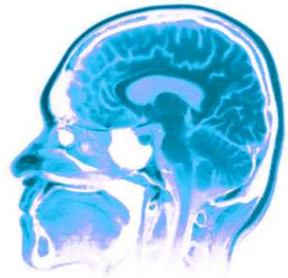
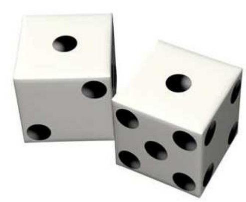
CPSC540



Discrete Probability and Bayesian Learning

Nando de Freitas
September, 2011
University of British Columbia

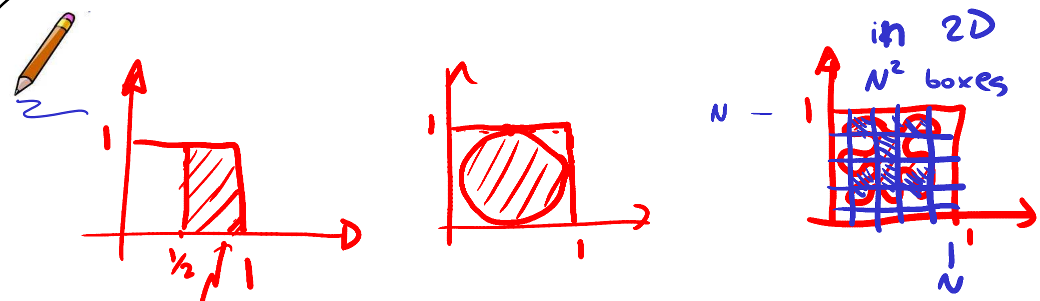
Probability



Probability theory is the formal study of the laws of chance. It is our tool for dealing with uncertainty. Notation:

- **Sample space:** is the set Ω of all outcomes of an experiment.
- **Outcome:** what we observed. We use $\omega \in \Omega$ to denote a particular outcome. *e.g.* for a die we have $\Omega = \{1, 2, 3, 4, 5, 6\}$ and ω could be any of these six numbers.
- **Event:** is a subset of Ω that is well defined (measurable). *e.g.* the event $A = \{even\}$ if $w \in \{2, 4, 6\}$

Frequentist interpretation



$P(\blacksquare) = \frac{1}{2}$

$P(\boxed{/}) = \frac{\pi (1/2)^2}{1} = \frac{\pi}{4}$

$P(\boxed{/}) = \frac{\# \text{ squares shaded}}{\# \text{ squares}}$

make squares smaller for more precision!

Axiomatic interpretation

The axiomatic view is a more elegant mathematical solution. Here, a **probabilistic model** consists of the triple $(\Omega, \mathcal{F}, P)$, where Ω is the sample space, $\mathcal{F}$ is the sigma-field (collection of measurable events) and P is a function mapping $\mathcal{F}$ to the interval $[0, 1]$. That is, with each event $A \in \mathcal{F}$ we associate a probability $P(A)$.

$$\begin{cases} \Omega = \{1, 2, 3, 4, 5, 6\} \\ \mathcal{F} = \text{Powerset} = \{\emptyset, 1, 2 \ldots, 6, \{1,2\}, \ldots\} \\ P(\text{even}) = 1/2 \\ P(\text{odd}) = 1/2 \end{cases}$$

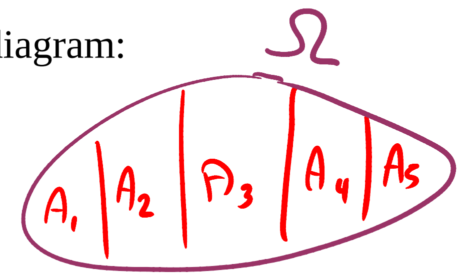
The axioms

1. $P(\emptyset) = 0 \leq p(A) \leq 1 = P(\Omega)$

2. For **disjoint sets** A_n, $n \geq 1$, we have

$$P\left(\sum_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

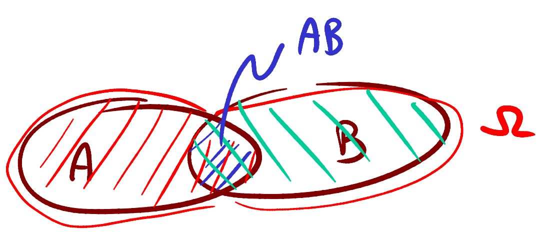
Venn diagram:

