Lecture 5:
Probability and statistics revision

Nando de Freitas
www.cs.ubc.ca/~nando/340-2009/
September 2009

Outline

In this lecture, we quickly revise the fundamental concepts of probability, including:

- Marginalization
- Conditioning
- Bayes rule
- Expectation

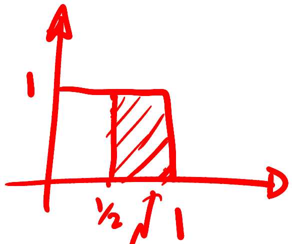
Probability

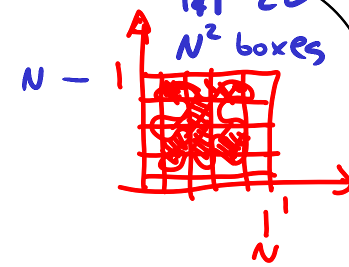
Probability theory is the formal study of the laws of chance. It is our tool for dealing with uncertainty. Notation:

- **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 $w \in \{2, 4, 6\}$

Measure interpretation

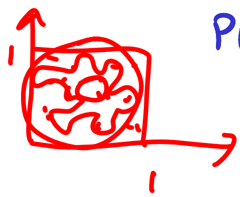
Frequentist interpretation





in 2D N^2 boxes

$$P(\blacksquare) = \frac{1}{2} \qquad P(\square) = \frac{\pi (1/2)^2}{1} = \frac{\pi}{4} \qquad P(\square) = \frac{\#\text{ squares shaded}}{\#\text{ squares}}$$



$$P(\square) = \frac{\#\text{ samples inside}}{\#\text{ samples}}$$

(Monte Carlo)

Make squares smaller for more precision!

- in 3D need N^3 boxes
- in 10D need N^{10} boxes

} CURSE OF DIMENSIONALITY

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 $P(A)$.

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

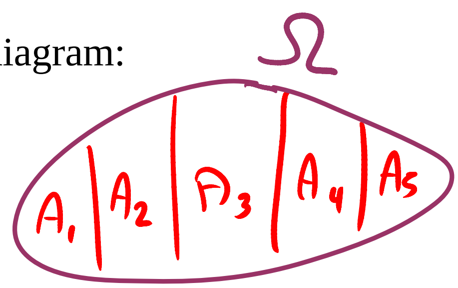
The axioms

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:

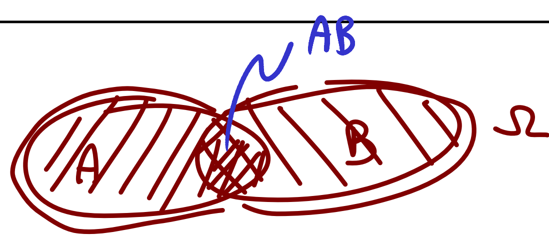


$P(\Omega) = P(A_1) + P(A_2) + \cdots + P(A_5)$

OR and AND operations

OR ... and

$$P(A + B) = P(A) + P(B) - P(AB)$$



Conditional probability

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

$$P(A\overset{\text{given}}{|}B) \triangleq \frac{P(A\overset{\text{and}}{B})}{P(B)}$$

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

$$P(\text{wet}|\text{rain}) = \frac{P(\text{wet}, \text{rain})}{P(\text{rain})}$$

Conditional probability example

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

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

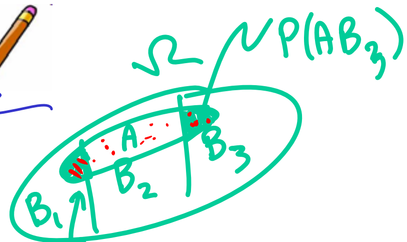
$$P(d_2 = r, d_1 = r) = P(d_2 = r | d_1 = r)\, P(d_1 = r)$$

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

Marginalization

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

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



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

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

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

$$P(A \cap B_1) = P(AB_1) = P(A, B_1)$$

Marginalization example

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

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

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

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

= Exercise!

Bayes rule

Bayes rule allows us to reverse probabilities:

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

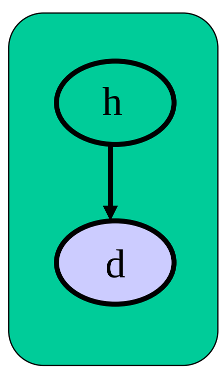
$$P(AB) = P(B|A)P(A) = P(A|B)P(B)$$

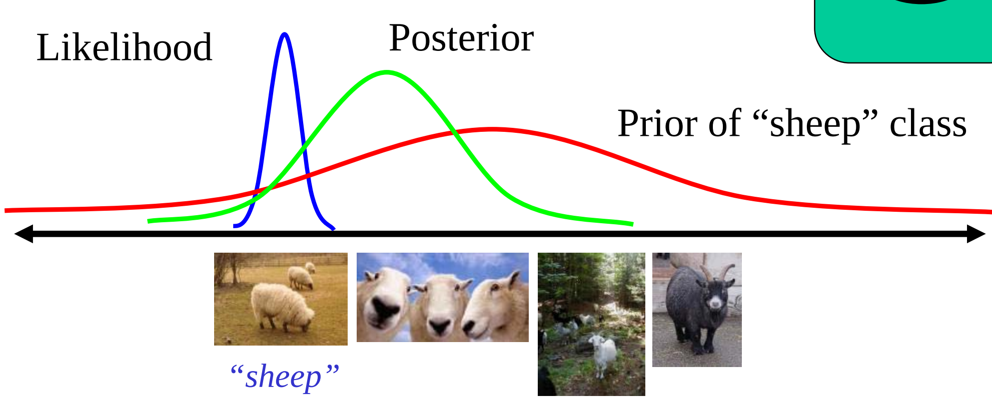
$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

