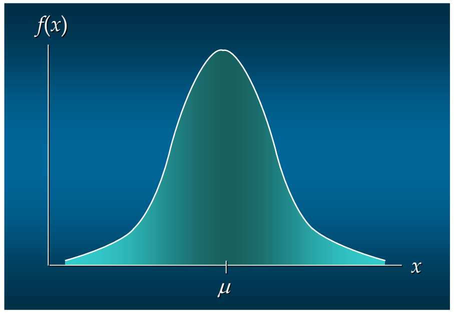
Machine Learning

10-701/15-781, Spring 2010

Tutorial on Basic Probability

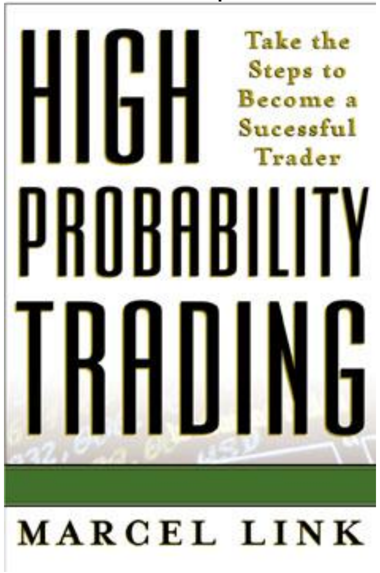
Field Cady



What is probability?

- Answer 1 : Our beliefs about the world
- Answer 2 : The random nature of the world

- “Probability theory is nothing but common sense reduced to calculation”
  - — Pierre Laplace, 1812.



- Either way, CRITICALLY important toolkit for ML and life
  - How confident is the robot that this object is a stapler?
  - My measurements have “noise”, i.e. random perturbations
  - What is the certainty threshold for acting on a belief?
  - Act so as to maximize “average” utility

Why use probability?

- There have been attempts to develop different methodologies for uncertainty:
  - Fuzzy logic
  - Qualitative reasoning (Qualitative physics)
  - …
- In 1931, de Finetti proved that it is irrational to have beliefs that violate probability axioms, in the following sense:
  - If you bet in accordance with your beliefs, but your beliefs violate the axioms, then you can be guaranteed to lose money to an opponent whose beliefs more accurately reflect the true state of the world. (Here, “betting” and “money” are proxies for “decision making” and “utilities”.)
- What if you refuse to bet? This is like refusing to allow time to pass: every action (including inaction) is a bet

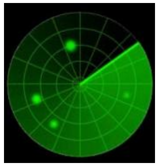
Basics of Formal Treatment of Probability

Basic Probability Concepts

- A *sample space* $\mathcal{S}$ is the set of all possible outcomes of a conceptual or physical, repeatable experiment. ($\mathcal{S}$ can be finite or infinite.)
  - E.g., $\mathcal{S}$ may be the set of all possible outcomes of a dice roll: $\mathcal{S} \equiv \{1,2,3,4,5,6\}$

  - E.g., $\mathcal{S}$ may be the set of all possible nucleotides of a DNA site: $\mathcal{S} \equiv \{A, T, C, G\}$

  - E.g., $\mathcal{S}$ may be the set of all possible time-space positions of a aircraft on a radar screen: $\mathcal{S} \equiv \{0, R_{max}\} \times \{0, 360^o\} \times \{0, +\infty\}$



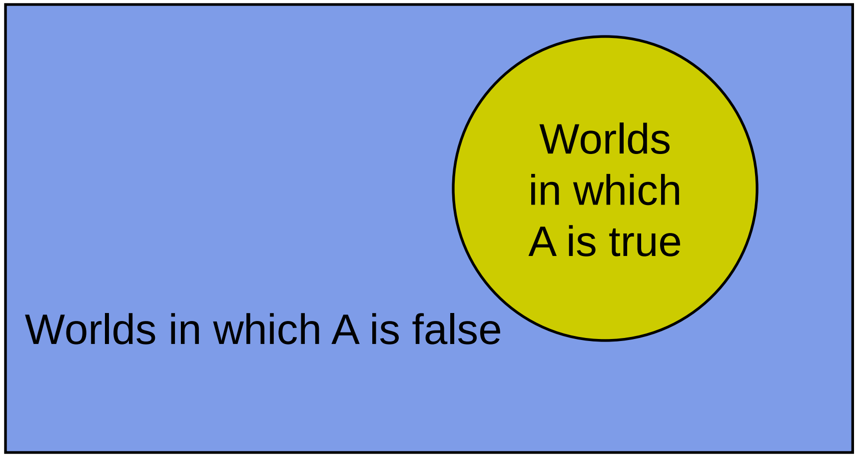
- An *event* $\mathcal{A}$ is any subset *of* $\mathcal{S}$:
  - Seeing "1" or "6" in a dice roll; observing a "G" at a site; UA007 in space-time interval
- If you want to be REALLY precise, use measure theory and set theory (I don't recommend this…)

Probability

- IMPORTANT HEURISTIC : Picture sample space as subset of the plane, and probabilities as areas
  - Most probability laws can be easily re-derived with this heuristic, and many are even obvious
- A *probability* $P(A)$ is a function that maps an event A onto the interval $[0, 1]$. $P(A)$ is also called the probability measure or probability mass of A.

Sample space of all possible worlds.

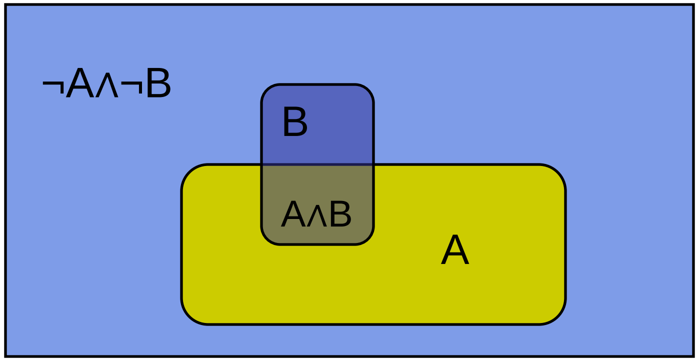
Its area is 1



$P(a)$ is the area of the oval

Kolmogorov Axioms

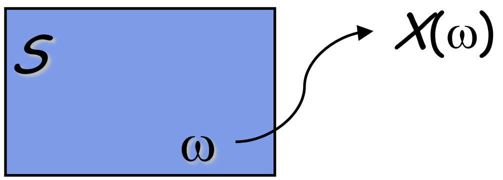
- All probabilities are between 0 and 1
  - $0 \leq P(A) \leq 1$
- $P(\mathcal{S}) = 1$
- $P(\Phi)=0$
- The probability of a disjunction is given by
  - $P(A \lor B) = P(A) + P(B) - P(A \land B)$



A∨B ?