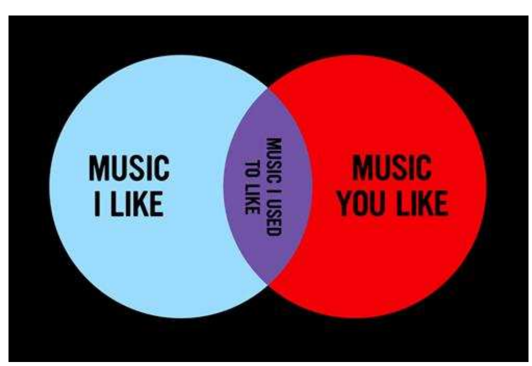


$P(\Omega) = P(A_1) + P(A_2) + \cdots + P(A_5)$

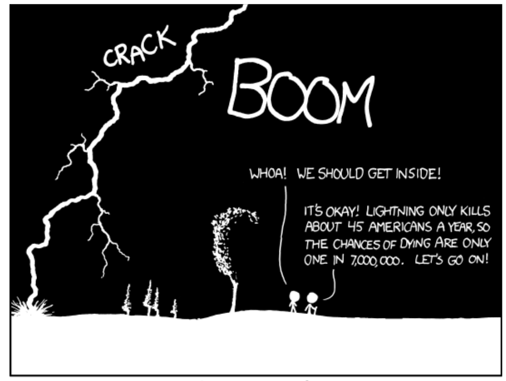
OR and AND operations

$$P(A \overset{\text{OR}}{+} B) = P(A) + P(B) - P(\overset{\text{and}}{AB})$$





Conditional probability



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

given, and

$$P(A|B) \triangleq \frac{P(AB)}{P(B)}$$

where $P(A|B)$ is the **conditional probability** of A given that B occurs, $P(B)$ is the **marginal probability** of B and $P(AB)$ is the **joint probability** of A and B. In general, we obtain a **chain rule**

$$P(A_{1:n}) = P(A_n|A_{1:n-1})P(A_{n-1}|A_{1:n-2}) \ldots P(A_2|A_1)P(A_1)$$

If the events A and B are **independent**, we have $P(AB) = P(A)P(B)$. $= P(A|\cancel{B})P(B)$ $P(A)$

Conditional probability example

⋆ Assume we have an urn with 3 red balls and 1 blue ball: $U = \{r, r, r, b\}$. What is the probability of drawing (without replacement) 2 red balls in the first 2 tries?

$$P(d_1 = r) = \frac{3}{4}$$

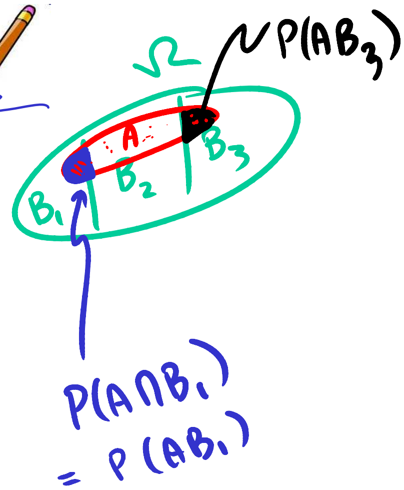
$$P(d_2 = r, d_1 = r) = P(d_2 = r | d_1 = r)\, P(d_1 = r)$$

$$= \frac{2}{3}\left(\frac{3}{4}\right) = \frac{1}{2}$$

Marginalization

Let the sets $B_{1:n}$ be disjoint and $\bigcup_{i=1}^{n} B_i = \Omega$. Then

$$\underline{P(A)} = \sum_{i=1}^{n} P(A, B_i)$$



$$P(A) = P(A \cap \Omega)$$

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3)$$

$$P(A) = P(AB_1) + P(AB_2) + P(AB_3)$$

Marginalization example

⋆ What is the probability that the second ball drawn from our urn will be red?

$$P(d_2 = r) = \sum_{d_1 \in \{b, r\}} P(d_2 = r, d_1)$$

$$= \sum_{d_1} P(d_2 = r \mid d_1) P(d_1)$$

$$= P(d_2 = r \mid d_1 = r) P(d_1 = r) + P(d_2 = r \mid d_1 = b) P(d_1 = b)$$

Bayes rule

Bayes rule allows us to reverse probabilities:

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

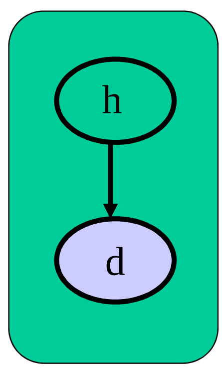
$$P(AB) = P(B|A)P(A) = P(A|B)P(B)$$

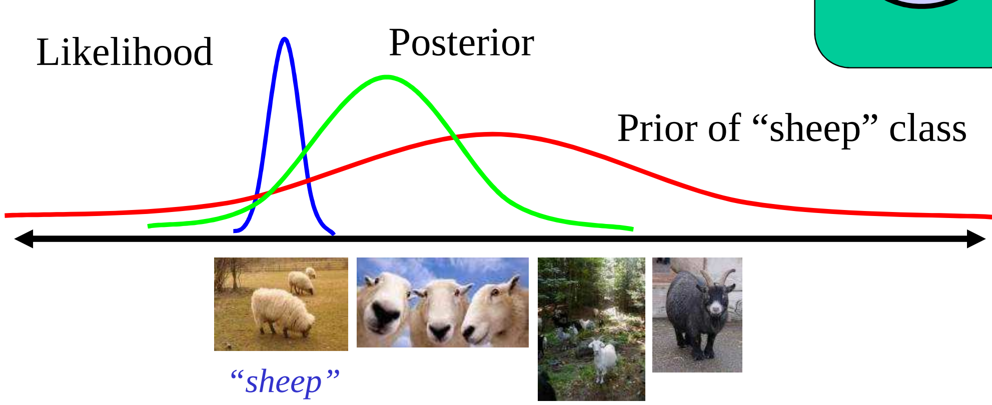
$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

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

Learning and Bayesian inference

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h' \in H} p(d|h')p(h')}$$





Speech recognition

P(words | sound) α P(sound | words) P(words)

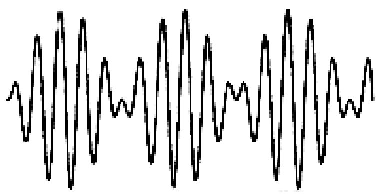
Final beliefs

Likelihood of data
eg mixture of Gaussians

Language model
eg Markov model

Hidden Markov Model (HMM)

"Recognize speech"



"Wreck a nice beach"

Definition of discrete r.v.s

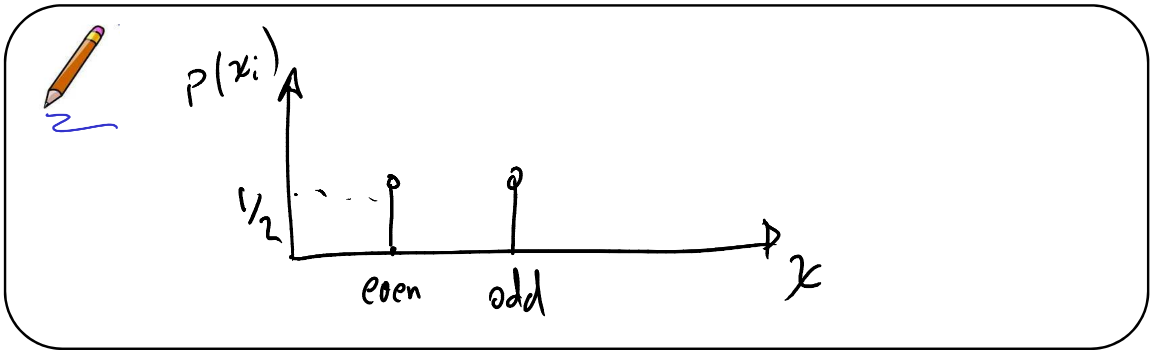
Let E be a discrete set, e.g. $E = \{0, 1\}$. A **discrete random variable** (r.v.) is a map from Ω to E:

$$X(w) : \Omega \mapsto E$$

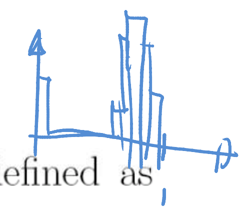
such that for all $x \in E$ we have $\{w | X(w) \leq x\} \in \mathcal{F}$. Since $\mathcal{F}$ denotes the measurable sets, this condition simply says that we can compute (measure) the probability $P(X = x)$.

