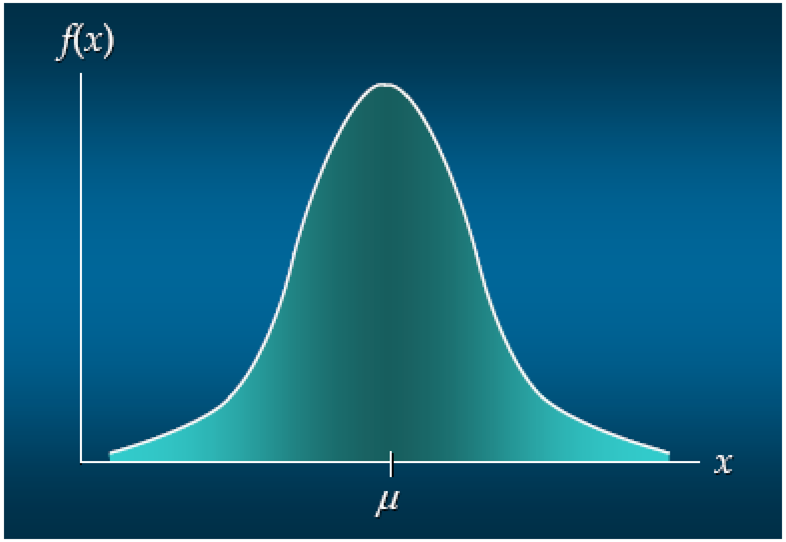
Machine Learning

Probability



Reading: Bishop: Chap 1,2

Probability in Machine Learning

- **Machine Learning tasks involve reasoning under uncertainty**

Sources of uncertainty/randomness:

- Noise – variability in sensor measurements, partial observability, incorrect labels
- Finite sample size - Training and test data are randomly drawn instances

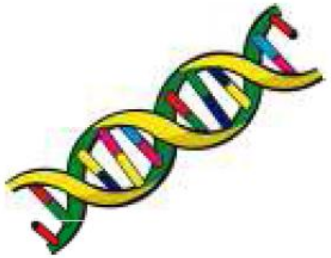
Hand-written digit recognition

Probability quantifies uncertainty!

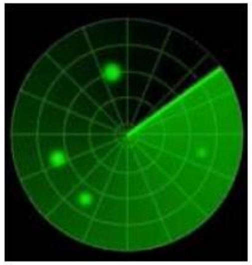
Basic Probability Concepts

- **Conceptual or physical, repeatable experiment with random outcome at any trial**

Roll of dice



Nucleotide present at a DNA site



Time-space position of an aircraft on a radar screen

- *Sample space S* - set of all possible outcomes. (can be finite or infinite.)

$S \equiv \{1,2,3,4,5,6\}$ $\quad S \equiv \{A,T,C,G\}$ $\quad S \equiv \{0, R_{\max}\} \times \{0, 360^{\circ}\} \times \{0, +\infty\}$

- *Event A* - any subset of *S* :

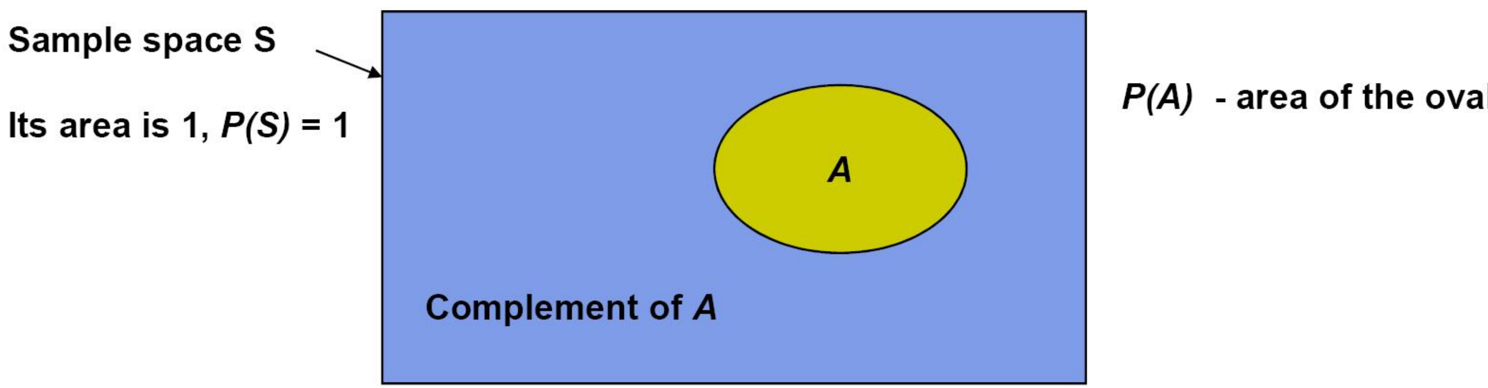
See “2“,”4” or “6” in a roll observe a "G" at a site UA007 in angular location {45°-60°}

Definition

- ***Classical***: Probability of an event A is the relative frequency (limiting ratio of number of occurrences of event A to the total number of trials)

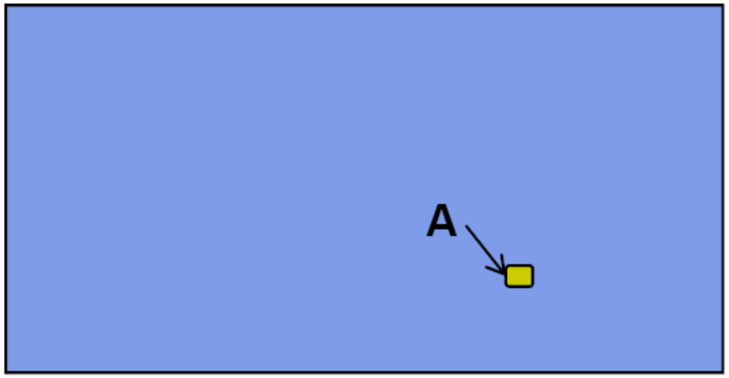
$$P(A) = \lim_{N \to \infty} \frac{N_A}{N}$$

