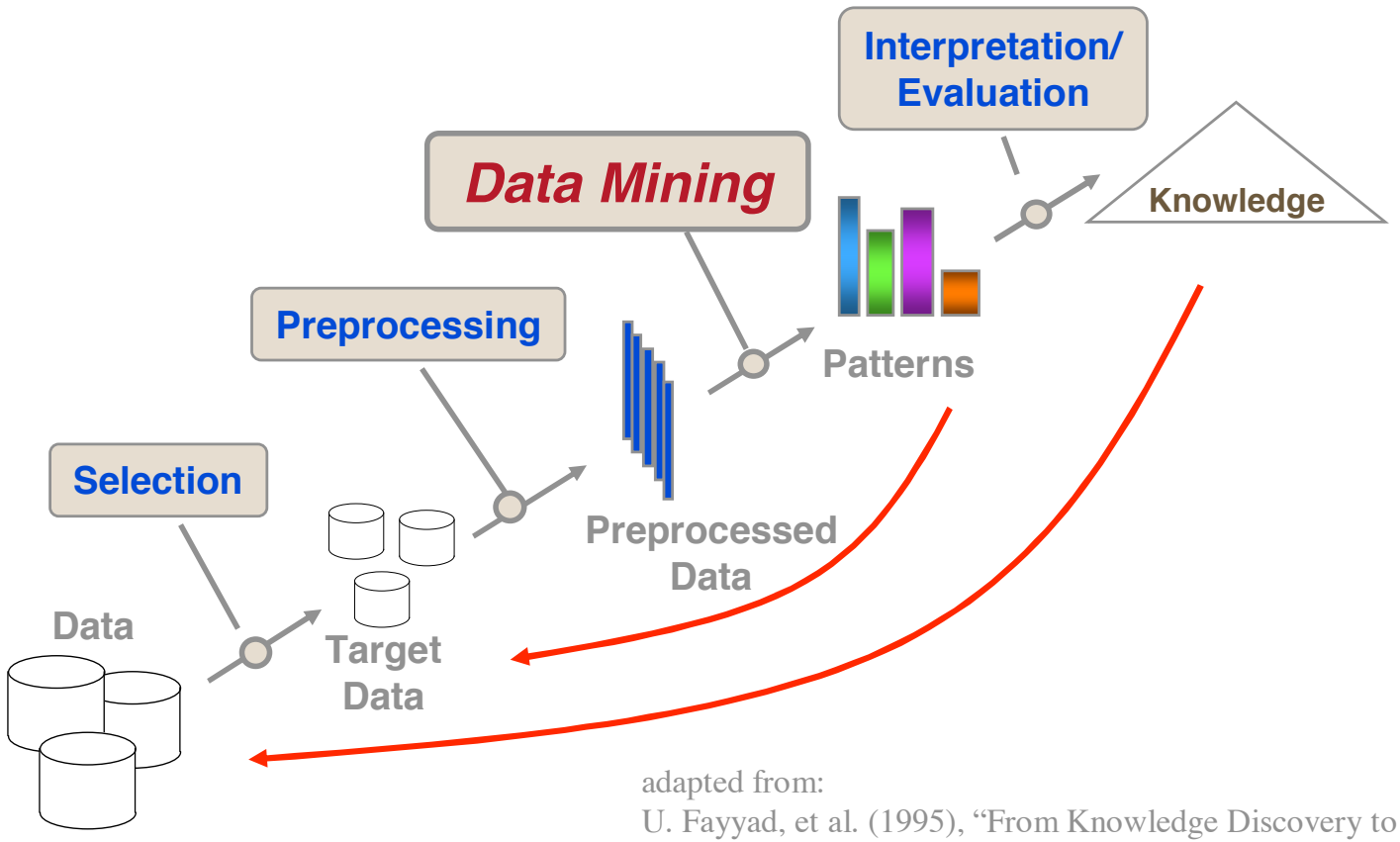
Multivariate RV

- A multivariate random variable ***X*** is a set $X_1, X_2, \ldots X_p$ of random variables
- **Joint** density function: $P(\mathbf{x}) = P(x_1, x_2, \ldots, x_p)$
- **Marginal** density function: the density of any subset of the complete set of variables, e.g.,:

$$P(x_1) = \sum_{x_2, x_3} p(x_1, x_2, x_3)$$

- **Conditional** density function: the density of a subset conditioned on particular values of the others, e.g.,:

