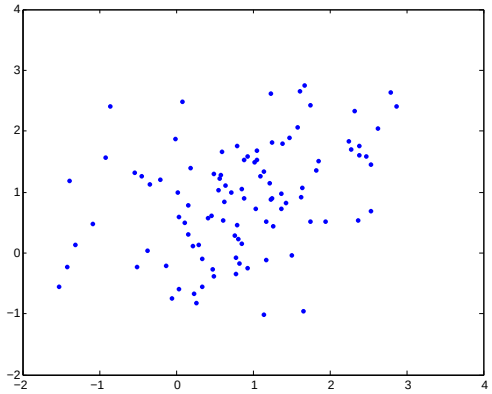
Cogsci 109

Virginia de Sa
desa at cogsci

Data Modeling

Suppose you are given some data and you would like to have a model that preserves properties of the data (e.g. interpoint distances, mean, variance, higher-order statistics) that you care about but can be easily operated on (e.g. compared to another dataset).



A very common approach is to model the data with a probability distribution (fit a probability distribution to the data). So we will now take an aside to delve into the basics of probability distributions.

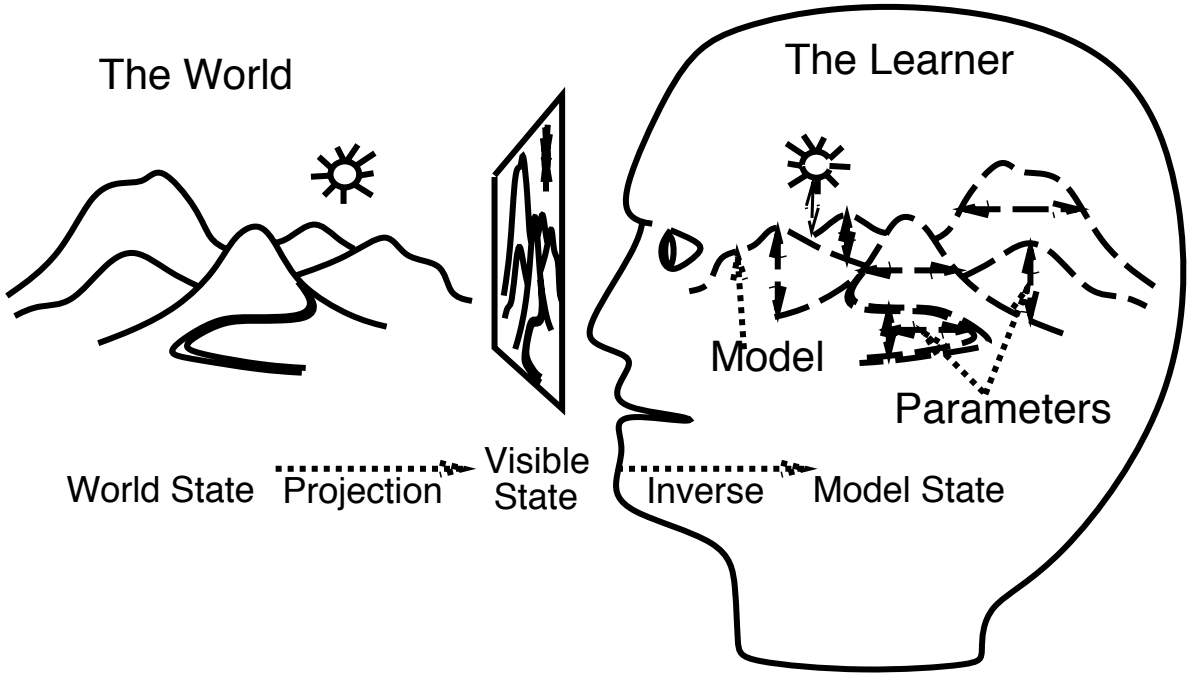
Probability

Probability theory is the natural way to deal with computations about uncertain events.