E.g. $P(\{1\}) = 1/6$ $P(\{2,4,6\}) = 1/2$

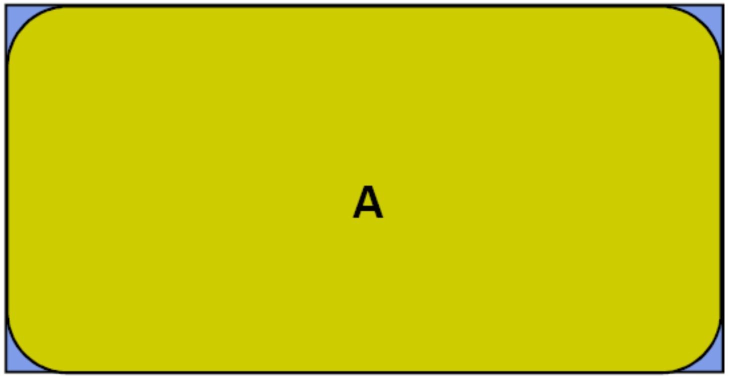


Definition

- *Axiomatic (Kolmogorov):* **Probability of an event A is a number assigned to this event such that**
- $0 \le P(A) \le 1$ all probabilities are between 0 and 1



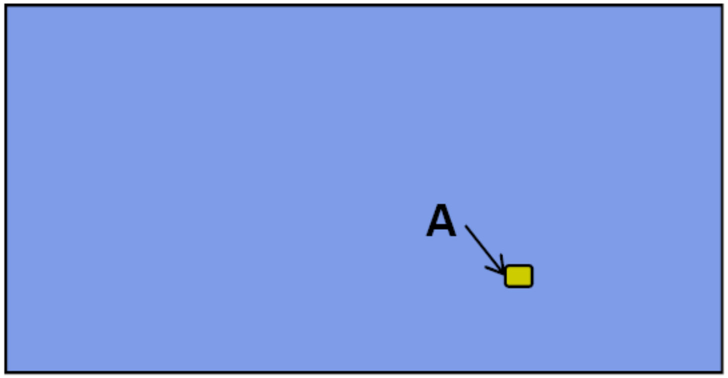
Area of A can't be smaller than 0



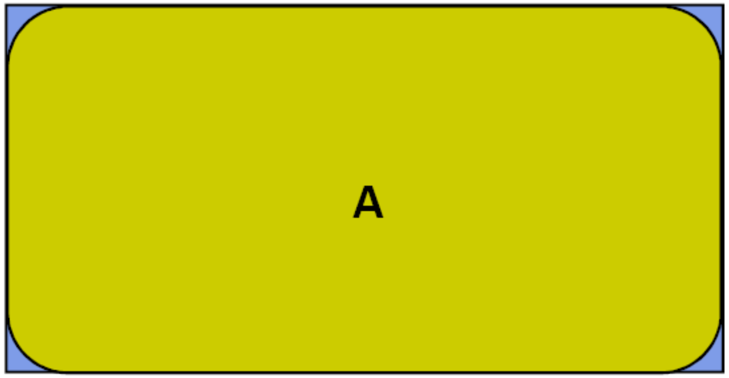
Area of A can't be larger than 1

Definition

- ***Axiomatic (Kolmogorov):*** **Probability of an event A is a number assigned to this event such that**
- $0 \leq P(A) \leq 1$ all probabilities are between 0 and 1
- $P(\phi) = 0$ probability of no outcome is 0



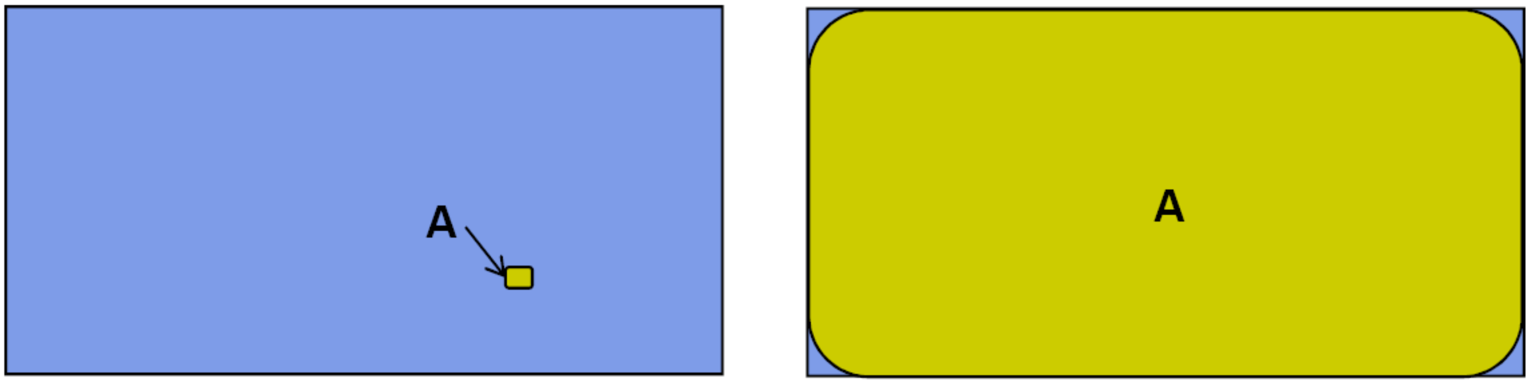
Area of A can’t be smaller than 0



Area of A can’t be larger than 1

Definition

- ***Axiomatic (Kolmogorov):*** **Probability of an event A is a number assigned to this event such that**
- $0 \le P(A) \le 1$ all probabilities are between 0 and 1
- $P(\phi) = 0$ probability of no outcome is 0
- $P(S) = 1$ probability of some outcome is 1