Random Variable

- A *random variable* is a function that associates a unique numerical value (a token) with every outcome of an experiment. (The value of the r.v. will vary from trial to trial as the experiment is repeated)



  - Discrete r.v.:
    - The outcome of a dice-roll
    - Or the square of the outcome; equally valid
  - Continuous r.v.:
    - The outcome of **recording** the **true** location of an aircraft: X_{true}
    - The outcome of **observing** the **measured** location of an aircraft X_{obs}
  - Indicator r.v.:
    - "Indicates" whether or not event H happened
    - 1 if H happens, 0 otherwise
    - Example : X is a dice roll, and Y indicates whether X is even
    - Like True/False
    - E[Indicator] = Probability of H

Discrete/Continuous Distributions and Important Distributions

Discrete Prob. Distribution

- A probability distribution P defined on a discrete sample space $\mathcal{S}$ is an assignment of a non-negative real number $P(s)$ to each sample $s \in \mathcal{S}$ such that $\sum_{s \in \mathcal{S}} P(s) = 1$. $(0 \le P(s) \le 1)$
  - intuitively, $P(s)$ corresponds to the *frequency* (or the likelihood) of getting a particular sample s in the experiments, if repeated multiple times.
  - call $\theta_s = P(s)$ the *parameters* in a discrete probability distribution

- A discrete probability distribution is sometimes called a *probability model*, in particular if several different distributions are under consideration
  - write models as M_1, M_2, probabilities as $P(X|M_1)$, $P(X|M_2)$
  - e.g., M_1 may be the appropriate prob. dist. if X is from "fair dice", M_2 is for the "loaded dice".
  - M is usually a two-tuple of {dist. family, dist. parameters}

Discrete Distributions

- Bernoulli distribution: Ber(p)

$$P(x)=\begin{cases}1-p & \text{for } x=0\\ p & \text{for } x=1\end{cases} \quad \Rightarrow \quad P(x)=p^{x}(1-p)^{1-x}$$

- Multinomial distribution: Mult($1,\theta$)
  - Multinomial (indicator) variable:

$$X=\begin{bmatrix}X_1\\X_2\\X_3\\X_4\\X_5\\X_6\end{bmatrix}, \quad \text{where} \quad \begin{matrix} X_j=[0,1], \quad \text{and} \quad \sum_{j\in[1,\dots,6]} X_j=1 \\ X_j=1 \text{ w.p. } \theta_j, \quad \sum_{j\in[1,\dots,6]}\theta_j=1 \quad . \end{matrix}$$

$$p(x(j))=P\left(\{X_j=1, \text{where } j \text{ index the dice-face}\}\right)$$
$$=\theta_j=\prod_k \theta_k^{x_k}$$

Discrete Distributions

- Multinomial distribution: Mult(n, θ)
  - Count variable:

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix}, \qquad \text{where } \sum_j x_j = n$$

$$p(x) = \frac{n!}{x_1! x_2! \cdots x_K!} \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_K^{x_K} = \frac{n!}{x_1! x_2! \cdots x_K!} \theta^x$$

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Continuous Prob. Distribution

- A continuous random variable X is defined on a continuous sample space: an interval on the real line, a region in a high dimensional space, etc.
  - X usually corresponds to a real-valued measurements of some property, e.g., length, position, …
  - It is meaningless to talk about the probability of the random variable assuming a particular value --- $P(x) = 0$
  - Instead, we talk about the probability of the random variable assuming a value within a given interval, or half interval, or arbitrary Boolean combination of basic propositions.
    - $P(X \in [x_1, x_2])$
    - $P(X < x) = P(X \in [-\infty, x])$
    - $P(X \in [x_1, x_2] \cup [x_3, x_4])$

Probability Density

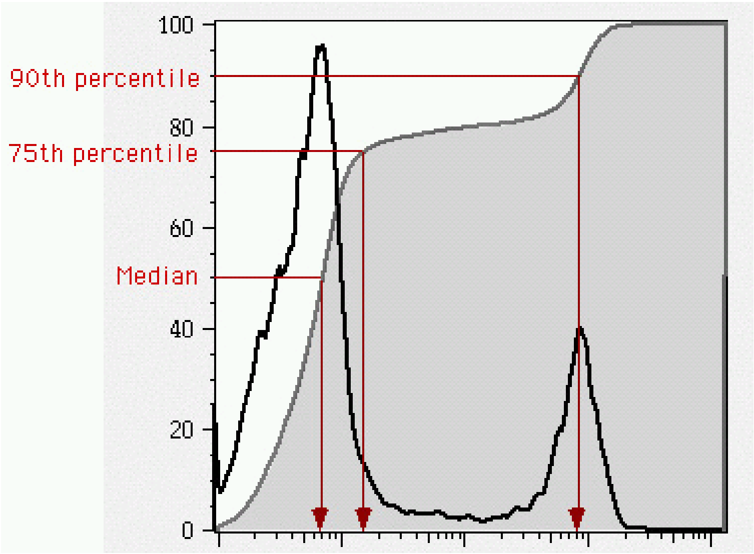
- If the prob. of *x* falling into *[x, x+dx]* is given by *p(x)dx* for *dx* , then *p(x)* is called the probability density over *x*.
- If the probability *P(x)* is differentiable, then the probability density over *x* is the derivative of *P(x)*.
  - The probability of the random variable assuming a value within some given interval from x_1 to x_2 is equivalent to the area under the graph of the probability density function between x_1 and x_2.
  - Probability mass: $P(X \in [x_1, x_2]) = \int_{x_1}^{x_2} p(x)dx$,

    note that $\int_{-\infty}^{+\infty} p(x)dx = 1.$
  - Cumulative distribution function (CDF):

    $$P(x) = P(X \le x) = \int_{-\infty}^{x} p(x')dx'$$
  - Probability density function (PDF):

    $$p(x) = \frac{d}{dx} P(x)$$

    $$\int_{-\infty}^{+\infty} p(x)dx = 1; \quad p(x) > 0, \forall x$$



Car flow on Liberty Bridge (cooked up!)

The intuitive meaning of *p(x)*

- If

  $p(x_1) = a$ and $p(x_2) = b$,

  then when a value X is sampled from the distribution with density p(x), you are a/b times as likely to find that X is "very close to" x_1 than that X is "very close to" x_2.

- That is :

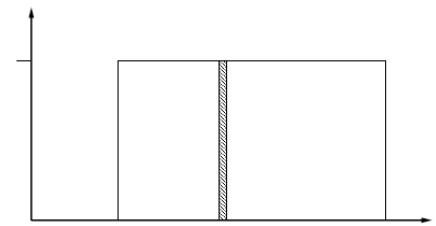
$$\lim_{h \to 0} \frac{P(x_1 - h < X < x_1 + h)}{P(x_2 - h < X < x_2 + h)} = \frac{\int_{x_1 - h}^{x_1 + h} p(x)dx}{\int_{x_2 - h}^{x_2 + h} p(x)dx} = \frac{p(x_1) \times 2h}{p(x_2) \times 2h} = a/b$$

- Alternately : $p(x_1)dx$ = Pr(X in the interval $[x_1, x_1+dx)$)

Continuous Distributions

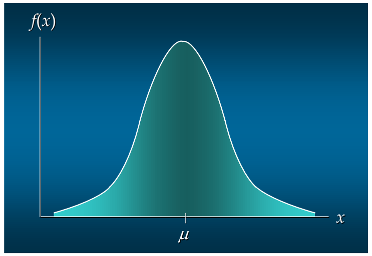
- Uniform Density Function

$$p(x) = 1/(b-a) \quad \text{for } a \le x \le b$$
$$= 0 \quad \text{elsewhere}$$



- Normal (Gaussian) Density Function

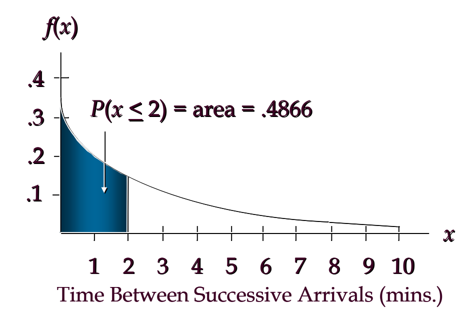
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$



  - The distribution is <u>symmetric</u>, and is often illustrated as a <u>bell-shaped curve</u>.
  - <u>Two parameters</u>, μ (mean) and σ (standard deviation), determine the location and shape of the distribution.
  - The <u>highest point</u> on the normal curve is at the mean, which is also the median and mode.