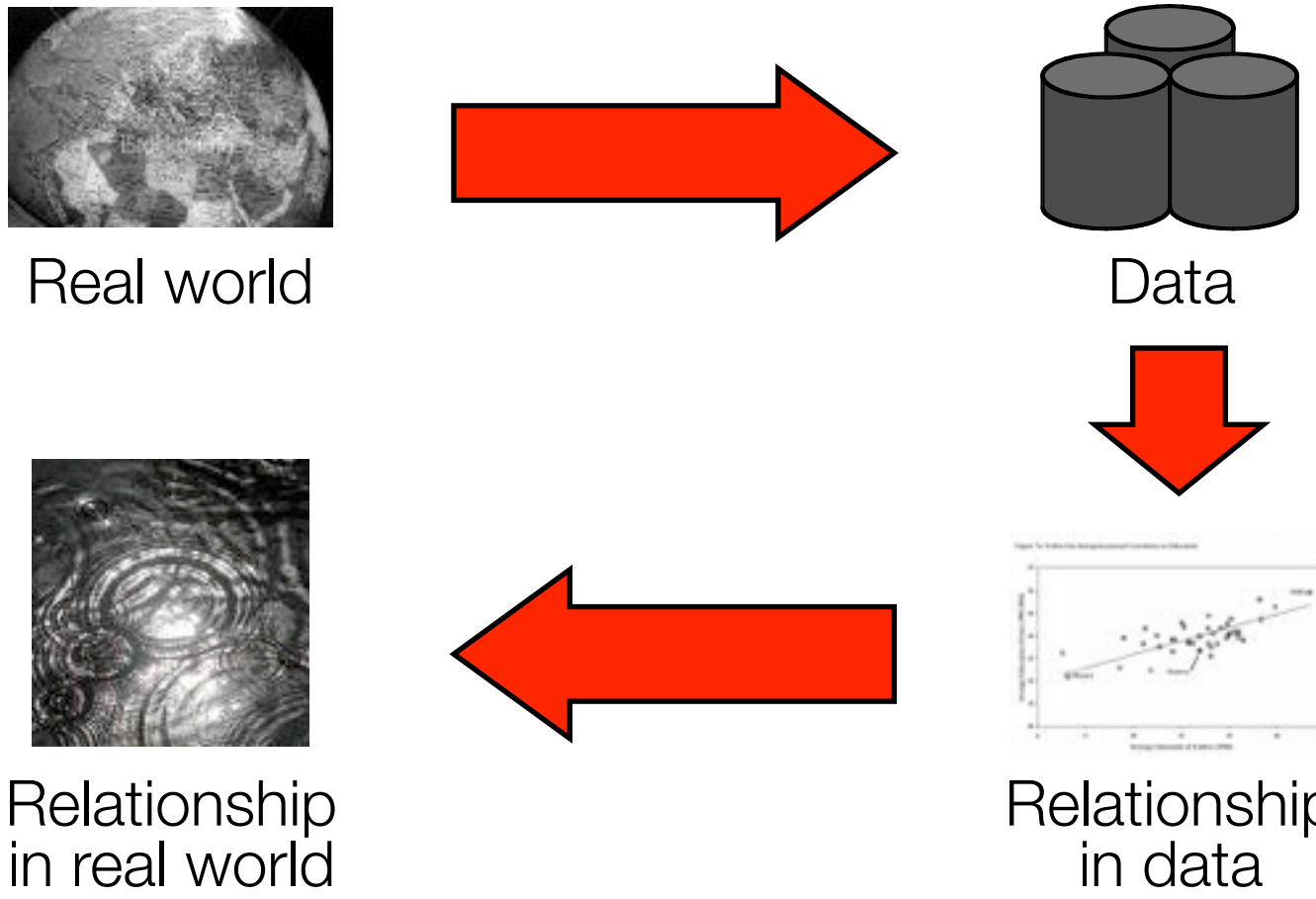
$$P(x_1 | x_2, x_3) = \frac{p(x_1, x_2, x_3)}{p(x_2, x_3)}$$



adapted from:
U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

Data and Measurement

Measurement



Goal: map domain entities to symbolic representations

What is data?