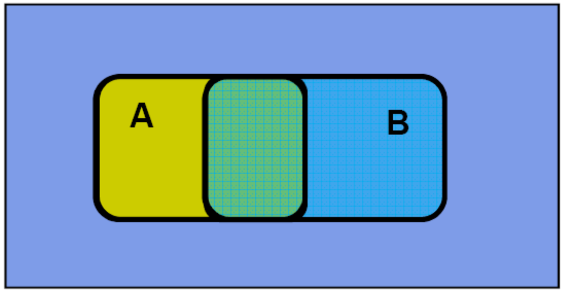
Area of A can't be smaller than 0

Area of A can't be larger than 1

Definition

- ***Axiomatic (Kolmogorov):*** **Probability of an event A is a number assigned to this event such that**
- $0 \le P(A) \le 1$ all probabilities are between 0 and 1
- $P(\phi) = 0$ probability of no outcome is 0
- $P(S) = 1$ probability of some outcome is 1
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

  probability of union of two events



Area of A U B = Area of A + Area of B – Area of A ∩ B

Definition

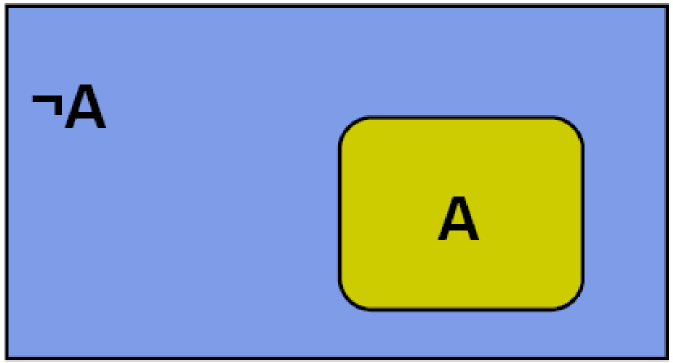
- ***Axiomatic (Kolmogorov):*** **Probability of an event A is a number assigned to this event such that**
- $0 \leq P(A) \leq 1$ all probabilities are between 0 and 1
- $P(\phi) = 0$ probability of no outcome is 0
- $P(S) = 1$ probability of some outcome is 1
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

  probability of union of two events

- ***Probability space*** is a sample space equipped with an assignment P(A) to every event $A \subset S$ such that P satisfies the Kolmogorov axioms.

Theorems from the Axioms

- $0 \le P(A) \le 1$
- $P(\phi) = 0$
- $P(S) = 1$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



$P(\neg A) = 1 - P(A)$

Proof: $P(A \cup \neg A) = P(S) = 1$

$P(A \cap \neg A) = P(\phi) = 0$

$1 = P(A) + P(\neg A) - 0 \quad \Rightarrow P(\neg A) = 1 - P(A)$

Theorems from the Axioms

- $0 \le P(A) \le 1$
- $P(\phi) = 0$
- $P(S) = 1$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

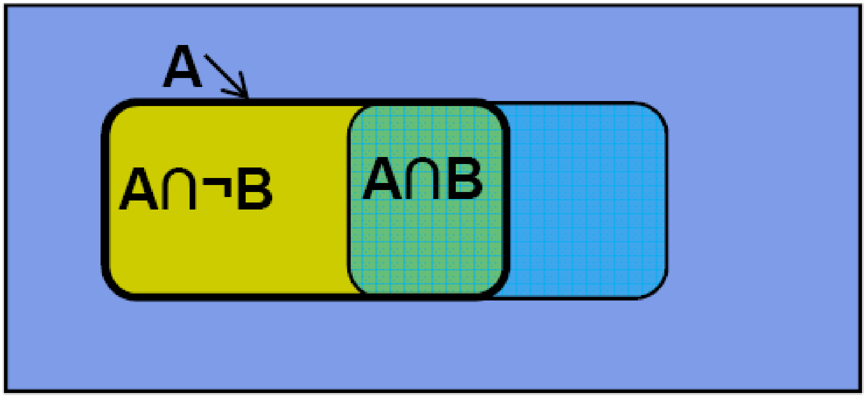
$$P(A) = P(A \cap B) + P(A \cap \neg B)$$

Proof: $P(A) = P(A \cap S) = P(A \cap (B \cup \neg B)) = P((A \cap B) \cup (A \cap \neg B))$

$= P(A \cap B) + P(A \cap \neg B) - P((A \cap B) \cap (A \cap \neg B))$

$= P(A \cap B) + P(A \cap \neg B) - P(\phi)$

$= P(A \cap B) + P(A \cap \neg B)$



Why use probability?

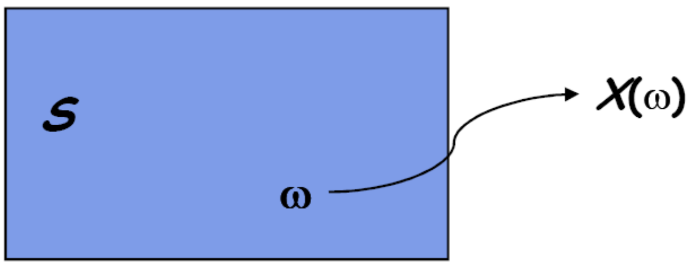
- **There have been many other approaches to handle uncertainty:**
  - **Fuzzy logic**
  - **Qualitative reasoning (Qualitative physics)**
- **“Probability theory is nothing but common sense reduced to calculation”**
  - **— Pierre Laplace, 1812.**

- **Any scheme for combining uncertain information really should obey these axioms**
  - **Di Finetti 1931 - If you gamble based on “uncertain beliefs” that satisfy these axioms, then you can’t be exploited by an opponent**

Random Variable

- A *random variable* is a function that associates a unique numerical value $X(\omega)$ with every outcome $\omega \in S$ of an experiment.

(The value of the r.v. will vary from trial to trial as the experiment is repeated)



$$P(X < 2) = P(\{\omega : X(\omega) < 2\})$$

- ***Discrete r.v.:***
  - The outcome of a coin-toss H = 1, T = 0 (Binary)
  - The outcome of a dice-roll 1-6
- ***Continuous r.v.:***
  - The location of an aircraft
- ***Univariate r.v.:***
  - The outcome of a dice-roll 1-6
- ***Multi-variate r.v.:***
  - The time-space position of an aircraft on radar screen
  - $X = \begin{pmatrix} R \\ \Theta \\ t \end{pmatrix}$

Discrete Probability Distribution

- In the discrete case, a probability distribution *P* on *S* (and hence on the domain of *X*) is an assignment of a non-negative real number P(s) to each $s \in S$ (or each valid value of *x*) such that

  $0 \leq P(X=x) \leq 1$ X – random variable

  $\Sigma_x P(X = x) = 1$ x – value it takes

- E.g. Bernoulli distribution with parameter θ

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

Discrete Probability Distribution

- **In the discrete case, a probability distribution *P* on *S* (and hence on the domain of *X*) is an assignment of a non-negative real number P(s) to each $s \in S$ (or each valid value of *x*) such that**

$0 \le P(X=x) \le 1$ X – random variable

$\Sigma_x P(X = x) = 1$ x – value it takes

E.g. Multinomial distribution with parameters $\theta_1, \ldots, \theta_K$

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

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

Continuous Prob. Distribution

- **A continuous random variable *X* can assume any value in an interval on the real line or in a region in a high dimensional space**
  - **X usually corresponds to a real-valued measurements of some property, e.g., length, position, …**
  - **It is not possible to talk about the probability of the random variable assuming a particular value --- P(X=x) = 0**
  - **Instead, we talk about the probability of the random variable assuming a value within a given interval, or half interval**

$$P(X \in [x1, x2])$$

$$P(X < x) = P(X \in [-\infty, x])$$

Continuous Prob. Distribution

- The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be 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 $\int_{-\infty}^{+\infty} p(x)dx = 1$.