- Exponential Distribution

$$\text{PDF: } p(x) = \frac{1}{\mu} e^{-x/\mu}, \qquad \text{CDF: } P(x \le x_0) = 1 - e^{-x_0/\mu}$$



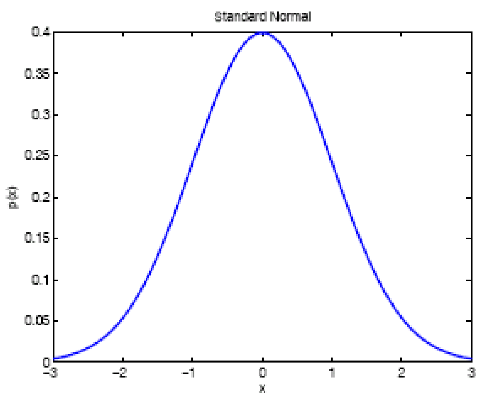
Gaussian (Normal) density in 1D

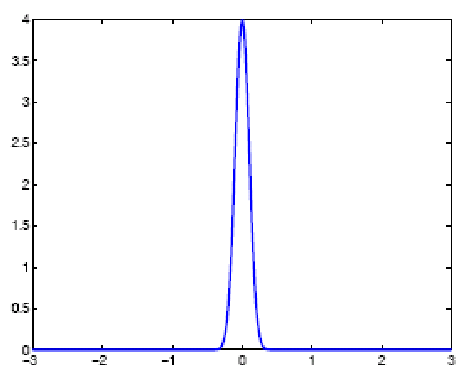
- If $X \sim N(\mu, \sigma^2)$, the probability density function (pdf) of X is defined as

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

  - We will often use the precision $\lambda = 1/\sigma^2$ instead of the variance σ^2.
  - Here is how we plot the pdf in matlab

```
xs=-3:0.01:3;
plot(xs,normpdf(xs,mu,sigma));
```





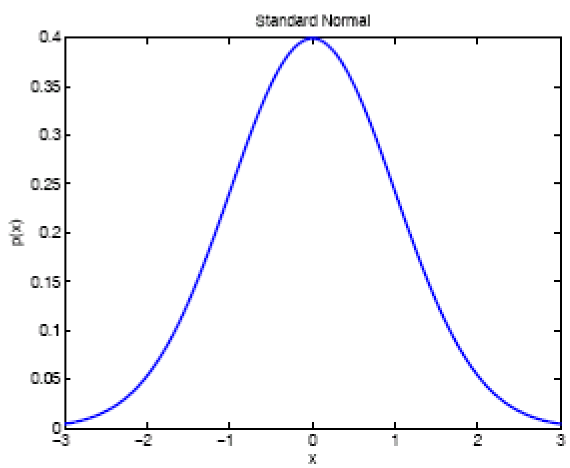
- Note that a density evaluated at a point can be larger than 1.

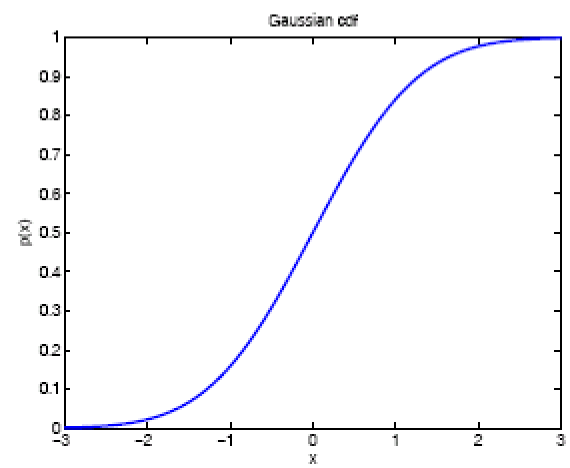
Gaussian CDF

- If $Z \sim N(0, 1)$, the cumulative density function is defined as

$$\Phi(x) = \int_{-\infty}^{x} p(z)dz$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2} dz$$

- This has no closed form expression, but is built in to most software packages (eg. normcdf in matlab stats toolbox).





More on Gaussian Distribution

- If $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0, 1)$.

- How much mass is contained inside the $[-2\sigma, 2\sigma]$ interval?

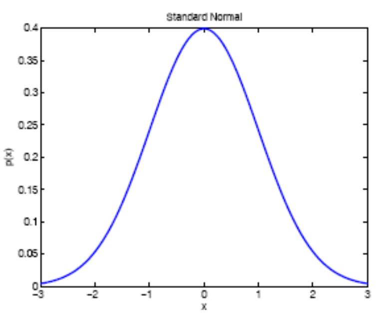
$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

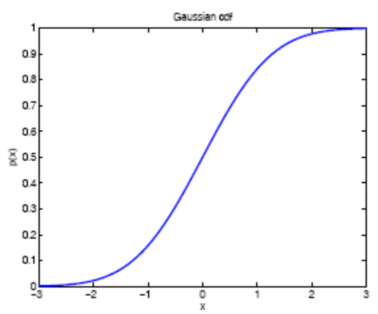
- Since

  $p(Z \leq -2) = \text{normcdf}(-2) = 0.025$

  we have

  $P(-2\sigma < X - \mu < 2\sigma) \approx 1 - 2 \times 0.025 = ($





Statistical Characterizations

- **Expectation:** the centre of mass, mean value, first moment):

$$E(X) = \begin{cases} \sum_{i \in S} x_i p(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} x p(x) dx & \text{continuous} \end{cases}$$

  - Sample mean: $$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- **Variance:** the spreadness:

$$Var(X) = \begin{cases} \sum_{x \in S} [x_i - E(X)]^2 p(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} [x - E(X)]^2 p(x) dx & \text{continuous} \end{cases}$$

  - Sample variance $$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2$$

Central limit theorem

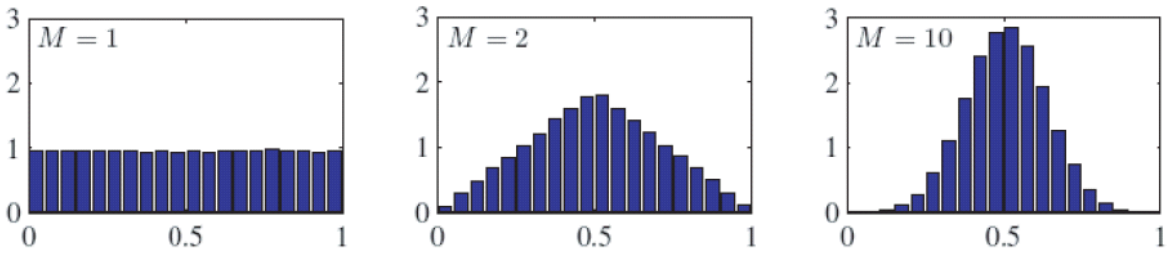
- If (X_1 ,X_2, … X_n) are i.i.d. continuous random variables

- Define

$$\bar{X} = f(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- As n → infinity,

$p(\bar{X})$ →Gaussian with mean E[X_i] and variance Var[X_i]



- Somewhat of a justification for assuming Gaussian noise is common

Hybrid Prob. Distributions

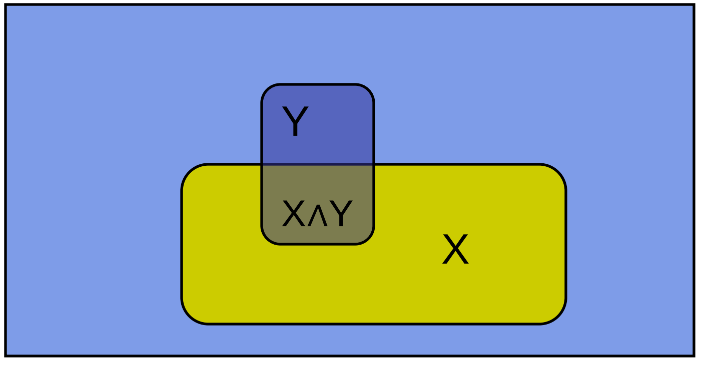
- World of probability not limited to Continuous and Discrete!
- Example : Elevation of airplane
  - Elevation=0 (airplane landed) with finite probability, like a discrete r.v.
  - But for >0, it's continuous
- Example : Measured data where detector can saturate
  - Non-zero probability you measure the maximum temperature on thermometer
  - Perhaps saturation yields string "ERROR"; not even a number!
- In practice, we rarely or never use hybrid distributions. But it's nice to know they're there ☺