Both brains and computers must deal with uncertain events. Many people argue that in many problems the brain is performing optimal given the uncertainties that it has to deal with. "Performing optimally" usually means following the rules of probability.

Motivation

Our world is full of regularities, structure. It is useful for brains to learn about these regularities: brains construct models of the world. Models allow to correctly interpret ambiguous sensory inputs or even allow to predict future events:



Random Variables

Loosely, a random variable is a variable that has many values with different probabilities

The outcome of a roll of a die is a discrete random variable The height of a randomly chosen person is a continuous random variable

An event is a set of outcomes (e.g. die roll is odd)

Probability Notation

$P(a)$ means probability that event a is true

- shorthand for Pr(a=true)
- usually called "probability of a"

more detailed info

e.g the probability of rolling a 6 on a die is 1/6

P(rolling 6) = 1/6

Probability Axioms

- probability of event A, is between zero and one $0 \leq P(A) \leq 1$
- probability of some event occuing from the entire sample space S is one $P(S) = 1$
- if events A and B are $mutually\ exclusive$, then $P(AorB) = P(A) + P(B)$

Example: Rolling two dice

What is the probability of the outcome "getting doubles"?

What is the probability of the event "the sum is less than 5"?

Interpretation of Probability

frequentist view: relative frequency

$$P(A) = \lim_{n->\infty} \frac{N_A}{N}$$

N number of experiments

N_A number of experiments where A happened

Bayesian view: degree of belief

"What is the probability that there is life on Mars"

Bayesian probability

From Wikipedia, the free encyclopedia.

Bayesianism is the philosophical tenet that the mathematical theory of probability applies to the degree of plausibility of statements, or to the degree of belief of rational agents in the truth of statements. This is in contrast to frequentism, which rejects degree-of-belief interpretations of mathematical probability, and assigns probabilities only to random events according to their relative frequencies of occurrence. The Bayesian interpretation of probability allows probabilities assigned to random events, but also allows the assignment of probabilities to any other kind of statement. Whereas a frequentist and a Bayesian might both assign probability $1/2$ to the event of getting a head when a coin is tossed, a Bayesian might assign probability $1/2$ to personal belief in the proposition that there was life on Mars a billion years ago, without intending to assert anything about any relative frequency.

Conditional Probability

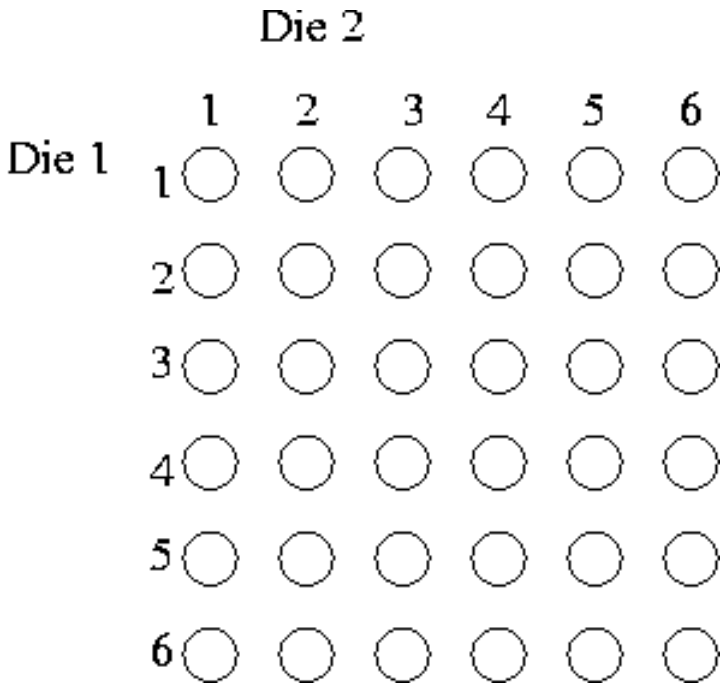
How does one event occuring affect the chance of occurence of another event occuring.

"What is the probability of voting for Bush, given that you have an IQ of 140?"