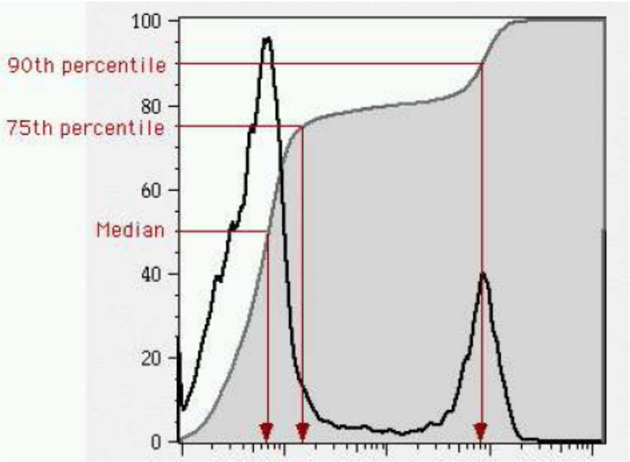
  - Cumulative distribution function (CDF):

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} p(x')dx'$$

  - Probability density function (PDF):

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

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



Car flow on Liberty Bridge (cooked up!)

What is the intuitive meaning of *p(x)*

- **If**

$$p(x_1) = a \text{ and } p(x_2) = b,$$

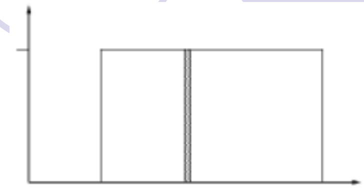
then when a value X is sampled from the distribution with density p(x), you are a/b times as likely to find that X is "very close to" x than that x_1 is "very close to" x_2.

- **That is:**

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

Continuous Distributions

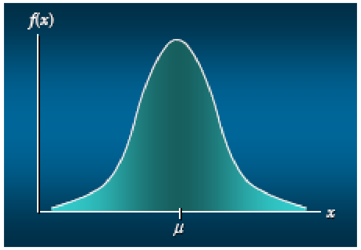
- **Uniform Probability Density Function**



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

- **Normal (Gaussian) Probability Density Function**

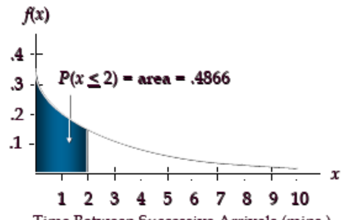
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$



  - **The distribution is symmetric, and is often illustrated as a bell-shaped curve.**
  - **Two parameters, μ (mean) and σ (standard deviation), determine the location and shape of the distribution.**

- **Exponential Probability Distribution**

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



Statistical Characterizations

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

$$E(X) = \begin{cases} \sum_x xp(x) & \text{discrete} \\ \int_{-\infty}^{\infty} xp(x)dx & \text{continuous} \end{cases}$$

- **Variance:** the spread

$$\text{Var}(X) = \begin{cases} \sum_x (x - E(X))^2 p(x) & \text{discrete} \\ \int_{-\infty}^{\infty} (x - E(X))^2 p(x)dx & \text{continuous} \end{cases}$$

Gaussian (Normal) density in 1D

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

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

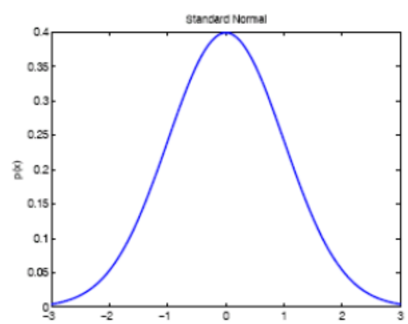
$$E(X) = \mu$$

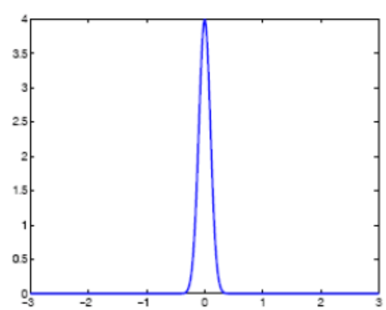
$$\mathrm{var}(X) = \sigma^2$$

  - Here is how we plot the pdf in matlab

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

Zero mean
Large variance





Zero mean
Small variance

Note that a density evaluated at a point can be bigger than 1!

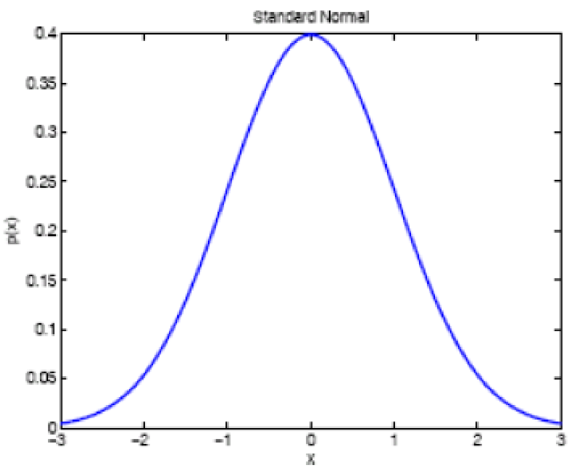
Gaussian CDF

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

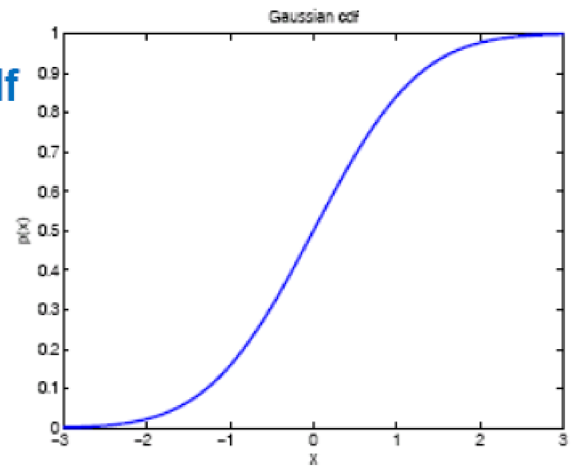
$$\Phi(x) = \int_{-\infty}^{x} p(z)dz$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2} dz$$