Things You Can Do with a Probability Distribution

Elementary manipulations of probabilities

- Set probability of multi-valued r.v.
  - $P(\{x=Odd\}) = P(1)+P(3)+P(5) = 1/6+1/6+1/6 = ½$
  - $P(X=x_1 \vee X=x_2,\ldots,\vee X=x_i)=\sum_{j=1}^{i} P(X=x_j)$
- Multi-variant distribution:
  - **Joint probability**: $P(X=true \wedge Y=true)$
  - **Marginal Probability:** $P(Y)=\sum_{j\in S} P(Y\wedge X=x_j)$

$$P(Y\wedge\{X=x_1\vee X=x_2,\ldots,\vee X=x_i\})=\sum_{j=1}^{i}P(Y\wedge X=x_j)$$



Joint Probability

- A joint probability distribution for a set of RVs gives the probability of every atomic event (sample point)

  - **P**(*Flu,HeadAche*) = a 2 $\times$ 2 matrix of values:

	H	¬H
F	0.005	0.02
¬F	0.195	0.78

  - **Every question about a domain can be answered by the joint distribution**, as we will see later.

Conditional Probability

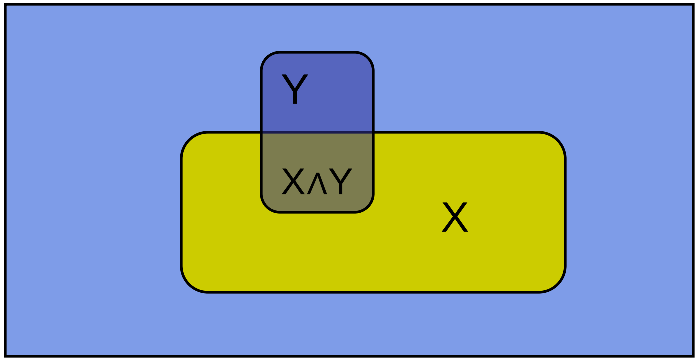
- P(X|Y) = Probability of X, IF we know that Y is true
  - H = "having a headache"
  - F = "coming down with Flu"
    - P(H)=1/10
    - P(F)=1/40
    - P(H|F)=1/2
  - P(H|F) = fraction of flu-inflicted worlds in which you have a headache

    = P(H∧F)/P(F)
- Equivalently : Fraction of worlds in which Y is true that also have X true

$$P(X \mid Y) = \frac{P(X \wedge Y)}{P(Y)}$$

- Definition:
  - Corollary: The Chain Rule

$$P(X \wedge Y) = P(X \mid Y) P(Y)$$



Condition formula is obvious when you use a picture!

- This is all fine and dandy if we already *know* the probability distribution

- But the real world we don’t have distributions : we have DATA

Guessing Probability Distributions (Educatedly)

Density Estimation

- You have some real-world data
- You need an “educated guess” about the distribution
- What do you do??

- There’s no one right answer, but two common approaches are
  - Bayesian : Start with a “reasonable guess” about the distribution Then update your guess based on observations.
  - Frequentist : YOU ARE IGNORANT! Only the data is ground truth! Choose a distribution for which the data you see is as likely as possible.

- Much of machine learning boils down to just estimating distributions

Bayesian

- How do you pick a “reasonable guess”?
  - “I go to CMU – of course I’m smart enough to come up with a great guess! Come to think of it, why even use data?” ☺
  - “ I know a domain expert – maybe I can use their knowledge”
  - “I know nothing! Use maximum entropy distribution” (more on entropy later)
- How do you update your guess?
  - Bayes rule : more on that later
- Advantages
  - Can use domain expertise
  - Not skewed as much by outlier data
- Disadvantages
  - You add your own bias

Frequentist

- Maximum Likelihood Estimation
  - Assume data follow a parameterized distribution
    - Bernoulli with probability p
    - Normal with some mean and variance
  - Choose parameters θ that minimize

$$P(X_1, X_2, \ldots, X_n \mid \theta)$$

$$= \prod_{j=1}^{n} P(X = x_j \mid \theta)$$ if data independent

- Advantages
  - Very principled
  - Not skewed as much by outlier data
- Disadvantages
  - Overfitting!

Maximum Likelihood Estimation

- Goal: estimate distribution parameters $\boldsymbol{\theta}$ from a dataset of N independent, identically distributed (*iid*), fully observed, training cases

$$D = \{x_1, \ldots, x_N\}$$

- Maximum likelihood estimation (MLE)
  1. Write probability of data as a function of parameters:

$$\begin{aligned} L(\theta) &= P(x_{1,} x_2, \ldots, x_N; \theta) \\ &= P(x; \theta) P(x_2; \theta), \ldots, P(x_N; \theta) \\ &= \prod_{i=1}^{N} P(x_i; \theta) \end{aligned}$$

  2. Find the maximum of this function, usually just using calculus

$$\theta^* = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \log L(\theta)$$

Often logs make the math easier; answer is the same

Example 1: Bernoulli model

- Data:
  - We observed *N* ***iid*** coin tossing: *D*={1, 0, 1, …, 0}
- Representation:

Binary r.v: $$x_n = \{0,1\}$$

- Model: $$P(x) = \begin{cases} 1-p & \text{for } x=0 \\ p & \text{for } x=1 \end{cases} \quad \Rightarrow \quad P(x) = \theta^x (1-\theta)^{1-x}$$

- How to write the likelihood of a single observation x_i ?

$$P(x_i) = \theta^{x_i} (1-\theta)^{1-x_i}$$

- The likelihood of dataset $D=\{x_1, \ldots, x_N\}$:

$$P(x_1, x_2, \ldots, x_N \mid \theta) = \prod_{i=1}^{N} P(x_i \mid \theta) = \prod_{i=1}^{N} \left(\theta^{x_i} (1-\theta)^{1-x_i} \right) = \theta^{\sum_{i=1}^{N} x_i} (1-\theta)^{\sum_{i=1}^{N} 1-x_i} = \theta^{\#\text{head}} (1-\theta)^{\#\text{tails}}$$

MLE for discrete (joint) distributions

- More generally, it is easy to show that

$$P(\text{event}_i) = \frac{\#\text{records in which event}_i \text{ is true}}{\text{total number of records}}$$

- This is an important (but sometimes not so effective) learning algorithm!

¬F	¬B	¬H	0.4	
¬F	¬B	H	0.1	
¬F	B	¬H	0.17	
¬F	B	H	0.2	
F	¬B	¬H	0.05	
F	¬B	H	0.05	
F	B	¬H	0.015	
F	B	H	0.015	

- Overfitting : what if, by chance, some event never occurs?
  - You flip a coin ONCE a get a head. Does that mean tails are *impossible*?

Example 2: univariate normal

- Data:
  - We observed N ***iid*** real samples:
    D={-0.1, 10, 1, -5.2, …, 3}
- Model: $$P(x) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-(x-\mu)^2 / 2\sigma^2\right\}$$
- Log likelihood:

$$\mathcal{L}(\mu, \sigma^2) = \log P(D \mid \theta) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{n=1}^{N}\frac{\left(x_n-\mu\right)^2}{\sigma^2}$$

- MLE: take derivatives and set to zero:

$$\frac{\partial \mathcal{L}}{\partial \mu} = (1/\sigma^2)\sum_n \left(x_n - \mu\right)$$

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_n \left(x_n - \mu\right)^2$$

$\Rightarrow$

$$\mu_{MLE} = \frac{1}{N}\sum_n \left(x_n\right)$$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_n \left(x_n - \mu_{ML}\right)^2$$

Overfitting

- Recall that for Bernoulli Distribution, we have

$$\widehat{\theta}_{ML}^{head} = \frac{n^{head}}{n^{head} + n^{tail}}$$

- What if we tossed too few times so that we saw zero head?