Probability of A given B

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

Jochen's dice example



What is P(D1=1,D2=2)?

What is P(D1=1 | D2=2)?

What is P(D1=1 | sum is even)?

What is P(D1=1 | sum is 4)?

Independence

Events A and B are independent iff

$P(A,B) = P(A)P(B)$

What are the alternate forms?

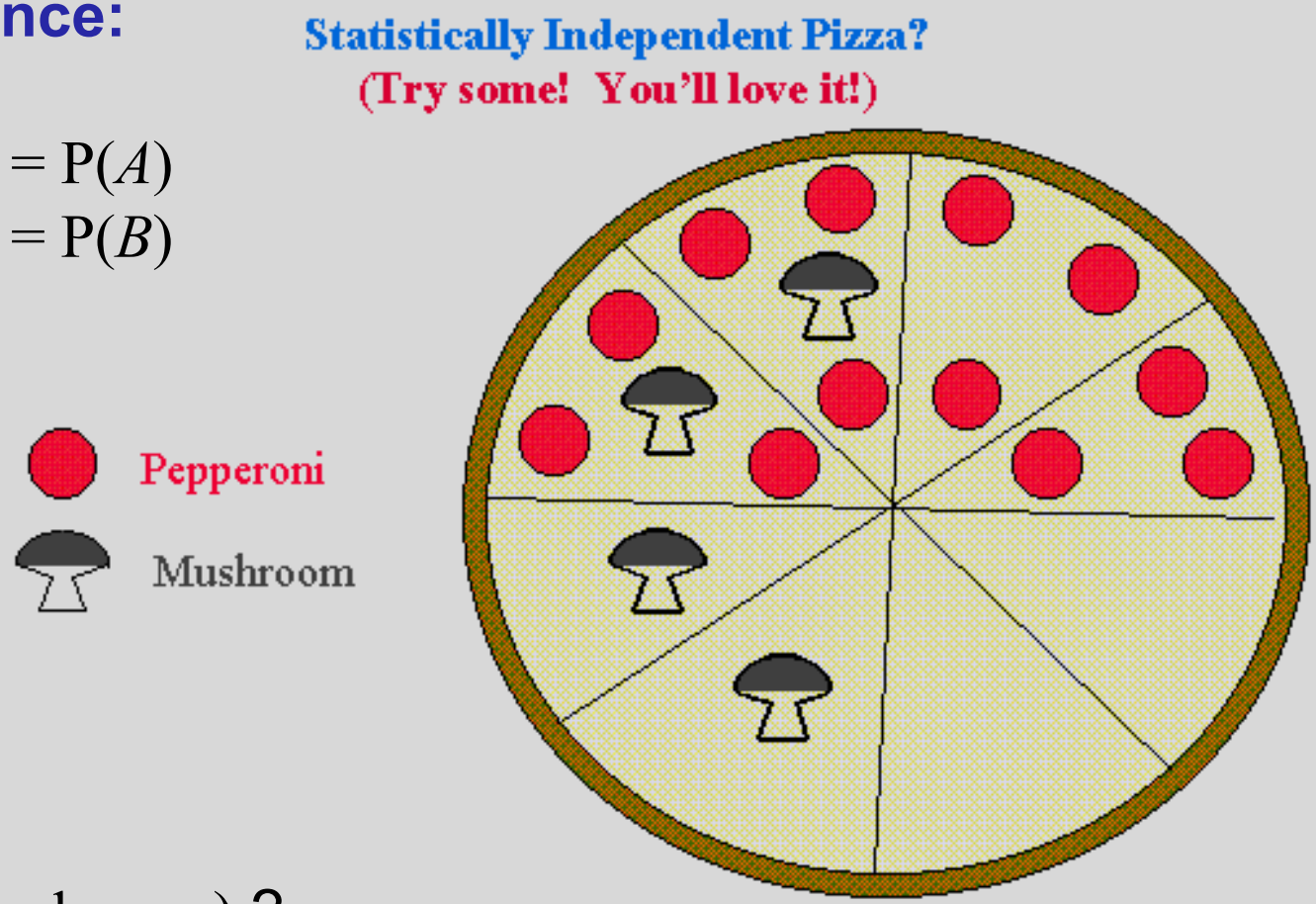
Jochen Triesch's famous pizza slide (modified from B. Warner)

Statistical Independence:

$P(A,B) = P(A)P(B)$

or equivalently: $P(A|B) = P(A)$

or equivalently: $P(B|A) = P(B)$



Pick slice at random!

