COMP30027 Machine Learning
Revision of Probability Theory

Semester 1, 2019

Jeremy Nicholson & Tim Baldwin & Karin Verspoor

Lecture Outline

1. Probability Theory
   - The Basics
   - Conditional Probability
   - Distributions
   - Entropy

2. Modelling
   - What is a model?
   - Probability models

Probability Theory

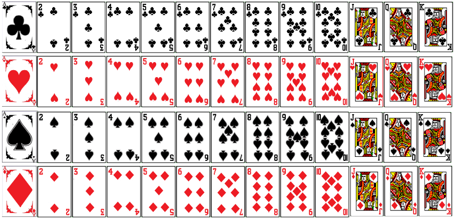
"The calculus of probability theory provides us with a formal framework for considering multiple possible outcomes and their likelihood. It defines a set of mutually exclusive and exhaustive possibilities, and associates each of them with a probability — a number between 0 and 1, so that the total probability of all possibilities is 1. This framework allows us to consider options that are unlikely, yet not impossible, without reducing our conclusions to content-free lists of every possibility."

From Probabilistic Graphical Models: Principles and Techniques (2009; Koller and Friedman) `http://pgm.stanford.edu/intro.pdf`

(Very) Basics of Probability Theory

- $P(A)$: the probability of A = proportion of successful outcomes out of possible outcomes

$$0 <= P(A) <= 1$$
$$P(True) = 1$$
$$P(False) = 0$$

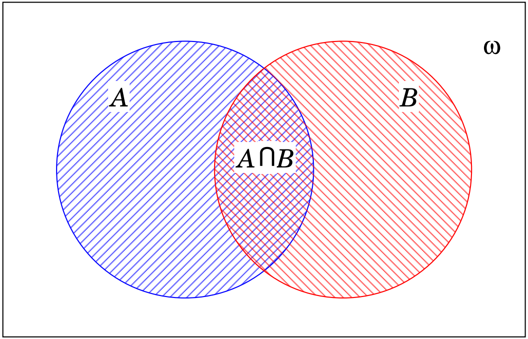


- Given a deck of 52 cards; 13 ranks (A, 2–10, J, Q, K) of each of four suits ($\clubsuit, \spadesuit$ = black; $\heartsuit, \diamondsuit$ = red)

$$P(K) = \frac{4}{52},\ P(red) = \frac{26}{52},\ P(\heartsuit) = \frac{13}{52}$$

Basics of Probability Theory

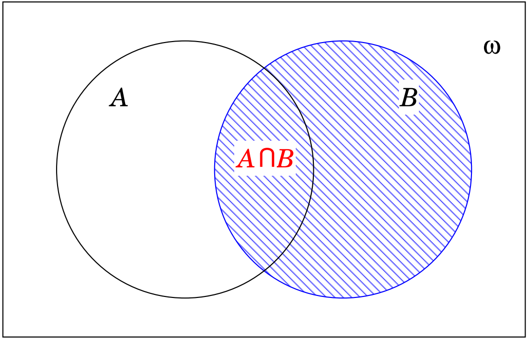
- **Joint probability** $(P(A, B))$:
  the probability of both A and B occurring $= P(A \cap B)$



$P(K, \heartsuit) = \frac{1}{52}$, $P(\heartsuit, red) = \frac{13}{52}$

Conditional Probability

- **Conditional probability** ($P(A|B) = \frac{P(A \cap B)}{P(B)}$):
  the probability of A occurring, given the occurrence of B



$P(K|\heartsuit) = \frac{1}{13}$, $P(\heartsuit|red) = \frac{13}{26}$

Working with Probability

- **Sum rule**: $P(A) = \sum_B P(A \cap B)$
- **Product rule**: $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$
- **Bayes' rule**: $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$
- **Chain rule**: $P(A_1 \cap ... \cap A_n) =$ $P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1) \dots P(A_n| \cap_{i=1}^{n-1} A_i)$

Probability terminology

- **Prior probability** ($P(A)$): the probability of A occurring, given no additional knowledge about A
- **Posterior probability** ($P(A|B)$): the probability of A occurring, given background knowledge about event(s) B leading up to A
- **Independence:** A and B are independent iff $P(A \cap B) = P(A)P(B)$; equivalently $P(A|B) = P(A)$
- **Conditional Independence:** A and B are independent, conditioned on C, iff $P(A \cap B|C) = P(A|C)P(B|C)$

Bayes' Rule

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

For proposition A and evidence B,

- $P(A)$, the prior, is the initial degree of belief in A.
- $P(A|B)$, the posterior, is the degree of belief in A, having accounted for B.
- $P(B|A)$ the likelihood of observing the evidence, given that the proposition is true

Bayes' Rule

Bayes' Rule is important because it allows us to compute $P(A|B)$, where $P(B|A)$ can be sensibly estimated (according to some model).

For instance, imagine we believe (from observed data), that $P(\mathrm{Smart}) = 0.3$, $P(\mathrm{H1}) = 0.2$ and $P(\mathrm{H1}|\mathrm{Smart}) = 0.6$.

What is the probability that a given student, who has received a mark of H1, is "smart"? i.e. $P(\mathrm{Smart}|\mathrm{H1})$?

(What if $P(\mathrm{H1}) = 0.4$? What if $P(\mathrm{Smart}) = 0.4$?)

Binomial Distributions

- A **binomial distribution** results from a series of independent trials with only two outcomes (**Bernoulli trials**) e.g. multiple coin tosses ($\langle H, T, H, H, ..., T \rangle$)
- The probability of an event with probability p occurring exactly m out of n times is given by

$$\begin{aligned} B(m; n, p) &= \binom{n}{m} p^m (1-p)^{n-m} \\ &= \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m} \end{aligned}$$

Intuition: we want exactly m successes (p^m) and $n - m$ failures ($(1-p)^{n-m}$). However, the successes can occur anywhere among the trials: there are C(n, m) different ways of distributing them.

Binomial Example: Coin Toss

If we toss a fair coin 3 times, what is the probability that we get exactly 2 heads?

X=number of heads when flipping coin 3 times; $P(X = 2)$

Possible outcomes from 3 coin flips $= 2 * 2 * 2 = 2^3 = 8$. Each possible outcome has $\frac{1}{8}$ probability.

Choose 2 heads out of 3 flips ($C(3, 2) = \frac{3!}{2!1!} = 3$).

So, 3 possible outcomes, $\frac{1}{8}$ for each, $P(X = 2) = \frac{3}{8}$

$$P\left(2; 3, \frac{1}{8}\right) = \frac{3!}{2!(3-2)!}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{3-2} = 3\left(\frac{1}{4}\right)\left(\frac{1}{2}\right)$$

Multinomial Distributions

- A **multinomial distribution** results from a series of independent trials with more than two outcomes

  e.g. two players in a chess tournament, 3 outcomes: (Player A wins, Player B wins, players draw); probability that Player A wins is 0.4, that player B wins is 0.35, probability of draw is 0.25

- The probability of events $X_1, X_2, ..., X_n$ with probabilities $p_1, p_2, ..., p_n$ occurring exactly $x_1, x_2, ..., x_n$ times, respectively, is given by

  $P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = (\sum_i x_i)! \prod_i \frac{p_i^{x_i}}{x_i!}$

If these two chess players played 12 games, what is the probability that Player A would win 7 games, Player B would win 2 games, and the remaining 3 games would be drawn?

Entropy (Information Theory)

- (Shannon) **Entropy**: a measure of **unpredictability** — the information required to predict an event
- The entropy (in bits) of a discrete random variable X with possible states $x_1, .., x_n$ is:

$$H(X) \;=\; -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

$$\text{where } 0 \log_2 0 \overset{\text{def}}{=} 0$$

Sidebar — Message Encoding I

- Consider a message M composed of distinct symbols $w_1, \ldots, w_n$, where each symbol w_i has a frequency f_i. The total length of the message is $|M| = \sum_i f_i$.
- Information theory tells us that the minimum length encoding of the message is to allocate $L(w_i) = -\log_2 \frac{f_i}{|M|}$ bits to symbol w_i.
- So, common symbols (high f_i) get assigned a small number of bits and rare symbols get a large number of bits.

Sidebar — Message Encoding II

The expected number of bits per symbol in the message:

$$\begin{aligned} E[W; M] &= \sum_i P(w_i) L(w_i) \\ &= \sum_i \frac{f_i}{|M|} \times - \log_2 \frac{f_i}{|M|} \end{aligned}$$

The theoretical minimum length of the message (in the context of the provided information) is $H(M) \times |M|$

Interpreting Entropy Values – High

In practice, entropy measures the *evenness* of a probability distribution.

- A high entropy value means x is unpredictable.
  - fair coin → impossible to predict outcome of coin toss ahead of time

$$\begin{aligned} H(x) &= -(P(X=h)\log_2 P(X=h) + P(X=t)\log_2 P(X= \\ &= -(0.5\log_2 0.5 + 0.5\log_2 0.5) \\ &= -((0.5*-1)+(0.5*-1)) = -(-1) = 1 \end{aligned}$$

  - Two possible outcomes with equal probability; Learning the outcome gives one bit of information

Interpreting Entropy Values – Low

- A low entropy value means x is (more) predictable.
    - A coin toss with two heads is perfectly predictable.

$$H(x) \quad = \quad -(1 \log_2 1 + 0 \log_2 0) = -(0+0) = 0$$

    - We don't learn anything once we see the outcome.
- Distributions are "peaky", uneven.

Entropy of an unfair coin

Let's say $P(X = h) = 0.9$ and $P(X = t) = 0.1$

$$\begin{aligned} H(x) &= -(P(X = h) \log_2 P(X = h) + P(X = t) \log_2 P(X = t)) \\ &= -(0.9 \log_2 0.9 + 0.1 \log_2 0.1) \\ &= 0.47 \end{aligned}$$

Entropy values

NB: The range of the entropy values is not $[0, 1]$.

- The range is determined by the possible number of outcomes.
- $0 \leq$ Entropy $\leq$ log(n), where n is number of outcomes
- Entropy=0 (minimum entropy) when one probability is 1, others 0
- Entropy=log(n) (maximum entropy): when all probabilities have equal values of 1/n

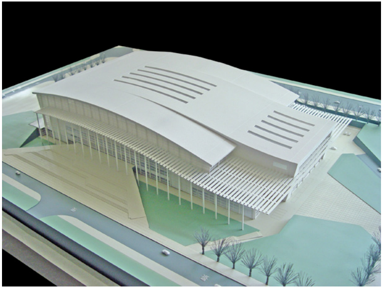
Lecture Outline

1. Probability Theory
   - The Basics
   - Conditional Probability
   - Distributions
   - Entropy

2. Modelling
   - What is a model?
   - Probability models

What is a model?

We talk about *modelling* data, building *models*.
So, what makes a *model* a model?



What is a model?

A model is our attempt to understand and represent the nature of reality through a particular lens, be it architectural, biological, or mathematical.

A model is an artificial construction where all extraneous detail has been removed or abstracted.

Attention must always be paid to these abstracted details after a model has been analysed to see what might have been overlooked.

What is a model?

A model tries to capture

- The component parts (parameters or variables).
- A representation of what influences/causes what.

to produce

- A mathematical function that approximates a set of observations (and can be estimated from data).

There is a trade-off in modelling between *simplicity* and *accuracy*.

Probability Models

A probability model is a mathematical representation of a random event.

It consists of:

- A sample space — the set of all possible outcomes
- Events — a subset of the sample space
- A **probability distribution** over the events

A *probability model* predicts the relative frequency of each event if the experiment is performed a large number of times. It captures a *theoretical probability*.

Estimating Probabilities I

In the examples above:

- we were given the "true" probabilities (e.g. $P(X = h) = 0.9$),
- or the sample space was exhaustively enumerated (e.g. the deck of cards)

Estimating Probabilities II

...But what if we don't know the "true" probabilities?

- Because we don't know which events are actually possible?
- Because we know which events are possible, but there are too many to enumerate?
- Because we know which events are possible, there are a manageable number of events, **but we don't have enough data to reliably estimate the relative distribution**?

So, we want to attempt to make a (good) estimate of the probabilities, based on an incomplete sample.

Estimating Probabilities III

For example:

- We are moving house; we have packed up our belongings into a number of boxes.
- At our new home, we open up one box, (we don't look inside!) and pull out three objects:
  - A book
  - Another book
  - A toothbrush
- What is the probability that the fourth object we pull out of this box is another book?

Estimating Probabilities IV

How do we determine the (theoretical) probability of an event?

- The most obvious (frequentist) way is to assume the incomplete sample is the complete sample:

$$\begin{aligned}
\hat{P}(x) &= \frac{freq(x)}{\sum_k freq(x_k)} = \frac{freq(x)}{N} \\
\hat{P}(x, y) &= \frac{freq(x, y)}{N} \\
\hat{P}(x|y) &= \frac{freq(x, y)}{freq(y)}
\end{aligned}$$

Estimating Probabilities V

How do we determine the (theoretical) probability of an event?

- Better statistical answer: choose a probability distribution which maximises the **likelihood** of observing the given sample (**Maximum Likelihood Estimation** (MLE))
- Interestingly, for a discrete distribution, the way to do this is by counting frequencies (as a Bernoulli trial for $X = x$) ... the obvious way!
- (For many continuous distributions — like *normal distributions* — the obvious answer also maximises the likelihood, but this is beyond the scope of this subject)

Estimating Probabilities VI

...But what happens if we try to estimate the probability of an event x (or combination of events x, y) that we've never seen before?

More tomorrow!

Summary

Probability forms the foundation of many machine learning methods.

- What are joint and conditional probabilities?
- What are prior and posterior probabilities?
- What is entropy, and how should you interpret entropy values?
- What is a probability model, and how can one be derived?