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

pdf



cdf

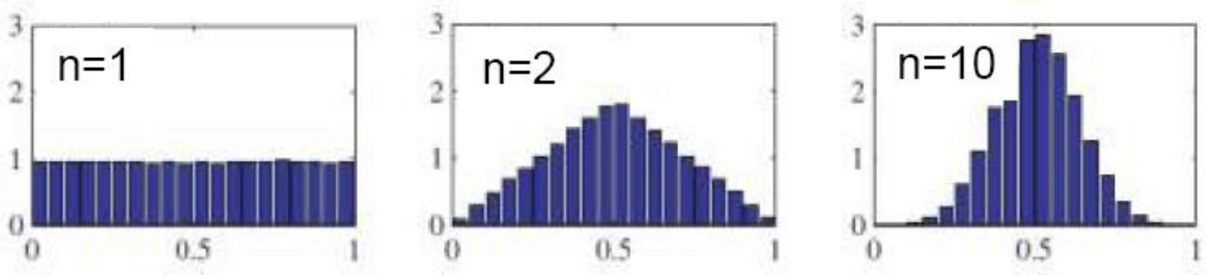


Central limit theorem

- If $(X_1, X_2, \ldots, X_n)$ are i.i.d. (independent and identically distributed – to be covered next) random variables
- Then define

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- As n→infinity,
- $p(\overline{X}) \rightarrow$ Gaussian with mean $E[X_i]$ and variance $Var[X_i]/n$



- **Somewhat of a justification for assuming Gaussian distribution**

Independence

- **Training and test samples typically assumed to be i.i.d. (independent and identically distributed)**

- **A and B are independent events if**

$$P(A \cap B) = P(A) * P(B)$$

- **Outcome of A has no effect on the outcome of B (and vice versa).**

E.g. Roll of two die

$P(\{1\},\{3\}) = 1/6*1/6 = 1/36$

Independence

- **A, B and C are pairwise independent events if**

$$P(A \cap B) = P(A) * P(B)$$

$$P(A \cap C) = P(A) * P(C)$$

$$P(B \cap C) = P(B) * P(C)$$

- **A, B and C are mutually independent events if, in addition to pairwise independence,**

$$P(A \cap B \cap C) = P(A) * P(B) * P(C)$$

Conditional Probability

- **P(A|B) = Probability of event A conditioned on event B having occurred**

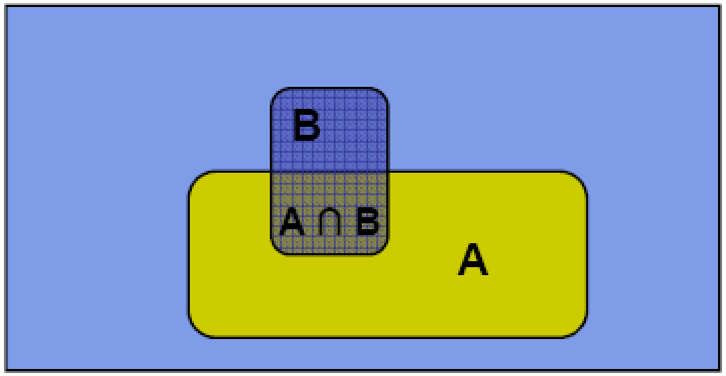
$$\textbf{If } P(B) > 0\textbf{, then } P(A|B) = \frac{P(A \cap B)}{P(B)}$$

E.g. H = "having a headache"

F = "coming down with Flu“

- P(H)=1/10
- P(F)=1/40
- P(H|F)=1/2

Fraction of people with flu that have a headache



- Corollary: The Chain Rule
- P(A ∩ B) = P(A|B) P(B)

If A and B are independent, P(A|B) = P(A)

Conditional Independence

- **A and B are independent if**

$$P(A \cap B) = P(A) * P(B) \quad \equiv \quad P(A|B) = P(A)$$

- **Outcome of B has no effect on the outcome of A (and vice versa).**

- **A and B are conditionally independent given C if**

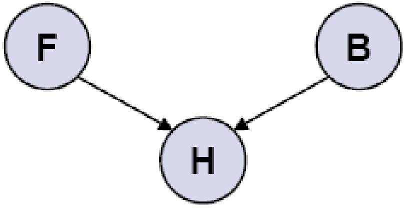
$$P(A \cap B|C) = P(A|C) * P(B|C) \quad \equiv \quad P(A|B,C) = P(A|C)$$

- **Outcome of B has no effect on the outcome of A (and vice versa) if C is true.**

Prior and Posterior Distribution

- **Suppose that our propositions have a "causal flow"**

e.g.,



- **Prior or unconditional probabilities of propositions**

e.g., P(Flu) = 0.025 and P(DrinkBeer) = 0.2

correspond to belief prior to arrival of any (new) evidence

- **Posterior or conditional probabilities of propositions**

e.g., P(Headache|Flu) = 0.5 and P(Headache|Flu,DrinkBeer) = 0.7

correspond to updated belief after arrival of new evidence

- **Not always useful: P(Headache|Flu, Steelers win) = 0.5**

Probabilistic Inference

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 with the following reasoning: "since 50% of flues are associated with headaches, so I must have a 50-50 chance of coming down with flu"**

Is this reasoning correct?

Probabilistic Inference

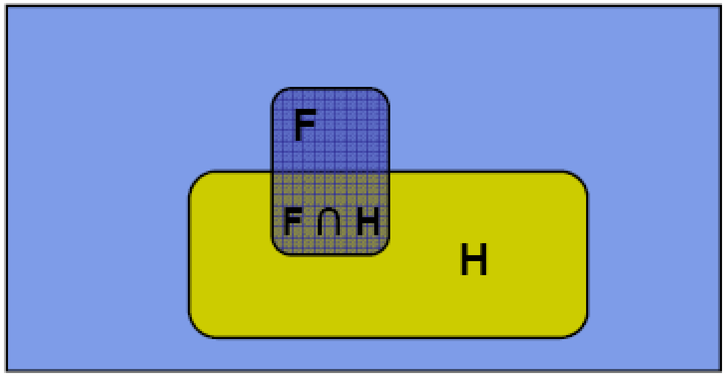
H = "having a headache"

F = "coming down with Flu"

- P(H)=1/10
- P(F)=1/40
- P(H|F)=1/2

- **The Problem:**

  $P(F|H) = ?$



Probabilistic Inference

H = "having a headache"
F = "coming down with Flu"

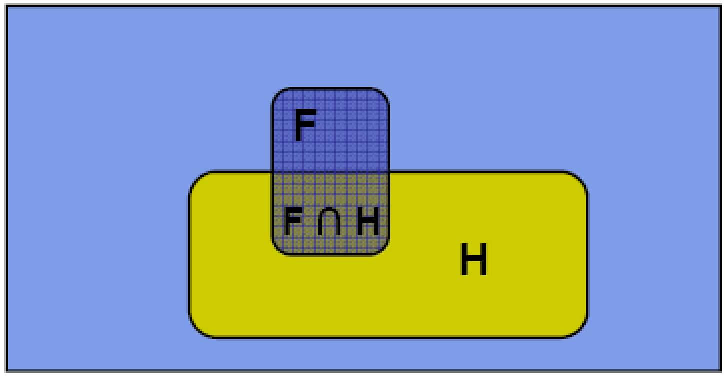
- **P(H)=1/10**
- **P(F)=1/40**
- **P(H|F)=1/2**

- **The Problem:**

$$P(F|H) = \frac{P(F \cap H)}{P(H)}$$

$$= \frac{P(H|F)P(F)}{P(H)}$$

$$= 1/8 \neq P(H|F)$$



The Bayes Rule

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

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

This is Bayes Rule

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**

Quiz

P(H)=1/10

P(F)=1/40

P(H|F)=1/2

P(F|H) = 1/8

- Which of the following statement is true?

$P(F|\neg H) = 1 - P(F|H)$ ✗

$P(\neg F|H) = 1 - P(F|H)$ ✓

$$P(F|\neg H) = \frac{P(\neg H|F)\,P(F)}{P(\neg H)} = \frac{(1 - P(H|F))\,P(F)}{1 - P(H)}$$

More General Forms of Bayes Rule

- $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- Law of total probability

$$P(B) = P(B \cap A) + P(B \cap \neg A)$$

$$= P(B|A)\,P(A) + P(B|\neg A)\,P(\neg A)$$

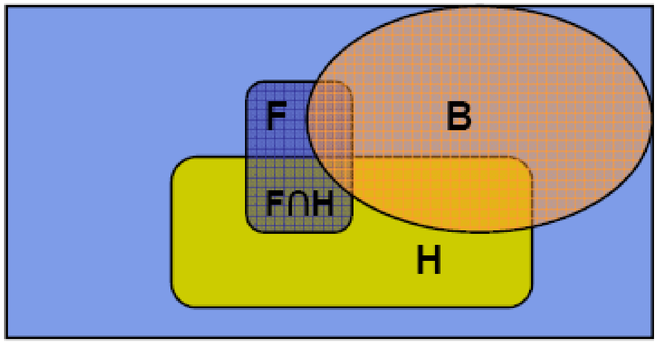
- $$P(A \mid B) = \frac{P(B|A)P(A)}{P(B|\,A)P(A) + P(B|\neg A)P(\neg A)}$$

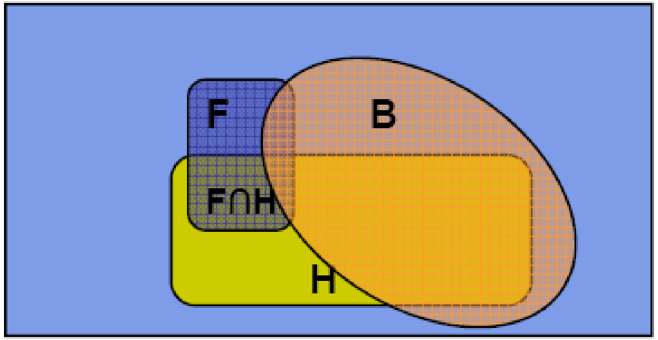
More General Forms of Bayes Rule

- $P(Y = y|X) = \frac{P(X|Y=y)P(Y=y)}{\sum_y P(X|Y = y)P(Y=y)}$

$$P(Y|X \wedge Z) = \frac{P(X \mid Y \wedge Z)p(Y \wedge Z)}{P(X \wedge Z)} = \frac{P(X \mid Y \wedge Z)p(Y \wedge Z)}{P(X \mid \neg Y \wedge Z)p(\neg Y \wedge Z) + P(X \mid Y \wedge Z)p(\neg Y \wedge Z)}$$

E.g. P(Flu | Headhead ∧ DrankBeer)





Joint and Marginal Probabilities

- **A *joint probability distribution* for a set of RVs (say X_1,X_2,X_3) gives the probability of every atomic event $P(X_1,X_2,X_3)$**
  - P(Flu,DrinkBeer) = a 2 × 2 matrix of values:

	B	¬B
F	0.005	0.02
¬F	0.195	0.78

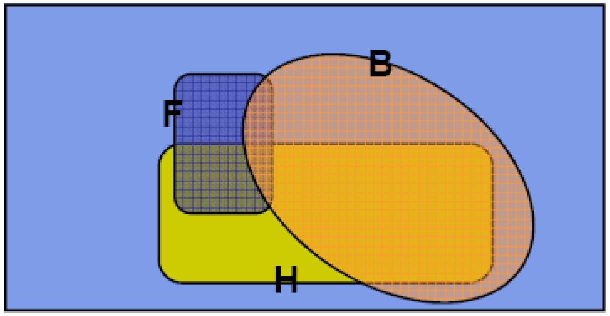
  - P(Flu,DrinkBeer, Headache) = ?
  - Every question about a domain can be answered by the joint distribution, as we will see later.
- A *marginal probability distribution* is the probability of every value that a single RV can take $P(X_1)$

P(Flu) = ?

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



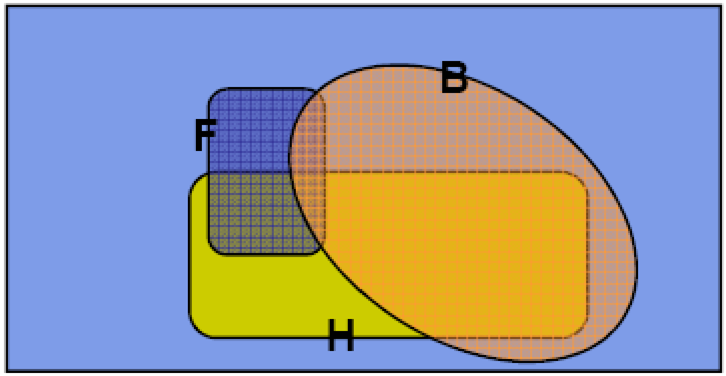
Inference with the Joint

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

¬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

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

E.g. E = {(¬F, ¬B,H),(¬F,B,H)}



Inference with the Joint

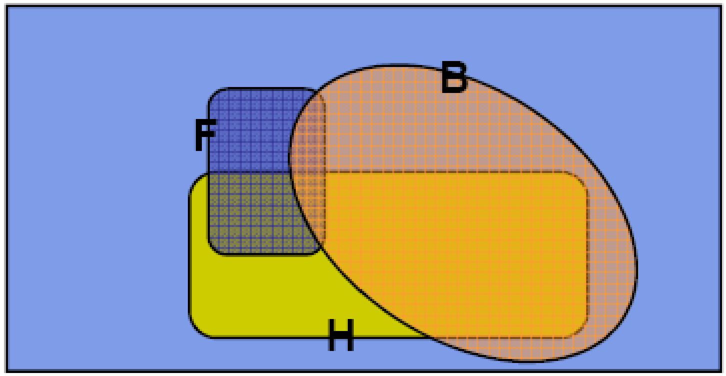
- Compute Marginals

P(Flu ∧ Headache)

$=P(F \wedge H \wedge B) + P(F \wedge H \wedge \neg B)$

Recall: Law of Total Probability

¬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

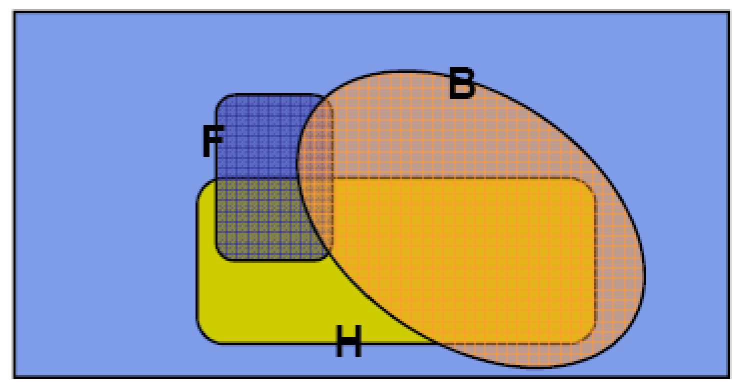
P(Headache)

$= P(H \wedge F) + P(H \wedge \neg F)$

$= P(H \wedge F \wedge B) + P(H \wedge F \wedge \neg B)$

$+ P(H \wedge \neg F \wedge B) + P(H \wedge \neg F \wedge \neg B)$

¬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



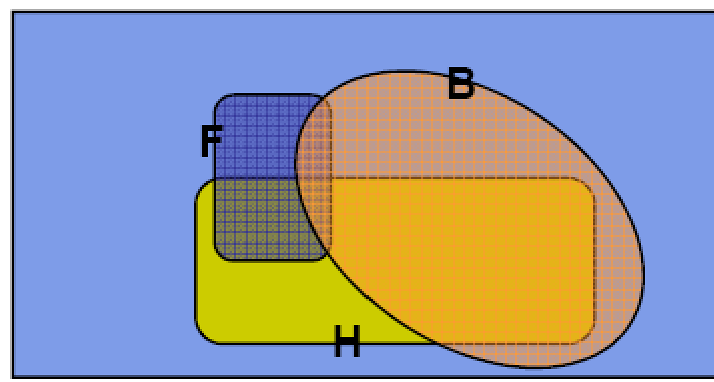
Inference with the Joint

- 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

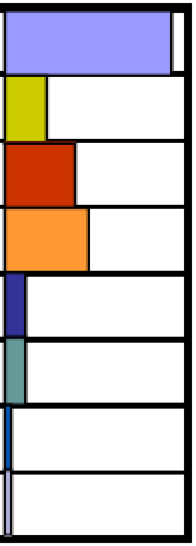


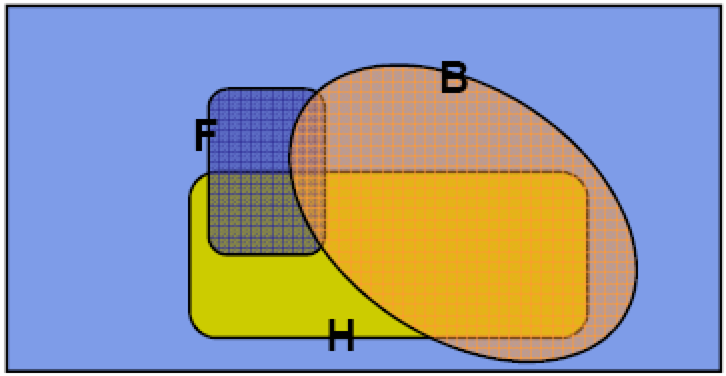
Inference with the Joint

- Compute Conditionals
- $P(Flu|Headache) = \frac{P(Flu \wedge Headache)}{P(Headache)} = ?$

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

¬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





Where do probability distributions come from?

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

e.g., $P(F)$

$P(B)$

$P(H|\neg F,B)$

$P(H|F,\neg B)$

¬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	

- **Use chain rule and independence assumptions to compute joint distribution**

Where do probability distributions come from?