What is P(pepperoni) ?

What is P(mushroom) ?

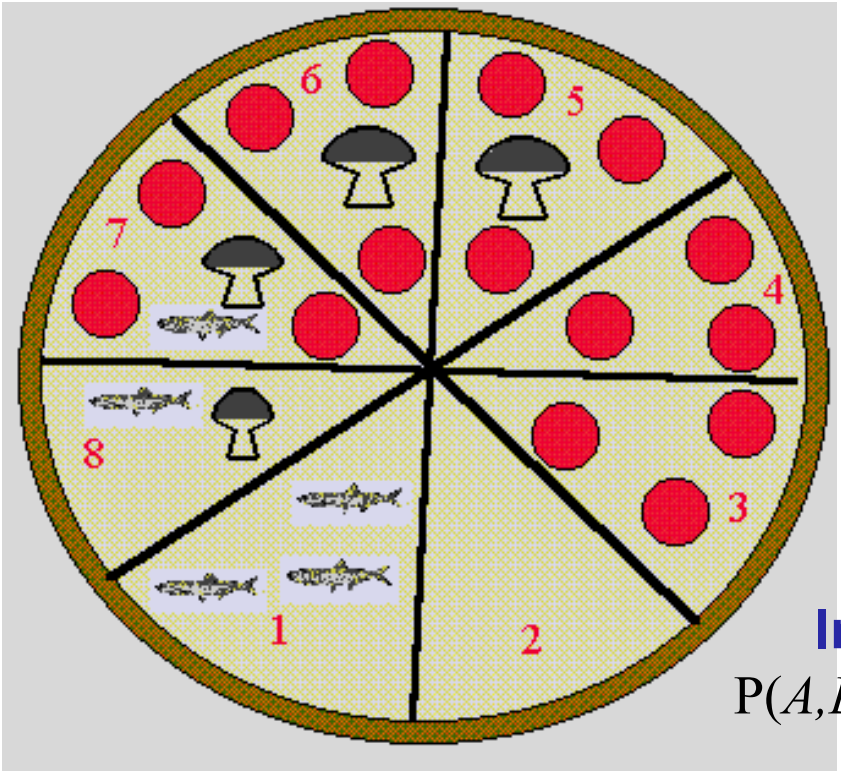
What is P(pepperoni | mushroom) ?

What is P(mushroom | pepperoni) ?

What is P(mushroom, pepperoni) ?

mushroom and pepperoni events are independent!