Probability distributions

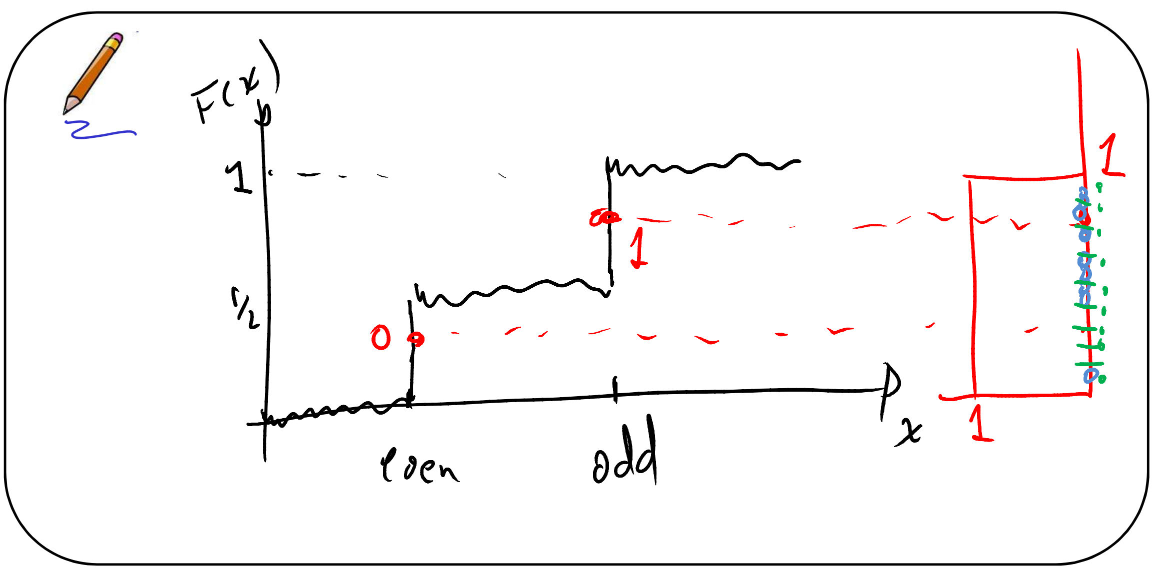
⋆ Assume we are throwing a die and are interested in the events $E = \{even, odd\}$. Here $\Omega = \{1, 2, 3, 4, 5, 6\}$. The r.v. takes the value $X(w) = even$ if $w \in \{2, 4, 6\}$ and $X(w) = odd$ if $w \in \{1, 3, 5\}$. We describe this r.v. with a **probability distribution** $p(x_i) = P(X = x_i) = \frac{1}{2}$, $i = 1, \ldots, 2$



The CDF



The **cumulative distribution function** is defined as $F(x) = P(X \leq x)$ and would for this example be:



Expectation

The expectation of a discrete random variable X is

$$\mathbb{E}[X] = \sum_{E} x_i p(x_i)$$

The expectation operator is linear, so $\mathbb{E}(ax_1+bx_2) = a\mathbb{E}(x_1)+b\mathbb{E}(x_2)$. In general, the expectation of a function $f(X)$ is

$$\mathbb{E}[f(X)] = \sum_{E} f(x_i)\, p(x_i)$$

Mean: $\mu \triangleq \mathbb{E}(X)$

Variance: $\sigma^2 \triangleq \mathbb{E}[(X-\mu)^2]$

Bernoulli r.v.s and the indicator function

Let $E = \{0, 1\}$, $P(X = 1) = \lambda$, and $P(X = 0) = 1 - \lambda$.

We now introduce the *set indicator variable*. (This is a very useful notation.)

$$\mathbb{I}_A(w) = \begin{cases} 1 \;\; if & w \in A; \\ 0 \;\; otherwise. & \end{cases}$$

Using this convention, the probability distribution of a **Bernoulli** random variable reads:

$$p(x) = \lambda^{\mathbb{I}_{\{1\}}(x)}(1-\lambda)^{\mathbb{I}_{\{0\}}(x)}.$$

if $x=1$ $\mathbb{I}_{\{1\}}(x)=1$ $P(X=1)=\lambda$

For **identical and independent distributed** (i.i.d.) data:

$$x_i \sim \underbrace{\theta^{\mathbb{I}_1(x_i)} (1-\theta)^{\mathbb{I}_0(x_i)}}_{P(x_i|\theta)}$$

$$p(x_{1:n}|\theta) = \prod_{i=1}^{n} P(x_i|\theta)$$

$$\mathcal{L}(\theta) = \log p(x_{1:n}|\theta) = \sum_{i=1}^{n} \mathrm{Log}\, P(x_i|\theta)$$

Maximum likelihood example

Let $x_{1:n}$, with $x_i \in \{0, 1\}$, be i.i.d. Bernoulli:

$$p(x_{1:n}|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

$$= \prod_{i=1}^{n} \theta^{\mathbb{I}_1(x_i)} (1-\theta)^{\mathbb{I}_0(x_i)}$$

$$= \theta^{\sum_{i=1}^{n} \mathbb{I}_1(x_i)} (1-\theta)^{\sum_{i=1}^{n} \mathbb{I}_0(x_i)}$$

$$= \theta^{m} (1-\theta)^{n-m}$$

$m \equiv$ # of 1's

$n-m =$ # of 0's

Maximum likelihood example

With $m \triangleq \sum x_i$, we have

$$\mathcal{L}(\theta) = \log P(x_{1:n}|\theta)$$

$$\mathcal{L}(\theta) = m \log \theta + (n-m)\log(1-\theta)$$

Differentiating, we get

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \frac{m}{\theta} + (n-m)\frac{1}{1-\theta}(-1)$$

$$= \frac{m}{\theta} - \frac{n-m}{1-\theta} \rightarrow 0$$

$$\boxed{\theta = \frac{m}{n}}$$

Bayesian learning

Given our **prior** knowledge $p(\theta)$ and the data **model** $p(\cdot|\theta)$, the Bayesian approach allows us to update our prior using the new data $x_{1:n}$ as follows:

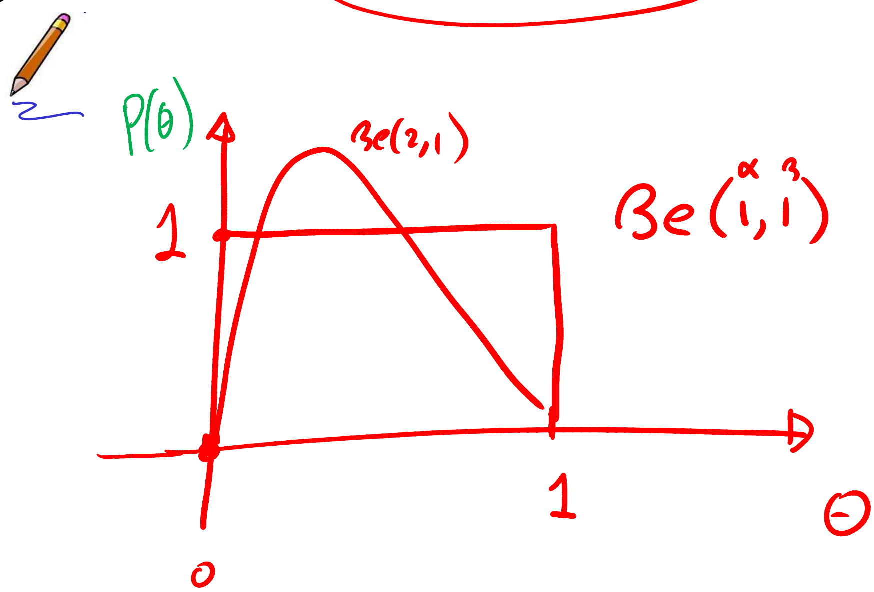
posterior

$$p(\theta|x_{1:n}) = \frac{\overbrace{p(x_{1:n}|\theta)}^{\text{lik}}\overbrace{p(\theta)}^{\text{prior}}}{p(x_{1:n})}$$

where $p(\theta|x_{1:n})$ is the **posterior distribution**, $p(x_{1:n}|\theta)$ is the likelihood and $p(x_{1:n})$ is the **marginal likelihood** (evidence). Note

$$p(x_{1:n}) = \int p(x_{1:n}|\theta)p(\theta)d\theta$$

Beta prior



$$\text{mean}\left[Be(\alpha,\beta)\right] = \frac{\alpha}{\alpha+\beta}$$

Example

Let $x_{1:n}$, with $x_i \in \{0,1\}$, be i.i.d. Bernoulli: $x_i \sim \mathcal{B}(1,\theta)$

$$p(x_{1:n}|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \theta^m (1-\theta)^{n-m}$$

Let us choose the following **Beta** prior distribution:

$$p(\theta) = \int \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta = 1$$

$$\int \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

where Γ denotes the Gamma-function. For the time being, α and β are fixed **hyper-parameters**. The posterior distribution is proportional to:

$$p(\theta|x_{1:n}) \propto P(x_{1:n}|\theta) \quad P(\theta)$$

$$= \theta^{m}(1-\theta)^{n-m} \quad \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$= \theta^{m+\alpha-1}(1-\theta)^{n-m+\beta-1}$$

with normalisation constant

$$P(\theta|x_{1:n}) = \frac{\Gamma(m+\alpha)\,\Gamma(n-m+\beta)}{\Gamma(n+\alpha+\beta)}$$

$$\times \; \theta^{m+\alpha-1}(1-\theta)^{n-m+\beta-1}$$