- Collection of entities and their attributes
- **Attribute**: property or characteristic of an entity (e.g., eye color, temperature)
- **Entity**: collection of attributes
  Aka: record, point, case, sample, object, or instance

Attributes

Entities

Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$29.09	Yes	486	Both
M24	2.55	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	$41.32	No	434	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat

Discrete and continuous attributes

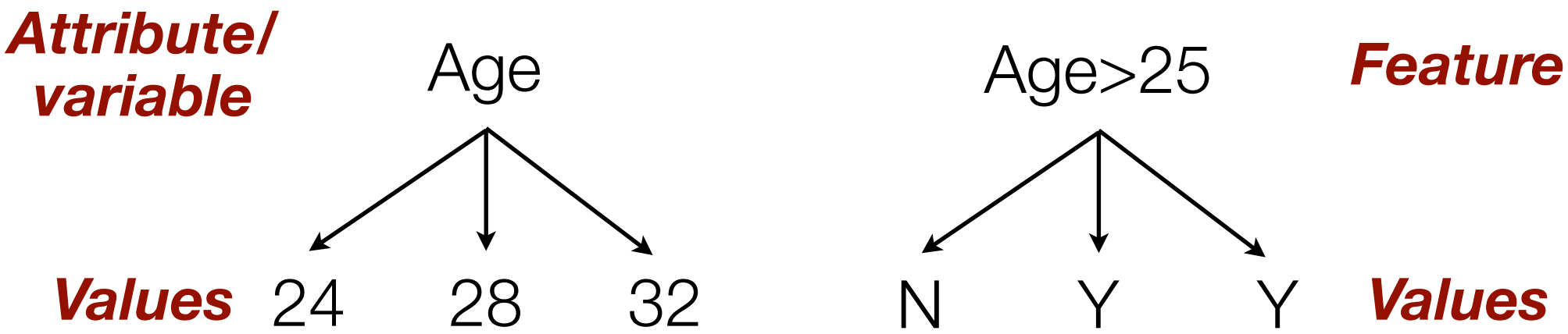
- Discrete
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, set of words in a collection of documents
  - Often represented as integer variables
- Continuous
  - Has real numbers as attribute values
  - Examples: temperature, height
  - Continuous attributes are typically represented as floating-point variables

Hierarchy of measurements

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Tan, Steinbach, Kumar. Introduction to Data Mining, 2004.

Naming conventions



Tabular data

- Collection of records, each of which consists of a fixed set of attributes

Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$29.09	Yes	486	Both
M24	2.55	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	$41.32	No	434	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat

Document data

- Each document is represented as a **term** vector, where each attribute records the number of times the term occurs in the document

Terms	Documents													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0

Transaction data

- Each record corresponds to a transaction that involves a set of items
- E.g., in a grocery store purchase, the set of products purchased by a customer constitute a transaction, while the individual products that were purchased are the items

Table 6.22. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	{a,d,e}
1	0024	{a,b,c,e}
2	0012	{a,b,d,e}
2	0031	{a,c,d,e}
3	0015	{b,c,e}
3	0022	{b,d,e}
4	0029	{c,d}
4	0040	{a,b,c}
5	0033	{a,d,e}
5	0038	{a,b,e}

User Trajectory Data

At Amazon:

- User 1: Browses items Point-and-shoot Canon, DLR Canon, DLR Nikon, Point-and-shoot Nikon, …
- User 2: Browses items Sony Laptop, Mac Laptop, Google Laptop, Samsung Laptop,…

At Last.fm:

- User 1 listens to Adele, Iron Maiden, Mozart, …
- …