Conditional Independence



Venn diagrams in
Pizza form from
B. Warner

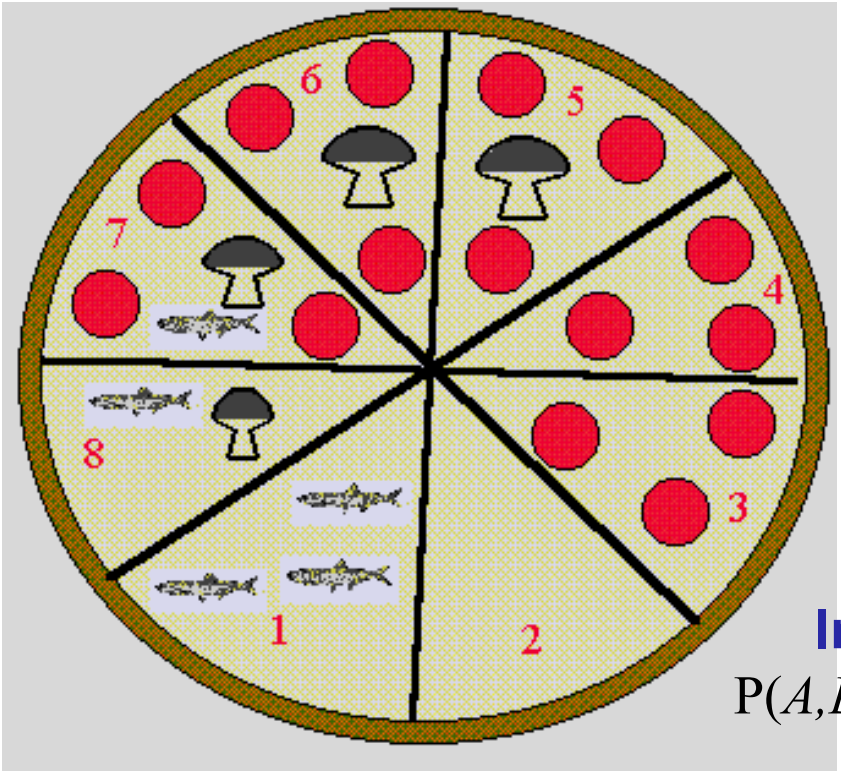
Conditional Independence:

$$\mathrm{P}(A,B|C) = \mathrm{P}(A|C)\mathrm{P}(B|C)$$

Are the presence of mushrooms and anchovies independent given pepperoni

$P(M|P) = ?$

Conditional Independence



Venn diagrams in
Pizza form from
B. Warner

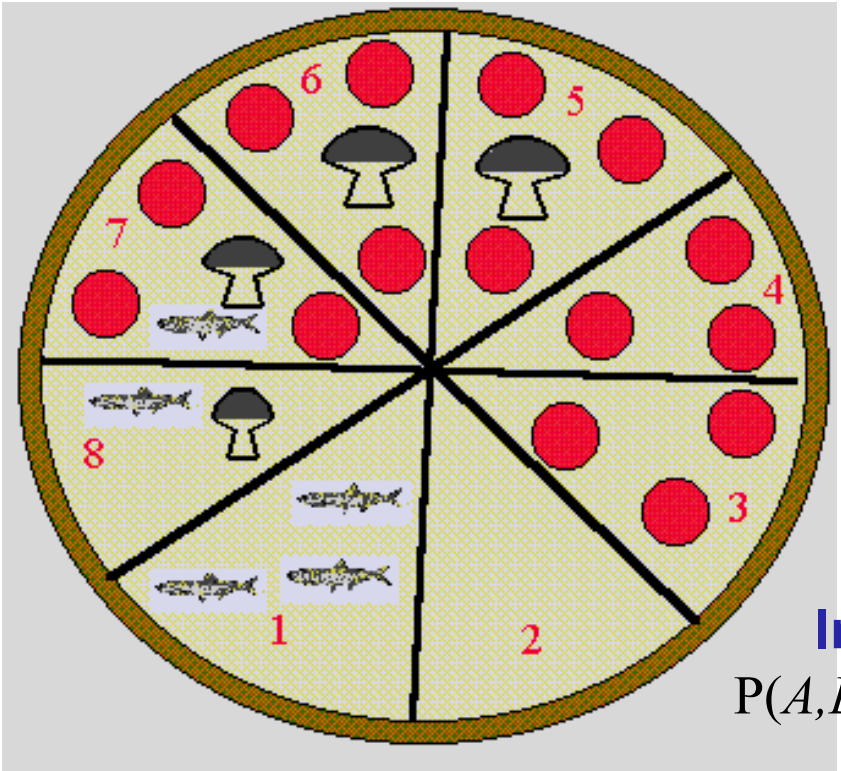
Conditional Independence:

$$\mathrm{P}(A,B|C) = \mathrm{P}(A|C)\mathrm{P}(B|C)$$

Are the presence of mushrooms and anchovies independent given pepperoni

$P(M|P) = 3/5$

Conditional Independence



Venn diagrams in
Pizza form from
B. Warner

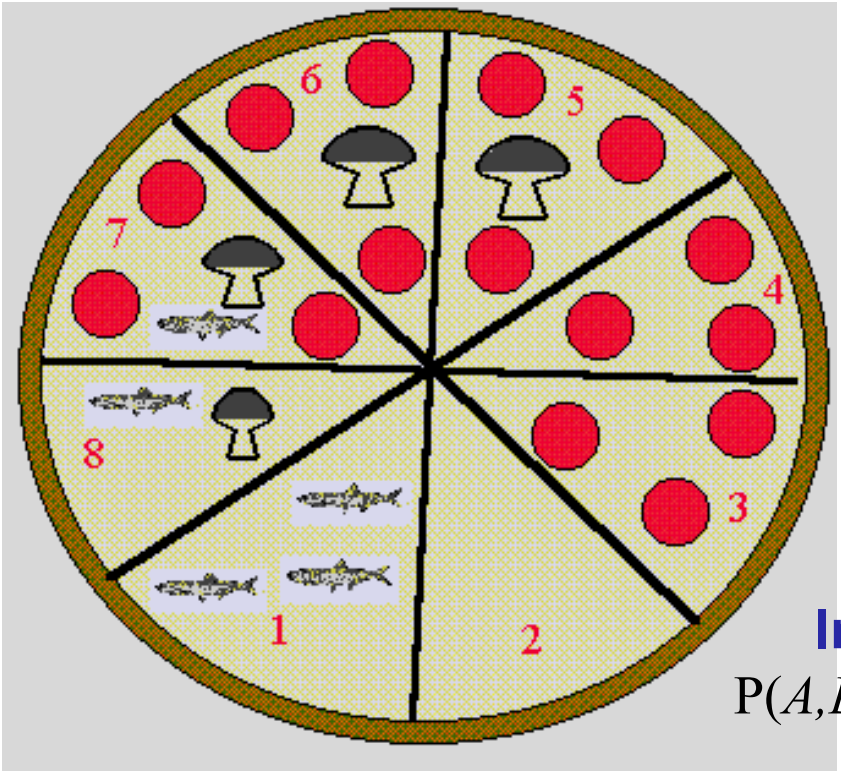
Conditional Independence:

$$\mathrm{P}(A,B|C) = \mathrm{P}(A|C)\mathrm{P}(B|C)$$

Are the presence of mushrooms and anchovies independent given pepperoni

$P(M|P) = 3/5 \; P(A|P) = ?$

Conditional Independence



Venn diagrams in
Pizza form from
B. Warner

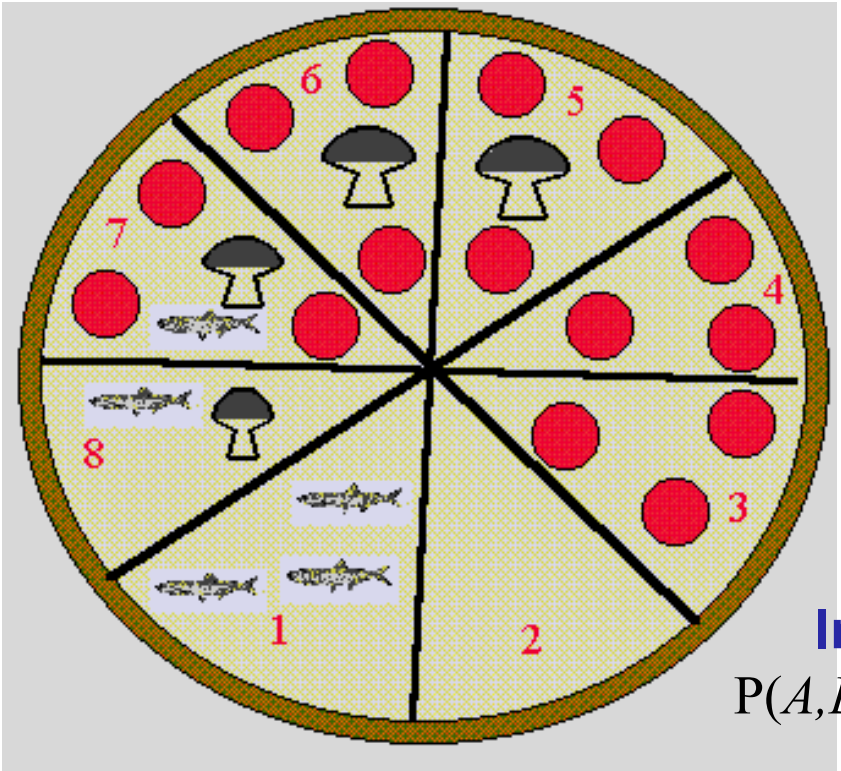
Conditional Independence:

$$\mathrm{P}(A,B|C) = \mathrm{P}(A|C)\mathrm{P}(B|C)$$

Are the presence of mushrooms and anchovies independent given pepperoni

$P(M|P) = 3/5$ $P(A|P) = 1/5$ $P(A, M|P) =$?

Conditional Independence



Venn diagrams in Pizza form from B. Warner

Conditional Independence:

$$\mathrm{P}(A,B|C) = \mathrm{P}(A|C)\mathrm{P}(B|C)$$

Are the presence of mushrooms and anchovies independent given pepperoni

$P(M|P) = 3/5 \; P(A|P) = 1/5 \; P(A, M|P) = 1/5$

so $P(A, M|P)$ is not equal to $P(M|P) \times P(A|P)$ thereform the presence of mushrooms and anchovies are not conditionally independent (given the presence of pepperoni).

Bayes Theorem

$$P(A, B) \quad = \quad P(A) * P(B|A)$$

Bayes Theorem

$$\begin{aligned} P(A,B) &= P(A) * P(B|A) \\ &= P(B) * P(A|B) \end{aligned}$$

Bayes Theorem

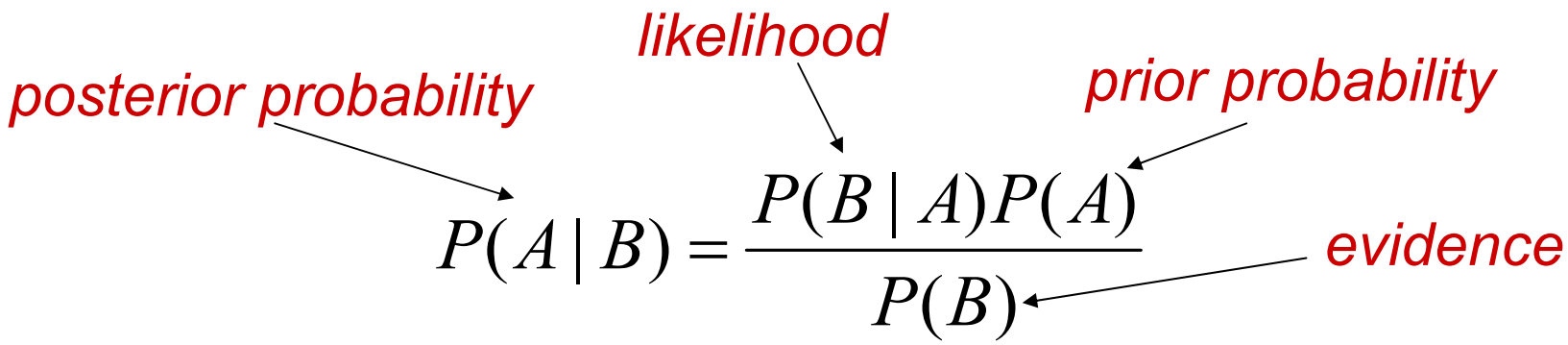
$$\begin{aligned} P(A,B) &= P(A) * P(B|A) \\ &= P(B) * P(A|B) \end{aligned}$$

Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

(proof above)

Bayes Theorem



P(B) is often computed as

$$P(B) = \sum_i P(B|A_i)P(A_i)$$

where A_i are all possible disjoint subsets

Bayes Rule in Biology

Given an image of an animal you have to determine whether the animal is a tiger or not

$$P(A = tiger|B = image_i) = \frac{P(B = image_i|A = tiger)P(A = tiger)}{P(B = image_i)}$$

Sample Problem (numbers made up)

The probability that an individual at the airport is a terrorist is 1 in 10 million ($1x10^{-6}$) Half the terrorists carry swiss army knifes. 10% of non-terrorists carry swiss army knifes. What's the probability that a knife carrier is a terrorist?

$P(terr) = .000001$ prior probability

$P(knife|terr) = .5$ likelihood

$P(knife| \sim terr) = .1$

compute

$$P(knife) = P(terr) * P(knife|terr) + P(\sim terr) * P(knife| \sim terr)$$

$$= .1$$

$$P(terr|knife) = \frac{P(knife|terr) * P(terr)}{P(knife)}$$

$$= .5 * .000001/.1$$
$$= 5 * 10^{-6}$$

Continuous probability density

For continuous random variables it does not make sense to talk about the probability of a particular value (which is equal to 0)

Instead we talk about probability density

$p(x)$ is a probability density over a continuous variable

$$Pr(x \in [a, b]) = \int_a^b p(x)dx$$

e.g. probability density of heights of females

Bayes Rule revisited

We can still have Bayes rule for continuous random variables. A common case is when we have different probability densities for different classes (ω_j)

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

- $P(\omega_j)$ = prior probability of ω_j
- $p(x)$ = evidence
- $P(\omega_j|x)$ = posterior probability of ω_j
- $p(x|\omega_j)$ = likelihood of ω_j with respect to x

Expected Value

Expected value or mean

$$E(X) = \sum xp(x)$$

$$E(X) = \int xp(x)dx$$

Variance

$$Var(X) = E(X - E(X))^2 = \sum P(X)(X - E(X))^2$$

$$Var(X) = \int (x - E(x))^2 p(x) dx$$

The Normal Density

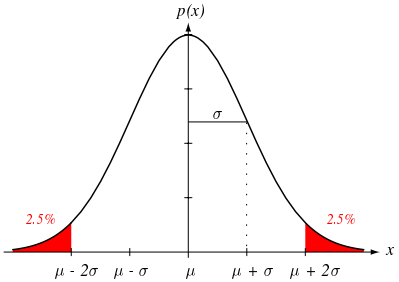


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Pattern Densities are commonly modeled by Normal Densities for several reasons

- Central Limit Theorem: sum of a large number of independent random variables is normally distributed applet
- It's analytically tractable!
- It's been well studied
- It has the maximum entropy of all distributions with a given mean and variance

Univariate normal density

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

has mean $= \mu$

variance $= \sigma^2$

has roughly 95% of its area within 2 standard deviations on either side of the mean (this is relevant for t-tests).

Standard Normal

Gaussian with mean 0 and variance 1

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

Multivariate normal density

$$p(\vec{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}(\det \Sigma)^{\frac{1}{2}}} e^{-\frac{(x-\mu)^T \Sigma^{-1}(x-\mu)}{2}}$$

Contours of constant density are defined by x such that

$$(x-\mu)'\Sigma^{-1}(x-\mu) = c^2$$

Ellipses are centered at μ with axes $\pm\sqrt{(\lambda_i)}e_i$ where λ_i and e_i are the eigenvalues and eigenvectors of Σ (this will be relevant for PCA and related algorithms)

- Linear combinations of the components of X are normally distributed
- All subsets of the components of X are normally distributed
- Zero covariance implies that the corresponding components are independent
- The conditional distributions of the components are multivariate normal

Multivariate Gaussians