  We have $\widehat{\theta}_{ML}^{head} = 0,$ and we will predict that the probability of seeing a head next is zero!!!

- The rescue:
  - Where n' is know as the pseudo- (imaginary) count

$$\widehat{\theta}_{ML}^{head} = \frac{n^{head} + n'}{n^{head} + n^{tail} + n'}$$

  - But that's pretty hacky…
  - It's related to "hierarchical Bayesian models", where you put a prior probability distribution on the parameters – more on that later

Overfitting

- So vanilla MLE is problematic for *discrete* distributions because, by chance, some event might not happen
  - That's an advantage to the Bayesian approach, IF your initial guess makes all events possible
- What about continuous distributions, where we estimate parameters?
  - Overfitting still a problem
  - For Normal distribution, for example, you underestimate the variance
- Unpleasant choice : Introduce our own biases, or over fit to the data? ☹
  - There are ways to partly work around this, but most of them are beyond this class
  - Bottom line : it's art as well as science, and nature doesn't furnish a "best answer". So we make due with what we have, which has been quite successful so far.

Entropy

"The tendency for entropy to increase in isolated systems is expressed in the second law of thermodynamics — perhaps the most pessimistic and amoral formulation in all human thought."

— Gregory Hill and Kerry Thornley

I tell you P(H) for a coin. How well can you predict it?

- P=1
  - You'll always guess H, and always be right
  - This distribution has "no uncertainty"
- P=.7 (or .3)
  - You'll guess H (or T), and you'll *usually* be right
  - "medium uncertainty"
- P=.5
  - Distribution is useless; you're right half the time no matter what
  - "high uncertainty"
- "Entropy" is a formal version of this uncertainty

Entropy of a Distribution

- Definition
  - Imagine you need to communicate observations of a random variable
  - Rare outcomes are more surprising; they contain more "information"
  - Useful definition $Information(X_i) = -\ln P(X_i)$
    - Information adds
    - Certain event has no information
  - Entropy is average information of a message

$$H(X) = E[-\ln P(X)] = -\sum_{x_j} P(x_j)\ln(P(x_j))$$

- High entropy = rare events more common
- Entropy comes from Information Theory
  - Sample space = letters to be encoded
  - Use fewer bits for common letters, more for rare to save space
  - Entropy = min. average number of bits to encode a letter

Claude Shannon invented information theory – and the motorized pogo stick

Entropy of a Distribution

- Definition valid for discrete distributions
- Does not generalize to continuous distributions
  - How would you encode a language with a continuum of letters in bits? That would be really weird.
- But there is a similar concept
  - Differential entropy $H(X) = -E[\ln p(X)] = -\int p(x) \ln(p(x)) dx$
  - Different in subtle but important ways
  - Some, but not all, of the same uses
- Hybrid distributions : Don't even try
- This class : Only discrete distributions

"You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. **In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage."**
- John von Neumann

Back to the Coin Example

- Bernoulli Random Variable
  - $H(X) = P(head)(-\ln P(head)) + P(tail)(-\ln P(tail))$
    $= p(-\ln p) + (1-p)(-\ln(1-p))$
    $= -p\ln p - q\ln q$
- P=1
  - H(X) = 0.0
- P=.7 (or .3)
  - H(X) = 0.360201221 + 0.521089678 = .88129
- P=.5
  - H(X) = 2(.5) = 1.0

Uses of Entropy

- Density Estimation
  - Distribution with max. entropy is "least informative"
  - If we have no idea what the distribution is but we need to make a guess, guess the one that is least informative

- Decision Trees :
  - How do you pick a root node for a decision tree?
  - Pick the "most informative" attribute A

    i.e. on average, distribution after seeing A has less entropy
  - $Gain(X, A) = H(X) - P(A=0)H(X \mid A=0) - P(A=1)H(X \mid A=1)$
  - Pick A with maximum gain

Bayes Theorem and How it will Change Your Life (in a good way!)

The Bayes Rule

- What we have just done leads to the following general expression:

$$P(Y \mid X) = \frac{P(X \mid Y)\, p(Y)}{P(X)} = p(Y) \times \left(\frac{P(X \mid Y)}{P(X)} \right)$$

  This is Bayes Rule.

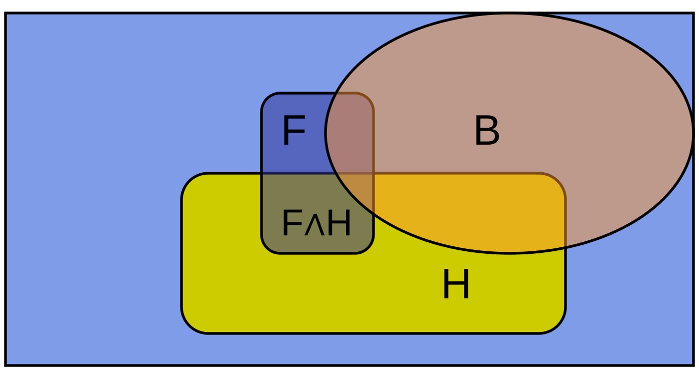
- We use it a *lot*
- Probability of Y is "updated" after observation of X
- Key element : direction of conditioning reversed
  - P(X|Y) is easy to calculate if Y is a parameter for a model of X

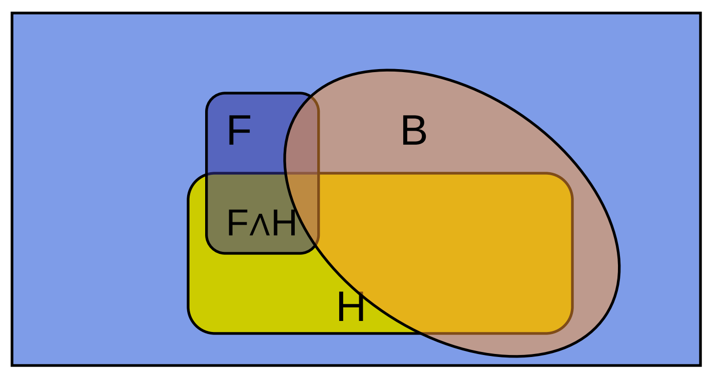
Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

More General Forms of Bayes Rule

- $$P(Y \mid X) = \frac{P(X \mid Y)p(Y)}{P(X \mid Y)p(Y) + P(X \mid \neg Y)p(\neg Y)}$$
- $$P(Y = y_i \mid X) = \frac{P(X \mid Y)p(Y)}{\sum_{i \in S} P(X \mid Y = y_i)p(Y = y_i)}$$
- $$P(Y \middle| X \wedge Z) = \frac{P(X \mid Y \wedge Z)p(Y \wedge Z)}{P(X \wedge Z)} = \frac{P(X \mid Y \wedge Z)p(Y \wedge Z)}{P(X \mid \neg Y \wedge Z)p(\neg Y \wedge Z) + P(X \mid \neg Y \wedge Z)p(\neg Y \wedge Z)}$$

  - P(Flu | HeadAche ∧ DrankBeer)





Probabilistic Inference : Using Observations

- H = "having a headache"
- F = "coming down with Flu"
  - P(H)=1/10
  - P(F)=1/40
  - P(H|F)=1/2

- One day you wake up with a headache. You come up with the following reasoning: "since 50% of flus are associated with headaches, I must have a 50-50 chance of coming down with flu”

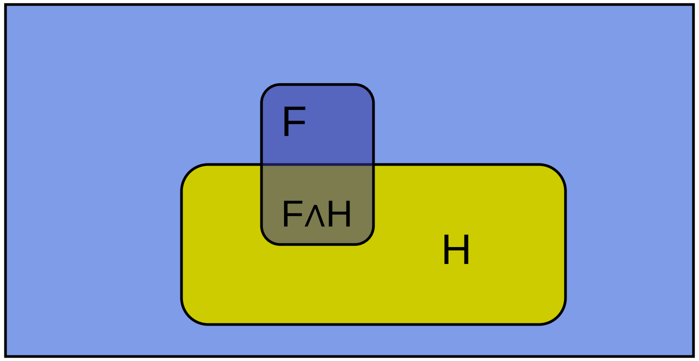
  Is this reasoning correct? NO!!!

Probabilistic Inference

- H = "having a headache"
- F = "coming down with Flu"
  - P(H)=1/10
  - P(F)=1/40
  - P(H|F)=1/2

- The Problem:

$$P(F \mid H) = ?$$



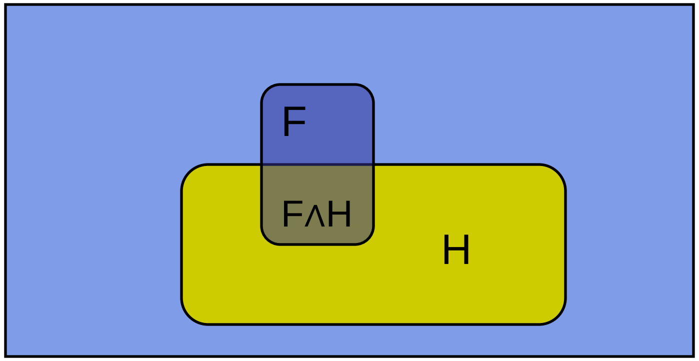
Probabilistic Inference

- H = "having a headache"
- F = "coming down with Flu"
  - P(H)=1/10
  - P(F)=1/40
  - P(H|F)=1/2

- The Answer:

$$P(F \mid H) = p(F)\frac{P(H \mid F)}{P(H)}$$

$$= (1/2)\frac{(1/40)}{(1/10)} = \frac{1}{8}$$



- 1/40 is the Prior, 1/8 is the Posterior
  - Probabilities before and after observation

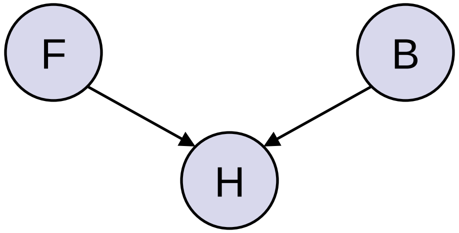
Posterior conditional probability

- Conditional or posterior (see later) probabilities
  - e.g., *P*(*Flu*|*Headache*) = 0.125
  - → **given that *flu* is all I know**

    NOT "if *flu* then 12.5% chance of *Headache*"
- Representation of conditional distributions:
  - **P**(*Flu*|*Headache*) = 2-element vector of 2-element vectors
- If we know more, e.g., DrinkBeer is also given, then we have
  - **P**(*Flu*|*Headache,DrinkBeer*) = 0.070 **This effect is known as explain away!**
  - **P**(*Flu*|*Headache,Flu*) = 1
  - Note: the less or more certain belief remains valid after more evidence arrives, but is not always useful
- New evidence may be irrelevant, allowing simplification, e.g.,
  - **P**(*Flu*|*Headache,StealersWin*) = **P**(*Flu*|*Headache*)
  - This kind of inference, sanctioned by domain knowledge, is crucial

Prior Distribution

- Suppose that our random variables have a "causal flow"
  - e.g.,

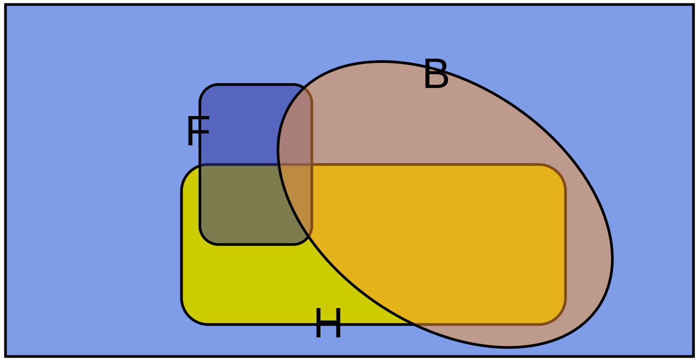


- Typically know probability distribution for a node, given the values of its parents
- Knowledge of one node gives information about its parents, via Bayes Rule
- Knowing F
  - by itself - says nothing about B
  - if you know H - gives information about B by explaining away

Inference by enumeration

- Start with a Joint Distribution
- Building a Joint Distribution of M=3 variables
  - Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
  - For each combination of values, say how probable it is.
  - Normalized, i.e., sums to 1

F	B	H	Prob
0	0	0	0.4
0	0	1	0.1
0	1	0	0.17
0	1	1	0.2
1	0	0	0.05
1	0	1	0.05
1	1	0	0.015
1	1	1	0.015

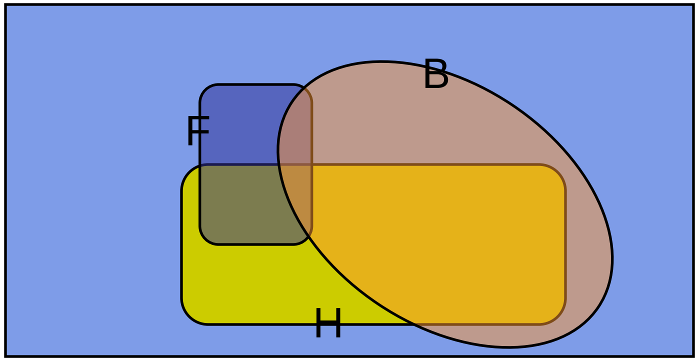


Inference with the Joint

- Once you have the JD you can ask for the probability of any atomic event consistent with you query

$$P(E)=\sum_{i\in E}P(row_i)$$

¬F	¬B	¬H	0.4
¬F	¬B	H	0.1
¬F	B	¬H	0.17
¬F	B	H	0.2
F	¬B	¬H	0.05
F	¬B	H	0.05
F	B	¬H	0.015
F	B	H	0.015

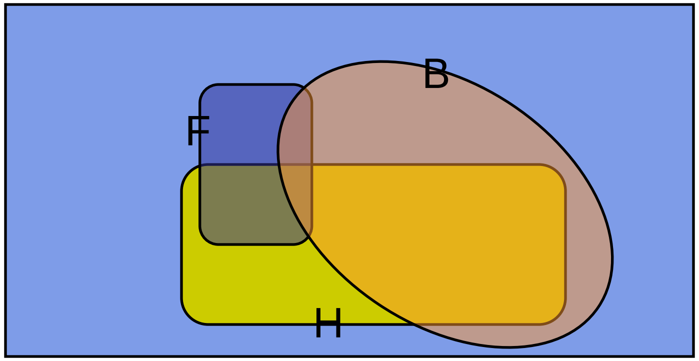


Inference with the Joint

- Compute Marginals

$$P(\text{Flu} \wedge \text{Headache}) =$$

¬F	¬B	¬H	0.4
¬F	¬B	H	0.1
¬F	B	¬H	0.17
¬F	B	H	0.2
F	¬B	¬H	0.05
F	¬B	H	0.05
F	B	¬H	0.015
F	B	H	0.015

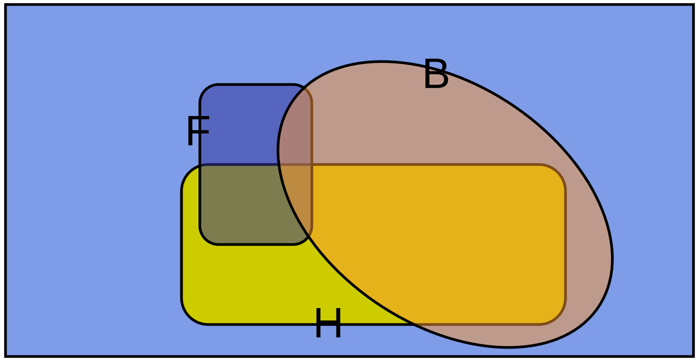


Inference with the Joint

- Compute Marginals

$$P(\text{Headache}) =$$

¬F	¬B	¬H	0.4
¬F	¬B	H	0.1
¬F	B	¬H	0.17
¬F	B	H	0.2
F	¬B	¬H	0.05
F	¬B	H	0.05
F	B	¬H	0.015
F	B	H	0.015



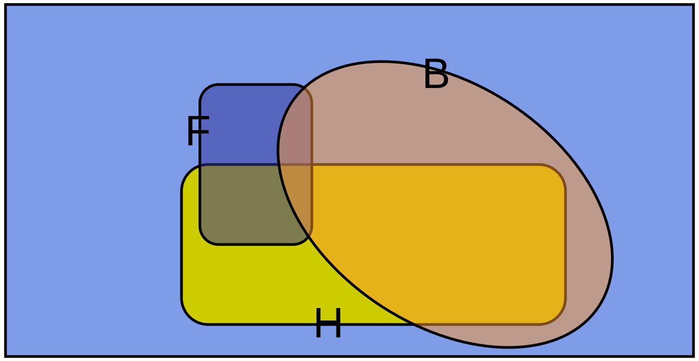
Inference with the Joint

- Compute Conditionals

$$P(E_1 | E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)}$$

$$= \frac{\sum_{i \in E_1 \cap E_2} P(row_i)}{\sum_{i \in E_2} P(row_i)}$$

¬F	¬B	¬H	0.4
¬F	¬B	H	0.1
¬F	B	¬H	0.17
¬F	B	H	0.2
F	¬B	¬H	0.05
F	¬B	H	0.05
F	B	¬H	0.015
F	B	H	0.015



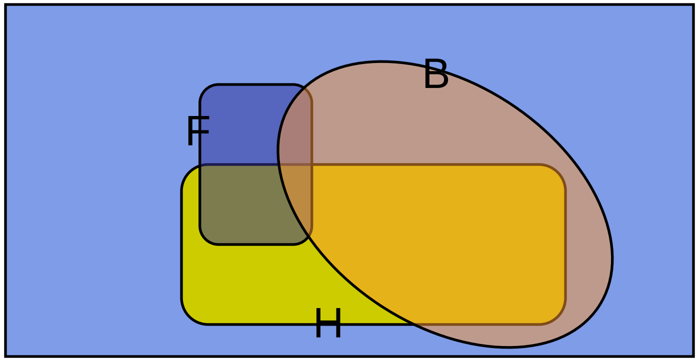
Inference with the Joint

- Compute Conditionals

$$P(\text{Flu} \mid \text{HeadAche}) = \frac{P(\text{Flu} \wedge \text{HeadAche})}{P(\text{HeadAche})}$$

¬F	¬B	¬H	0.4
¬F	¬B	H	0.1
¬F	B	¬H	0.17
¬F	B	H	0.2
F	¬B	¬H	0.05
F	¬B	H	0.05
F	B	¬H	0.015
F	B	H	0.015

- General idea: compute distribution on query variable by **fixing** **evidence variables** and **summing** over **hidden variables**



Summary: Inference by enumeration

- Let X be all the variables. Typically, we want
  - the posterior joint distribution of the query variables Y
  - given specific values e for the evidence variables E
  - Let the hidden variables be H = X-Y-E
- Then the required summation of joint entries is done by summing out the hidden variables:

$$P(Y|E=e)=\alpha P(Y,E=e)=\alpha\sum_h P(Y,E=e, H=h)$$

- The terms in the summation are joint entries because Y, E, and H together exhaust the set of random variables
- Obvious problems:
  - Worst-case time complexity *$O(d^n)$* where d is the largest arity
  - Space complexity *$O(d^n)$* to store the joint distribution
  - How to find the numbers for *$O(d^n)$* entries???

Using Independence to Simplify Calculations

Rules of Independence --- by examples

- P(Virus | DrinkBeer) = **P(Virus)**

  iff Virus is independent of DrinkBeer

- P(Flu | Virus;DrinkBeer) = **P(Flu|Virus)**

  iff Flu is independent of DrinkBeer, given Virus

- P(Headache | Flu;Virus;DrinkBeer) = **P(Headache|Flu;DrinkBeer)**

  iff Headache is independent of Virus, given Flu and DrinkBeer

Conditional independence

- Write out full joint distribution using chain rule:

  P(Headache;Flu;Virus;DrinkBeer)

  = P(Headache | Flu;Virus;DrinkBeer) P(Flu;Virus;DrinkBeer)

  = P(Headache | Flu;Virus;DrinkBeer) P(Flu | Virus;DrinkBeer) P(Virus | DrinkBeer) P(DrinkBeer)

Assume independence and conditional independence

= **P(Headache|Flu;DrinkBeer) P(Flu|Virus) P(Virus) P(DrinkBeer)**

I.e., ? independent parameters

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from **exponential** in *n* to **linear** in *n*.
- Conditional independence is our most basic and robust form of knowledge about uncertain environments.

Marginal and Conditional Independence

- Recall that for events E (i.e. $X=x$) and H (say, $Y=y$), the conditional probability of E given H, written as $P(E|H)$, is

$$P(E \text{ and } H)/P(H)$$

(= the probability of both E and H are true, given H is true)

- E and H are (statistically) independent if

$$P(E) = P(E|H)$$

(i.e., prob. E is true doesn't depend on whether H is true); or equivalently

$$P(E \text{ and } H)=P(E)P(H).$$

- E and F are *conditionally* independent given H if

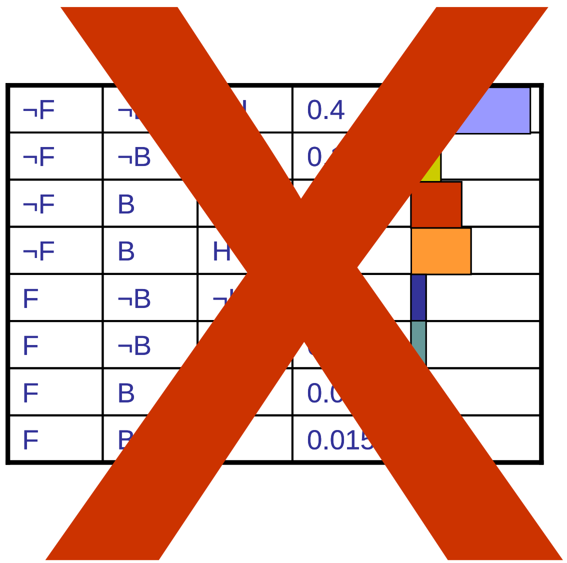
$$P(E|H,F) = P(E|H)$$

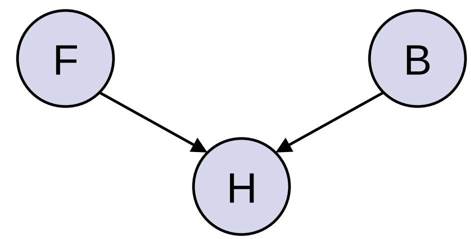
or equivalently

$$P(E,F|H) = P(E|H)P(F|H)$$

Why knowledge of Independence is useful

- Lower complexity (time, space, search …)





- Motivates efficient inference for all kinds of queries

  Stay tuned !!

- Structured knowledge about the domain
  - easy to learning (both from expert and from data)
  - easy to grow

Where do probability distributions come from?

- Idea One: Human, Domain Experts
- Idea Two: Simpler probability facts and some algebra

e.g., P(F)

P(B)

P(H|¬F,B)

P(H|F,¬B)

…

⇒

¬F	¬B	¬H	0.4
¬F	¬B	H	0.1
¬F	B	¬H	0.17
¬F	B	H	0.2
F	¬B	¬H	0.05
F	¬B	H	0.05
F	B	¬H	0.015
F	B	H	0.015

- Idea Three: Learn them from data!
  - A good chunk of this course is essentially about various ways of learning various forms of them!

The Bayesian Theory

- The Bayesian Theory: (e.g., for data D and model M)

$$P(M|D) = P(D|M)P(M)/P(D)$$

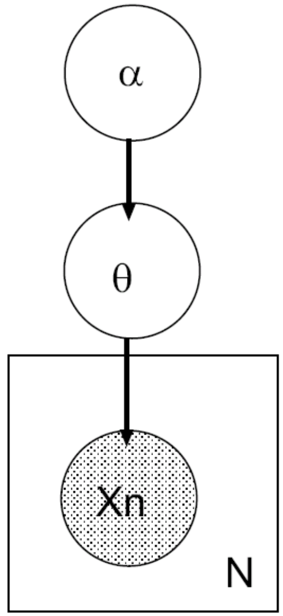
  - the **posterior** equals to the **likelihood** times the **prior**, up to a constant.

- This allows us to capture uncertainty about the model in a principled way

Hierarchical Bayesian Models

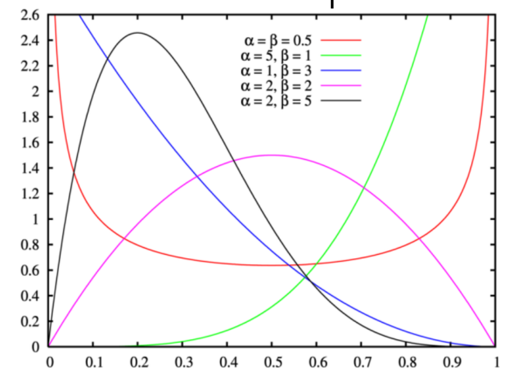
- θ are the parameters for the likelihood $p(x|\theta)$
- α are the parameters for the prior $p(\theta|\alpha)$.
- We can have hyper-hyper-parameters, etc.
- We stop when the choice of hyper-parameters makes no difference to the marginal likelihood; typically make hyper-parameters constants.
- Where do we get the prior?
  - Intelligent guesses
  - Empirical Bayes (Type-II maximum likelihood)
  - → computing point estimates of α :

$$\hat{\vec{\alpha}}_{MLE} = \arg\max_{\vec{\alpha}} = p(\vec{n} \mid \vec{\alpha})$$



Bayesian estimation for Bernoulli

- Beta distribution:



$$P(\theta;\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}=B(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Posterior distribution of θ:

$$P(\theta\,|\,x_1,\ldots,x_N)=\frac{p(x_1,\ldots,x_N\,|\,\theta)\,p(\theta)}{p(x_1,\ldots,x_N)}\propto\theta^{n_h}(1-\theta)^{n_t}\times\theta^{\alpha-1}(1-\theta)^{\beta-1}=\theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1}$$

  - Notice the isomorphism of the posterior to the prior,
  - such a prior is called a **conjugate prior**

Bayesian estimation for Bernoulli, con'd

- Posterior distribution of θ:

$$P(\theta \mid x_1,\ldots,x_N) = \frac{p(x_1,\ldots,x_N \mid \theta)\, p(\theta)}{p(x_1,\ldots,x_N)} \propto \theta^{n_h}(1-\theta)^{n_t} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1}$$

- Maximum *a posteriori* (MAP) estimation:

$$\theta_{MAP} = \arg\max_{\theta} \log P(\theta \mid x_1,\ldots,x_N)$$

- Posterior mean estimation:

Beta parameters can be understood as pseudo-counts

$$\theta_{Bayes} = \int \theta p(\theta \mid D)\, d\theta = C\int \theta \times \theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1}\, d\theta = \frac{n_h+\alpha}{N+\alpha+\beta}$$

- Prior strength: A=α+β
  - A can be interoperated as the size of an imaginary data set from which we obtain the **pseudo-counts**

Effect of Prior Strength

- Suppose we have a uniform prior ($\alpha=\beta=1/2$), and we observe $\vec{n}=(n_h=2, n_t=8)$
- Weak prior A = 2. Posterior prediction:

$$p(x=h \mid n_h=2, n_t=8, \vec{\alpha}=\vec{\alpha}'\times 2)=\frac{1+2}{2+10}=0.25$$

- Strong prior A = 20. Posterior prediction:

$$p(x=h \mid n_h=2, n_t=8, \vec{\alpha}=\vec{\alpha}'\times 20)=\frac{10+2}{20+10}=0.40$$

- However, if we have enough data, it washes away the prior. e.g., $\vec{n}=(n_h=200, n_t=800)$. Then the estimates under weak and strong prior are $\frac{1+200}{2+1000}$ and $\frac{10+200}{20+1000}$, respectively, both of which are close to 0.2

Bayesian estimation for normal distribution

- Normal Prior:

$$P(\mu) = \left(2\pi\tau^2\right)^{-1/2} \exp\left\{-(\mu - \mu_0)^2 / 2\tau^2\right\}$$

- Joint probability:

$$P(x,\mu) = \left(2\pi\sigma^2\right)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}$$
$$\times \left(2\pi\tau^2\right)^{-1/2} \exp\left\{-(\mu - \mu_0)^2 / 2\tau^2\right\}$$

- Posterior:

$$P(\mu \mid x) = \left(2\pi\tilde{\sigma}^2\right)^{-1/2} \exp\left\{-(\mu - \tilde{\mu})^2 / 2\tilde{\sigma}^2\right\}$$

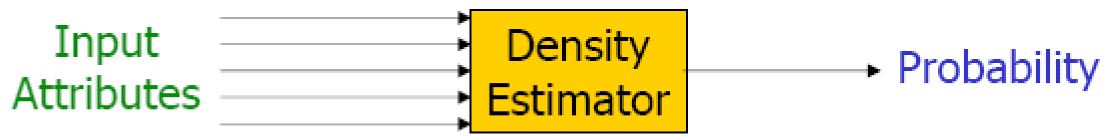
where $\tilde{\mu} = \frac{N/\sigma^2}{N/\sigma^2 + 1/\tau^2}\bar{x} + \frac{1/\tau^2}{N/\sigma^2 + 1/\tau^2}\mu_0$, and $\tilde{\sigma}^2 = \left(\frac{N}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}$

Sample mean

- AFTER THIS POINT ARE OLD SLIDES

Density Estimation

- A Density Estimator learns a mapping from a set of attributes to a Probability



- Often know as parameter estimation if the distribution form is specified
  - Binomial, Gaussian …

- Three important issues:

  - Nature of the data (iid, correlated, …)
  - Objective function (MLE, MAP, …)
  - Algorithm (simple algebra, gradient methods, EM, …)
  - Evaluation scheme (likelihood on test data, predictability, consistency, …)

Parameter Learning from iid data

- Goal: estimate distribution parameters θ from a dataset of N independent, identically distributed (*iid*), fully observed, training cases

$$D = \{x_1, \ldots, x_N\}$$

- Maximum likelihood estimation (MLE)
  1. One of the most common estimators
  2. With iid and full-observability assumption, write $L(\theta)$ as the likelihood of the data:

$$\begin{aligned} L(\theta) &= P(x_1, x_2, \ldots, x_N; \theta) \\ &= P(x;\theta)P(x_2;\theta), \ldots, P(x_N;\theta) \\ &= \prod_{i=1}^{N} P(x_i;\theta) \end{aligned}$$

  3. pick the setting of parameters most likely to have generated the data we saw:

$$\theta^* = \arg\max_\theta L(\theta) \; = \arg\max_\theta \log L(\theta)$$