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

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…

- Some 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 independent, identically distributed (iid), fully observed, training cases**

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

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

$$L(\theta)= P(D; \theta)=P(x_1\ x_2\ ,\ ,\ x_N; \theta\ )$$

$$= P(X_1;\theta)\ P(X_2;\theta)\ldots P(X_N;\theta)$$

$$= \prod_i^N P(X_i;\theta)$$

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

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \boldsymbol{L}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \boldsymbol{log}(\boldsymbol{L}(\boldsymbol{\theta}))$$

Example 1: Bernoulli model

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

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

- **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 = $\{x_1, \ldots, x_N\}$:**

$$L(\theta) = P(x_1, x_2, \ldots, x_N; \theta) = \prod_{i=1}^{N} P(x_i; \theta) = \prod_{i=1}^{N} \left(\theta^{x_i}(1-\theta)^{1-x_i}\right)$$

$$= \theta^{\sum_{i=1}^{N} x_i} (1-\theta)^{\sum_{i=1}^{N} 1-x_i} = \theta^{\#\text{head}}(1-\theta)^{\#\text{tails}}$$

MLE

- **Objective function:**

$$\ell(\theta) = \log L(\theta) = \log \theta^{n_h}(1-\theta)^{n_t} = n_h \log\theta + (N - n_h)\log(1-\theta)$$

- **We need to maximize this w.r.t. *θ***
- **Take derivatives w.r.t *θ***

$$\frac{\partial \ell}{\partial \theta} = \frac{n_h}{\theta} - \frac{N - n_h}{1-\theta} = 0 \quad \Rightarrow \quad \widehat{\theta}_{MLE} = \frac{n_h}{N} \quad \text{or} \quad \widehat{\theta}_{MLE} = \frac{1}{N}\sum_i x_i$$

Frequency as sample mean

- **Sufficient statistics**

The counts, n_h, where $n_h = \sum_i x_i$, are **sufficient statistics** of data D

Example 2: univariate normal

- Data:
  - We observed *N* ***iid*** real samples:

    D={-0.1, 10, 1, -5.2, …, 3}
- Model: $P(x) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-(x-\mu)^2/2\sigma^2\right\}$ $\theta = (\mu, \sigma^2)$
- Log likelihood:

$$\ell(\theta) = \log L(\theta) = \prod_{i=1}^{N} P(x_i) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^2}$$

- MLE: take derivative and set to zero:

$$\frac{\partial \ell}{\partial \mu} = (1/\sigma^2)\sum_n (x_n - \mu)$$

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

$$\Rightarrow$$

$$\mu_{MLE} = \frac{1}{N}\sum_n x_n$$

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

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 "smoothing":**
  - Where n' is know as the pseudo- (imaginary) count

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

  - But can we make this more formal?

Bayesian Learning

- **The Bayesian Rule:**

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

- **Or equivalently,**

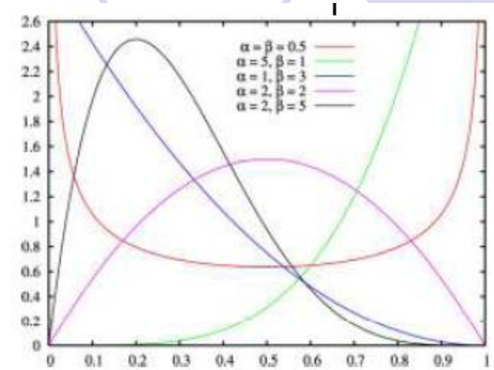
$$\underbrace{P(\theta \mid \mathcal{D})}_{\text{posterior}} \propto \underbrace{P(\mathcal{D} \mid \theta)}_{\text{likelihood}} \underbrace{P(\theta)}_{\text{prior}}$$

(Belief about coin toss probability)

- MAP estimate: $\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D)$
- If prior is uniform, MLE = MAP

Bayesian estimation for Bernoulli

- **Beta(α,β) distribution:**



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

- **Posterior distribution of *θ* :**

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

Beta(α+n_h,β+n_t)

  - Notice the **isomorphism** of the **posterior to the prior**,
  - such a prior is called a ***conjugate prior***
  - ***α*** and ***β*** are ***hyperparameters*** (parameters of the prior) and correspond to the number of “virtual” heads/tails (pseudo counts)

MAP

- 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:

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

- Posterior mean estimation:

$$\hat{\theta}_{\mathrm{MAP}} = \frac{n_h + \alpha}{N + \alpha + \beta}$$

Beta parameters
can be understood
as pseudo-counts

- With enough data, prior is forgotten

Dirichlet distribution

- number of heads in N flips of a two-sided coin
  - follows a binomial distribution
  - Beta is a good prior (conjugate prior for binomial)
- what it's not two-sided, but k-sided?
  - follows a multinomial distribution
  - Dirichlet distribution is the conjugate prior

$$P(\theta_1, \theta_2, \dots \theta_K) = \frac{1}{B(\alpha)} \prod_i^K \theta_i^{(\alpha_1 - 1)}$$

Lejeune Dirichlet

Johann Peter Gustav Lejeune Dirichlet

Born	13 February 1805 Düren, French Empire
Died	5 May 1859 (aged 54) Göttingen, Hanover
Residence	Germany
Nationality	German
Fields	Mathematician
Institutions	University of Berlin University of Breslau University of Göttingen
Alma mater	University of Bonn
Doctoral advisor	Simeon Poisson Joseph Fourier
Doctoral students	Ferdinand Eisenstein Leopold Kronecker Rudolf Lipschitz Carl Wilhelm Borchardt
Known for	Dirichlet function Dirichlet eta function

Estimating the parameters of a distribution

- Maximum Likelihood estimation (MLE) Choose value that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

- Maximum a posteriori (MAP) estimation Choose value that is most probable given observed data and prior belief

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D) = \arg\max_{\theta} P(D \mid \theta) P(\theta)$$

MLE vs MAP (Frequentist vs Bayesian)

- Frequentist/MLE approach:
  - θ is unknown constant, estimate from data
- Bayesian/MAP approach:
  - θ is a random variable, assume a probability distribution
- Drawbacks
  - MLE: Overfits if dataset is too small
  - MAP: Two people with different priors will end up with different estimates

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

Probability Review

What you should know:

- Probability basics
  - random variables, events, sample space, conditional probs, ...
  - independence of random variables
  - Bayes rule
  - Joint probability distributions
  - calculating probabilities from the joint distribution
- Point estimation
  - maximum likelihood estimates
  - maximum a posteriori estimates
  - distributions – binomial, Beta, Dirichlet, ...