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

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)

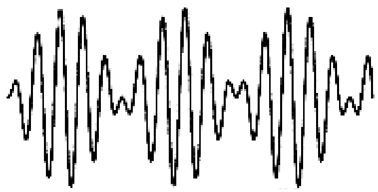
Final beliefs

Likelihood of data
eg mixture of Gaussians

Language model
eg Markov model

Hidden Markov Model (HMM)

"Recognize speech"



"Wreck a nice beach"

Bayes rule: Inverting probabilities

Combinining this with marginalisation, we obtain a powerful tool for statistical modelling:

$$P(model_i|data) = \frac{P(data|model_i)P(model_i)}{\sum_{j=1}^{M} P(data|model_j)P(model_j)}$$

That is, if we have **prior** probabilities for each model and generative data models, we can compute how likely each model is **a posteriori** (in light of our prior knowledge and the evidence brought in by the data).

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

$\star$ 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:

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

Continuous r.v.s

A continuous r.v. is a map to a continuous space, $X(w) : \Omega \mapsto \mathbb{R}$, under the usual measurability conditions. The **cumulative distribution function** $F(x)$ (cdf) is defined by

$$F(x) \triangleq \int_{-\infty}^{x} p(y)\, dy = P(X \leq x)$$

where $p(x)$ denotes the **probability density function** (pdf). For an infinitesimal measure dx in the real line, distributions F and densities p are related as follows:

$$F(dx) = p(x)dx = P(X \in dx).$$

Continuous r.v.s

Lecture 5: Maximum likelihood and Bayesian learning

Nando de Freitas
www.cs.ubc.ca/~nando/340-2009/

October 2009

Outline

We revise maximum likelihood (ML) for a simple binary model. We then introduce Bayesian learning for this simple model.

The key difference between the two approaches is that the frequentist view assumes there is one true model responsible for the observations, while the Bayesian view assumes that the model is a random variable with a certain prior distribution. Computationally, the ML problem is one of optimization, while Bayesian learning is one of integration.

Frequentist learning

Frequentist Learning assumes that there is a true model (say a parametric model with parameters θ_0). The estimate is denoted $\widehat{\theta}$. It can be found by maximising the **likelihood**:

$$\widehat{\theta} = \arg\max_{\theta} \quad p(x_{1:n}|\theta)$$

distributed according

For **identical and independent distributed** (i.i.d.) data: $X_i \sim \underbrace{\theta^{\mathbb{I}_1(x_i)} (1-\theta)^{\mathbb{I}_0(x_i)}}_{P(x_i|\theta)}$

$x_1 = 1$, $x_2 = 1$, $x_3 = 0$; $m = 2$, $n = 3$

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

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

$P(x_i = 1) = \theta \qquad P(x_i = 0) = 1-\theta$

$2^3 2^4 = 2^{3+4}$

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 = # 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

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

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

$$\boxed{\theta = \frac{m}{n}} = \frac{2}{3}$$

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:

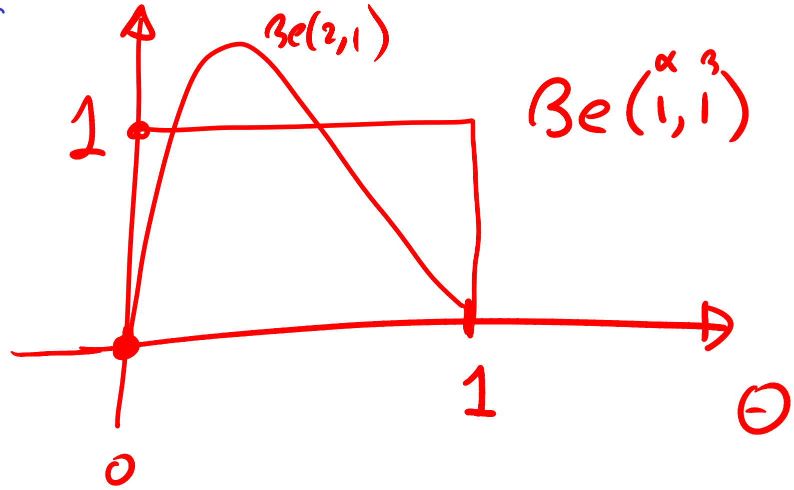
$$\underset{\text{posterior}}{p(\theta|x_{1:n})} = \frac{\overbrace{p(x_{1:n}|\theta)}^{\text{lik}}\overbrace{p(\theta)}^{\text{prior}}}{p(x_{1:n})}$$

110

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



$Be(\overset{\alpha}{1}, \overset{\beta}{1})$

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

Beta prior

Bayesian model selection

For a particular model structure M_i, we have

$$p(\theta|x_{1:n}, M_i) = \frac{p(x_{1:n}|\theta, M_i)p(\theta|M_i)}{p(x_{1:n}|M_i)}$$

Models are selected according to their posterior:

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

The ratio $P(x_{1:n}|M_i)/P(x_{1:n}|M_j)$ is known as the **Bayes Factor**.

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:

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

$$\int \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta$$

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

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

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

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

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

beta!

with normalisation constant

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

Conjugate analysis

Since the posterior is also Beta, we say that the Beta prior is **conjugate** with respect to the binomial likelihood. Conjugate priors lead to the same form of posterior.

Different hyper-parameters of the Beta $\mathcal{B}e(\alpha, \beta)$ distribution give rise to different prior specifications: