C24 Advanced Probability Theory

Michael A. Osborne

mosb@robots.ox.ac.uk
www.robots.ox.ac.uk/∼mosb/c24

Michaelmas 2014

Welcome to Advanced Probability Theory!

www.robots.ox.ac.uk/∼mosb/c24 will hold copies of the

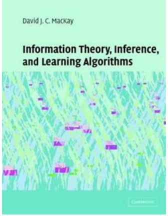
- lectures slides/notes
- tutorial sheet
- FAQs

Please get in touch if you spot anything unclear or incorrect. The reply (if generally useful) will get added to the web FAQs.

There are many useful texts.

- **Bayesian Reasoning and Machine Learning**
  D. Barber, CUP, 2012
  Up-to-date and comprehensive.
  Available free online (legally)!

- **Information Theory, Inference, and Learning Algorithms**
  D.J.C. MacKay, CUP, 2003
  Covers all the course material, though at an advanced level.
  Available free online (legally)!



Topic 1: Bayesian Model Comparison and the Value of Information

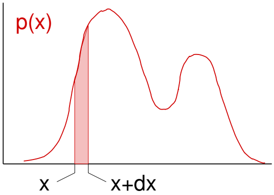
Recall that probability theory is specified by two rules in either continuous or discrete cases.

The probability density function (pdf) is defined as

$$p(x) = p(X = x) = \lim_{\delta x \to 0} \frac{P(x < X < x + \delta x)}{\delta x}$$

Note the the lowercase p for pdfs.
This gives the continuous sum rule

$$\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1\,.$$



The product rule is (for joint $p(a,\ b)$)

$$p(a,\ b) = p(a)\, p(b|a) = p(b)\, p(a|b)$$

Bayes' rule is an important reformulation of the product rule.

$$p(a \mid b) = \frac{p(b \mid a)\, p(a)}{p(b)}$$

1. $p(a \mid b)$ is called the posterior for a.
2. $p(a)$ is called the prior for a.
3. $p(b \mid a)$ is called the likelihood of a, and is usually considered as a function of a, $\mathcal{L}(a) = p(b \mid a)$.
4. $p(b)$ is called the evidence, or marginal likelihood.

   The latter name is due to the fact that $p(b) = \int p(a', b)\mathrm{d}a' = \int p(b \mid a')p(a')\mathrm{d}a'$.

♣ Consider a trial of a new component in an aircraft engine.

The component is put into two types of engine, A and $\neg A$, and the compiled results from both engines are below.

	R	$\neg R$	Total	Reliability Rate
N	200	200	400	50%
$\neg N$	160	240	400	40%

1. N means the new component was used, $\neg N$ means it wasn't.
2. R means the engine was sufficiently reliable, $\neg R$ means it wasn't.

Hence we would say $P(R \mid N) = 0.5$ and $P(R \mid \neg N) = 0.4$ and so advise the use of the component.

♣ Now consider the results for A alone.

	R	$\neg R$	Total	Reliability Rate
N	180	120	300	60%
$\neg N$	70	30	100	70%

1. N means the new component was used, $\neg N$ means it wasn't.
2. R means the engine was sufficiently reliable, $\neg R$ means it wasn't.

Hence we would say $P(R \mid N, A) = 0.6$ and $P(R \mid \neg N, A) = 0.7$ and so advise **against** the use of the component for A.

♣ Now consider the results for $\neg A$ alone.

	R	$\neg R$	Total	Reliability Rate
N	20	80	100	20%
$\neg N$	90	210	300	30%

1. N means the new component was used, $\neg N$ means it wasn't.
2. R means the engine was sufficiently reliable, $\neg R$ means it wasn't.

Hence we would say $P(R \mid N, \neg A) = 0.2$ and $P(R \mid \neg N, \neg A) = 0.3$ and so advise **against** the use of the component for $\neg A$.

Hang on a second: the component is good overall, but bad for each type of engine?

This phenomenon is known as Simpson's "paradox".

The explanation can be found by noting that

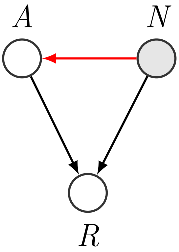
$$
\begin{aligned}
P(A \mid N) &= \frac{P(N \mid A)P(A)}{P(N)} \\
&= \frac{{}^{300}\!/\!_{300+100} \times {}^{300+100}\!/\!_{400+400}}{{}^{300+100}\!/\!_{400+400}} \\
&= 0.75.
\end{aligned}
$$

That is, many more new components were put into A engines than $\neg A$ engines in this trial.

Hence we must be cautious about interpreting results from the aggregated population! Here, we should advise **against** the use of the component.

Understanding Simpson's paradox is aided with a Bayes net.

Our intuitive notion of using $P(R \mid N)$ alone to decide on the benefits of N ignores the red edge, $P(A \mid N)$.



Even if we weren't able to control or observe the type of engine, A, we should use our prior information to determine $P(A \mid N)$.
NB: the probabilities above were empirical frequencies.

Let’s consider the influence of race on death penalties for murder in the US.

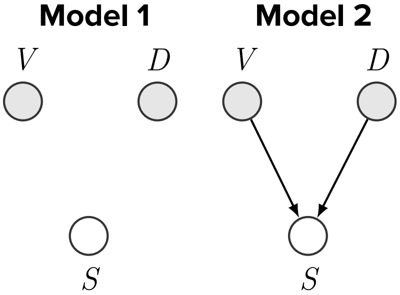
Taken from Mackay (2003, p354), using data from Radelet (2001).

White defendant			Black defendant		
	Death Penalty			Death Penalty	
	Yes	No		Yes	No
White victim	19	132	White victim	11	52
Black victim	0	9	Black victim	6	97

Note that $^{19}/_{160} = 12\%$ of white defendants are sentenced to death compared to $^{17}/_{166} = 10\%$ of black defendants. However, where the victim is white, a black defendant is more likely to receive the death penalty ($^{11}/_{63} > ^{19}/_{151}$), as is also true for black victims ($^{6}/_{103} > ^{0}/_{9}$).

Which model provides the best explanation?

V is the race of the victim.
D is the race of the defendant.
S is the sentence awarded.



Model 1 says that neither the defendent's race nor the victim's race affect the sentence.
Model 2 says that both the defendent's race and the victim's race affect the sentence.

We ideally want $P(\mathcal{M} = 1 \mid \mathcal{D})$.

That is, we want the probability of the thing we're interested in (the true model, $\mathcal{M}$) given what we know (which is data $\mathcal{D}$, such as the counts in the table above),

$$P(\mathcal{M} = 1 \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{M} = 1) P(\mathcal{M} = 1)}{\sum_i P(\mathcal{D} \mid \mathcal{M} = i) P(\mathcal{M} = i)}.$$

Unfortunately, there are two quite profound problems.

1. What's the prior $P(\mathcal{M} = i)$?
2. How do you work out the sum over all possible models i?

Hence we settle for $P(\mathcal{D} \mid \mathcal{M} = 1)$, as

$$P(\mathcal{M} = 1 \mid \mathcal{D}) \propto P(\mathcal{D} \mid \mathcal{M} = 1).$$

$P(\mathcal{D} \mid \mathcal{M})$ has a name: it is the evidence, or marginal likelihood.

Hang on: wasn't the evidence some normalisation factor? Yes! This arises from

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

where $P(\mathcal{D} \mid \mathcal{M}) = \int P(\mathcal{D} \mid \theta, \mathcal{M})\ p(\theta \mid \mathcal{M})\ \mathrm{d}\theta$.

Let's derive that by starting with the probability of everything.

$$
\begin{aligned}
p(\theta \mid \mathcal{D}, \mathcal{M}) &= \frac{p(\mathcal{D}, \theta, \mathcal{M})}{p(\mathcal{D}, \mathcal{M})} \\
&= \frac{p(\mathcal{D} \mid \theta, \mathcal{M})\ p(\theta \mid \mathcal{M})\ p(\mathcal{M})}{p(\mathcal{D} \mid \mathcal{M})\ p(\mathcal{M})} \\
&= \frac{p(\mathcal{D} \mid \theta, \mathcal{M})\ p(\theta \mid \mathcal{M})}{p(\mathcal{D} \mid \mathcal{M})}
\end{aligned}
$$

θ are the parameters of the model.

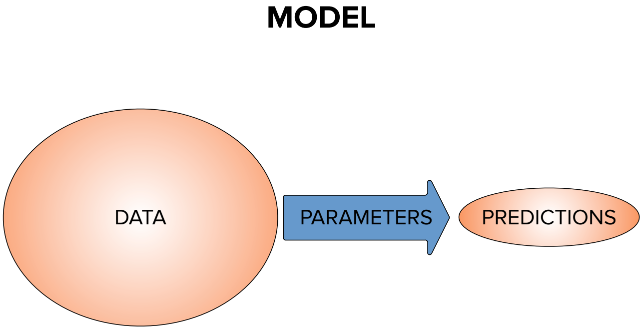
Let's predict for $f_\star$ in the presence of parameters θ.

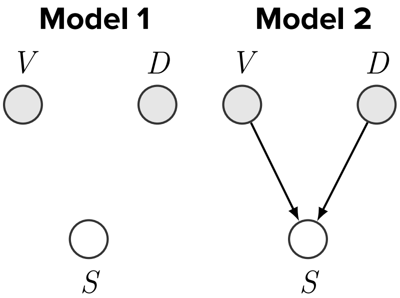
$$p(f_\star \mid \mathcal{D}, \mathcal{M}) = \int p(f_\star \mid \mathcal{D}, \theta, \mathcal{M})\, p(\theta \mid \mathcal{D}, \mathcal{M})\, \mathrm{d}\theta$$

1. $p(f_\star \mid \mathcal{D}, \mathcal{M})$ is the posterior for $f_\star$; this is our goal.
2. $p(f_\star \mid \mathcal{D}, \theta, \mathcal{M})$ are the predictions given θ.
3. $p(\theta \mid \mathcal{D}, \mathcal{M})$ is the posterior for θ, from above.

The predictions are averaged over, weighted by the parameter posterior.

A model is just a framework for allowing data to influence predictions via some parameters.

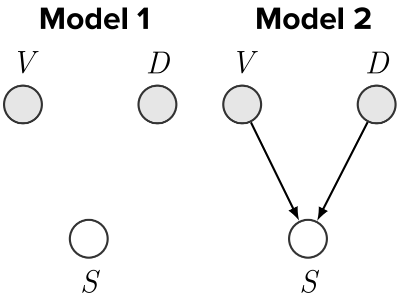




How many parameters do we need to specify

$$P(S \mid V, D, \mathcal{M} = 1)?$$

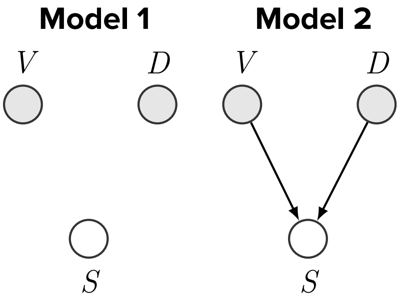
1. 1.
2. 4.
3. 6.
4. 8.



How many parameters do we need to specify

$$P(S \mid V, D, \mathcal{M} = 1)?$$

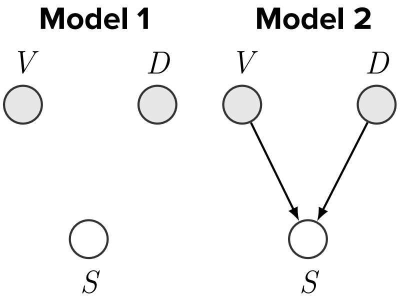
1. $1.$
2. $4.$
3. $6.$
4. $8.$



How many parameters do we need to specify

$$P(S \mid V, D, \mathcal{M} = 2)?$$

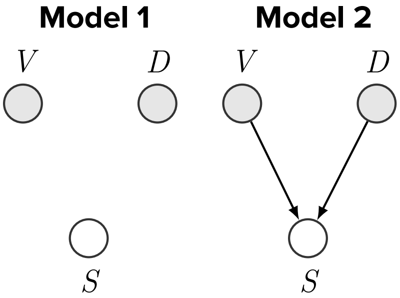
1. $1.$
2. $4.$
3. $6.$
4. $8.$



How many parameters do we need to specify

$$P(S \mid V, D, \mathcal{M} = 2)?$$

1. $1.$
2. $4.$
3. $6.$
4. $8.$

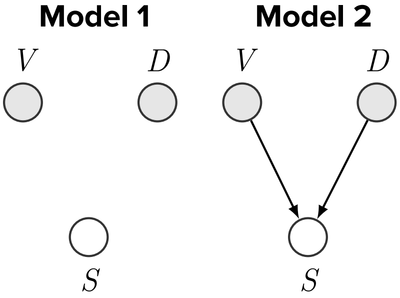


$P(S \mid V, D, \mathcal{M} = 1) = \theta_0 \Rightarrow$ one parameter for Model 1.

$P(S \mid V, D, \mathcal{M} = 2)$	$V =$ White	$V =$ Black
$D =$ Black	θ_1	θ_2
$D =$ White	θ_3	θ_4

$\Rightarrow$ four parameters for Model 2.

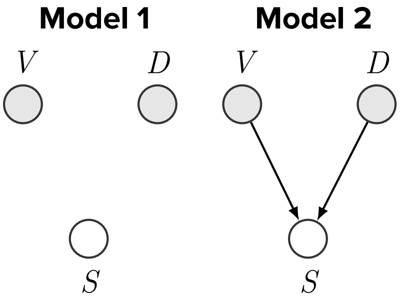
Model 1 is a special case of Model 2.



That is, you can represent any joint probability with Model 2 that you can with Model 1, just by choosing all parameters to be equal.

So aren't we always going to pick Model 2? That would be bad: we would lose the ability to distinguish structures!

Model 1 is a special case of Model 2, but Model 2 has more parameters.



We need some way to penalise complex models for their additional parameters, or else they will tend to overfit. That is, complex models will slavishly match structures in the data that we may not expect to be reproduced in new data (we'll return to this!).

Let's return to the evidence, $P(\mathcal{D} \mid \mathcal{M})$,

which is computed as

$$\boxed{P(\mathcal{D} \mid \mathcal{M}) = \int P(\mathcal{D} \mid \theta, \mathcal{M})\; p(\theta \mid \mathcal{M})\, \mathrm{d}\theta.}$$

Our general strategy is to pick the model (amongst those we can imagine) with the highest evidence.

Let’s compute the evidence for Model 1.

Under Model 1, we’ll assume that all sentences are independent, that $p(S) = \sigma$, and that the prior for sigma is a uniform distribution over $[0, 1]$.

$$
\begin{aligned}
&P(\mathcal{D} \mid \mathcal{M} = 1) \\
&= \int p(\mathcal{D} \mid \sigma, \mathcal{M} = 1) p(\sigma \mid \mathcal{M} = 1) \mathrm{d}\sigma \\
&= \int_0^1 p(\mathcal{D} \mid \sigma, \mathcal{M} = 1) \mathrm{d}\sigma \\
&= \int_0^1 \prod_i p(\mathcal{D}_i \mid \sigma, \mathcal{M} = 1) \mathrm{d}\sigma
\end{aligned}
$$

Let's compute the evidence for Model 1.

	White defendant		Black defendant		
	Death Penalty			Death Penalty	
	Yes	No		Yes	No
White victim	19	132	White victim	11	52
Black victim	0	9	Black victim	6	97

$$
\begin{aligned}
P(\mathcal{D} \mid \mathcal{M} = 1) &= \int_0^1 \prod_i p(\mathcal{D}_i \mid \sigma, \mathcal{M} = 1) \mathrm{d}\sigma \\
&= \int_0^1 \sigma^{19+0+11+6} (1-\sigma)^{132+9+52+97} \mathrm{d}\sigma \\
&= \int_0^1 \sigma^{36} (1-\sigma)^{290} \mathrm{d}\sigma = 2.8 \times 10^{-51},
\end{aligned}
$$

where the numerical result was calculated using the Beta function. The evidence for Model 2 is left as an exercise.

Let's find the evidence for a new Model 0.

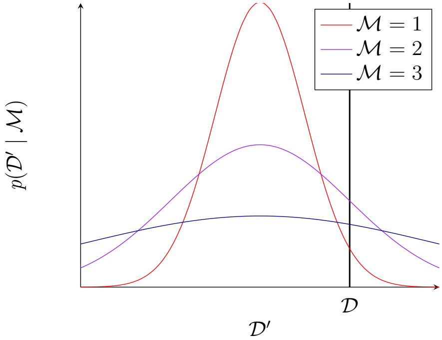
Let's build a new Model 0 that assumes that the sentence is set by a fair coin flip: it has no parameters.

	White defendant			Black defendant	
	Death Penalty			Death Penalty	
	Yes	No		Yes	No
White victim	19	132	White victim	11	52
Black victim	0	9	Black victim	6	97

$$P(\mathcal{D} \mid \mathcal{M} = 0) = (1/2)^{36} (1/2)^{290}$$
$$= 7.3 \times 10^{-99} \ll 2.8 \times 10^{-51} = P(\mathcal{D} \mid \mathcal{M} = 1) :$$

Hence Model 1 is to be preferred to Model 0.

The evidence penalises models that can explain too many different datasets.



Here $\mathcal{D}$ is the data actually observed, and $\mathcal{D}'$ is the random variable, spanning all possible datasets.
$\mathcal{M} = 1$ is too simple, $\mathcal{M} = 3$ is too complex and $\mathcal{M} = 2$ is just right.

A model must have just enough complexity.

Imagine an annoying friend who always knows exactly why nation X won the world cup after the event.

"It was because of that one particular player's club performances and the weather and the national track record over the last seven games" etc.

The problem is: this model, where any of a million different things can be combined to provide explanation (that is, the model has a million parameters), could predict almost anything.

Essentially, this model is not a lot different from that of my Aussie-rules-loving granny, who knows nothing about soccer: Both essentially place a uniform prior over all possible winners.

A model must have just enough complexity.

Let's go to the opposite extreme: a one-eyed-zealot who only predicts England as the winner, no matter what.

This is an exceedingly simple model: the only problem is that it's never right.

Most trustworthy is the expert who is only ever willing to entertain, say, three nations as possible winners beforehand, and is consistently proven right.

Simpson's paradox is a good example of the importance of experimental design.

The choice of experiments can have an important impact on what we can conclude.

Example: in our engine component example, the small number of examples of new components in $\neg A$ engines should reduce our confidence in the estimates for $P(R \mid N, \neg A)$ and $P(R \mid \neg N, \neg A)$.

e.g. If we had no $\neg A$ engines with new components N, it would be folly to conclude much about the influence of N.

How can we select the best experiments?

Recall that decision theory specifies how an agent should take action.

An agent needs to be equipped with a loss function, $\lambda(x; a)$, which gives the cost of any realisation of (all relevant) random variables X, given action A.

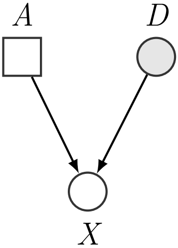
Decision theory then simply states that an agent should take the action that minimises its expected loss. That is, the agent should choose the action

$$\operatorname*{argmin}_{a} \int \lambda(x; a)\ p(x \mid a)\ \mathrm{d}x.$$

Equivalently, you can define a utility function (a negative loss), and maximise the expected utility.

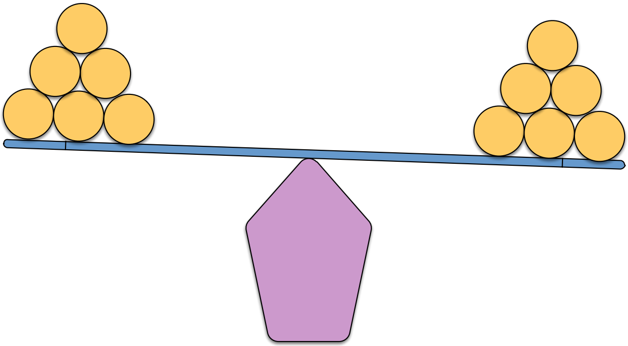
A loss function assigns a cost to every possible state of the world.

We then weight these costs by their probabilities in light of what we know and sum to compute an expected loss.



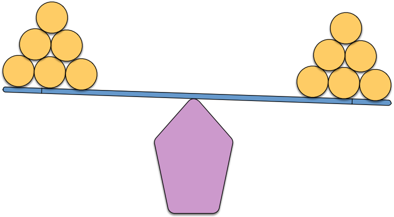
The square node is our decision, which we assume is always independent of everything else (we have complete autonomy.)

♣ We must design weighing experiments to find an odd ball.



We have 12 balls: all are the same weight, except for one that is either lighter or heavier. Using a scale, find the odd ball, and whether it is lighter or heavier.

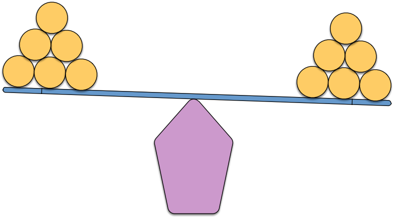
We have 12 balls: all are the same weight, except for one that is either lighter or heavier.



What should our first measurement be?

1. Five balls against seven balls.
2. Six balls against six balls.
3. One ball against one ball, with ten left out.
4. Four balls against four balls, with four left out.

We have 12 balls: all are the same weight, except for one that is either lighter or heavier.



What should our first measurement be?

1. Five balls against seven balls.
2. Six balls against six balls.
3. One ball against one ball, with ten left out.
4. Four balls against four balls, with four left out.

We want experiments whose outcomes are as close as possible to equiprobable.

For a weighing experiment, there are three possible outcomes:

1. both sides are balanced;
2. the left side is heavier; and
3. the left side is lighter.

Note that only weighing one ball against one ball is most likely to return a balanced measurement: this is sub-optimal.

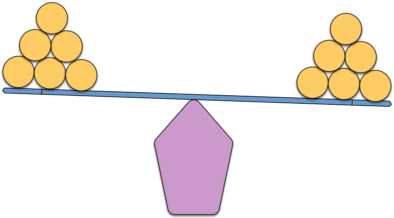
Conversely, weighing six balls against the other six can never return a balanced measurement: this too is sub-optimal.

We want to maximise the expected amount we learn with each experiment.

Our four-against-four test has three possible outcomes:

1. balanced, with probability $1/3 \Rightarrow$ the odd ball must be in the held-out four and it
   1. is heavy, with probability $1/2$, or
   2. is light, with probability $1/2$;
2. left heavier, with probability $1/3 \Rightarrow$ either the odd ball is
   1. is heavy and is amongst the left four, with prob. $1/2$, or
   2. is light and is amongst the right four, with prob. $1/2$;
3. left lighter, with probability $1/3 \Rightarrow$ either the odd ball is
   1. is light and is amongst the left four, with prob. $1/2$, or
   2. is heavy and is amongst the right four, with prob. $1/2$.

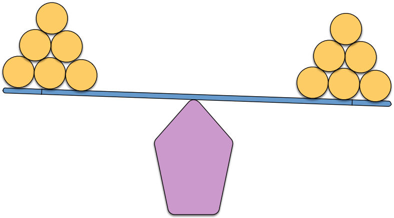
We have 12 balls: all are the same weight, except for one that is either lighter or heavier.



How many measurements in total do we need to perfectly identify the ball?

1. 3.
2. 4.
3. Maybe 3, maybe 4, depending.
4. 5.

We have 12 balls: all are the same weight, except for one that is either lighter or heavier.



How many measurements in total do we need to perfectly identify the ball?

1. $3.$
2. $4.$
3. Maybe 3, maybe 4, depending.
4. $5.$

Our first measurement narrows the 12×2 possibilities for the odd ball to 8.

This is because our four-against-four test has three possible outcomes:

1. balanced, with probability $1/3$;
2. left heavier, with probability $1/3$;
3. left lighter, with probability $1/3$.

Likewise, our second measurement narrows the possibilities from 8 to either 2 or 3 (again, we divide by three, the number of possible outcomes).

Our third measurement is sufficient to fully identify the odd ball.

A good utility function for the outcomes of experiments is the Shannon information.

Also known simply as the information, this is defined as

$$h(x) = \log \frac{1}{P(x)}.$$

This can be used as the utility of obtaining outcome x from an experiment where our prior is $P(x)$: it rewards surprising outcomes.

Given our twelve balls, if we'd weighed one ball against one other, and found an (improbable) imbalance, this would be very informative.

The logarithm renders the information gained from independent experiments additive.

If we have two independent experiments, such that the distribution over their two outcomes is separable, the information will be simply the sum of that gained from the outcomes in each experiment:

$$
\begin{aligned}
h(x, y) &= \log \frac{1}{P(x, y)} \\
&= \log \frac{1}{P(x)\,P(y)} \\
&= \log \frac{1}{P(x)} + \log \frac{1}{P(y)} \\
&= h(x) + h(y).
\end{aligned}
$$

The expected utility is then the entropy.

$$
\begin{aligned}
H[X] &= \sum_i P(x_i) h(x_i) \\
&= \sum_i P(x_i) \log \frac{1}{P(x_i)}
\end{aligned}
$$

By convention, we take $0 \log {}^1/_0 = 0$ if $P(x) = 0$.

Note that the X in $H[X]$ is a random variable: the entropy is a function of an entire distribution (such as for an experiment!), not a particular realisation (or outcome).

The entropy is maximised when the distribution is as flat as possible: another name for entropy is uncertainty.

Recall that a probability distribution is conditional on some information: your entropy depends on what you know.

Let's compute the entropy of our distribution over the odd ball.

Let's first simplify the problem to identifying X, the index of the odd ball (we don't care if it's heavier or lighter).

Our current state of knowledge is $P(x) = {}^1/_{12}$ for $x \in [1, \ldots, 12]$. Hence

$$H[X] = \sum_{i=1}^{12} {}^1/_{12} \log 12 = \log 12 = 2.5 \text{ nats} = 3.6 \text{ bits},$$

where nats are derived from using $\log_e$ and bits use $\log_2$.

The joint entropy is exactly what you expect:

$$H[X, Y] = \sum_{i,j} P(x_i, y_j) \log \frac{1}{P(x_i, y_j)}.$$

Hence, as with information, entropy is additive for independent random variables, for which

$$\begin{aligned} H[X, Y] &= \sum_{i,j} P(x_i) P(y_j) \log \frac{1}{P(x_i) P(y_j)}. \\ &= \sum_{i,j} P(x_i) P(y_j) \log \frac{1}{P(x_i)} + \sum_{i,j} P(x_i) P(y_j) \log \frac{1}{P(y_j)} \\ &= \sum_{i} P(x_i) \log \frac{1}{P(x_i)} \sum_{j} P(y_j) + \sum_{j} P(y_j) \log \frac{1}{P(y_j)} \sum_{i} P(x_i) \\ &= H[X] + H[Y]. \end{aligned}$$

The conditional entropy is the uncertainty given something you know something.

$$H[X \mid Y = y] = \sum_i P(x_i \mid Y = y) \log \frac{1}{P(x_i \mid Y = y)}$$

is the entropy for X given you know that Y takes particular value y. In fact, given that all probabilities are conditional on something, so are all entropies of this form.

$$H[X \mid Y] = \sum_j P(y_j) \sum_i P(x_i \mid Y = y_j) \log \frac{1}{P(x_i \mid Y = y_j)}$$

is the expected entropy for X you will have after an experiment Y whose result you don't yet know.

We have the chain rule for entropies:

$$
\begin{aligned}
&H[X, Y] \\
&= \sum_{i,j} P(x_i \mid y_j)P(y_j) \log \frac{1}{P(x_i \mid y_j)P(y_j)}. \\
&= \sum_{i,j} P(x_i \mid y_j)P(y_j) \left(\log \frac{1}{P(x_i \mid y_j)} + \log \frac{1}{P(y_j)} \right) \\
&= \sum_{i,j} P(x_i \mid y_j)P(y_j) \log \frac{1}{P(x_i \mid y_j)} + \sum_{i,j} P(x_i \mid y_j)P(y_j) \log \frac{1}{P(y_j)} \\
&= \sum_{i,j} P(x_i \mid y_j)P(y_j) \log \frac{1}{P(x_i \mid y_j)} + \sum_{j} P(y_j) \log \frac{1}{P(y_j)} \sum_{i} P(x_i \mid y_j) \\
&= \sum_{i,j} P(x_i \mid y_j)P(y_j) \log \frac{1}{P(x_i \mid y_j)} + \sum_{j} P(y_j) \log \frac{1}{P(y_j)} \\
&= H[X \mid Y] + H[Y].
\end{aligned}
$$

Now let's compute the conditional entropy of the one-against-one strategy.

Here, as before, X will be the index of the odd ball, and Y will now be the outcome of the experiment.
For the one-against-one strategy, there are two possible outcomes:

1. $Y = 1$ is that the scales are unbalanced, with $P(Y = 1) = 1/6$, and
$$P(x \mid Y = 1) = \begin{cases} \frac{1}{2}, & \text{if } x \in [1, 2] \\ 0, & \text{otherwise;} \end{cases}$$
2. $Y = 2$ is that the scales are balanced, with $P(Y = 2) = 5/6$, and
$$P(x \mid Y = 2) = \begin{cases} \frac{1}{10}, & \text{if } x \in [3, 4, \ldots, 12] \\ 0, & \text{otherwise.} \end{cases}$$

1. $Y = 1$ is that the scales are unbalanced, with $P(Y = 1) = 1/6$, and

$$P(x \mid Y = 1) = \begin{cases} \frac{1}{2}, & \text{if } x \in [1, 2] \\ 0, & \text{otherwise;} \end{cases}$$

2. $Y = 2$ is that the scales are balanced, with $P(Y = 2) = 5/6$, and

$$P(x \mid Y = 2) = \begin{cases} \frac{1}{10}, & \text{if } x \in [3, 4, \ldots, 12] \\ 0, & \text{otherwise.} \end{cases}$$

$$\begin{aligned} H[X \mid Y] &= \sum_j P(y_j) \sum_i P(x_i \mid Y = y_j) \log \frac{1}{P(x_i \mid Y = y_j)} \\ &= \frac{1}{6} 2 \frac{1}{2} \log 2 + \frac{5}{6} 10 \frac{1}{10} \log 10 = 2.0 \text{ nats} \end{aligned}$$

The mutual information is the average reduction in the uncertainty about X after learning Y, and is defined as

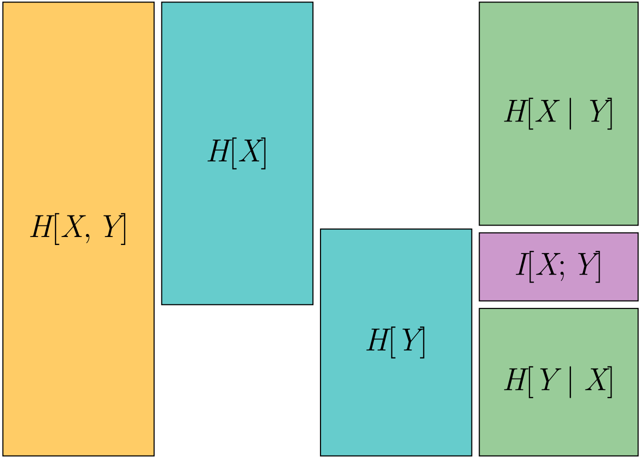
$$I[X;Y] = H[X] - H[X \mid Y] = H[Y] - H[Y \mid X] = I[Y;X].$$

Note the symmetry of the mutual information, and also that $I[X;Y] \geq 0$.

The mutual information of the one-against-one experiment is

$$I[X;Y] = H[X] - H[X \mid Y] = 2.5 \text{ nats} - 2.0 \text{ nats} = 0.5 \text{ nats}.$$

Using height to represent magnitude, we can plot the relationships between entropies.



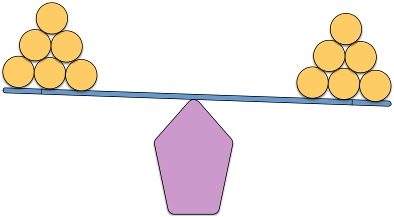
Plot inspired by Mackay (2003).

Entropy rewards experiments that have high expected informativeness.

Conditional entropy and mutual information are identically good utility functions for experiments Y designed to learn about X.

The one-against-one weighing strategy will, rarely, narrow down the odd ball to being one of two on the first trial.

The four-against-four strategy will never return such an informative outcome: however, its average informativeness is much higher.



The entropy of a continuous variable needs some care, as pdfs have units.

The differential entropy is defined as

$$H[X] = \int_{-\infty}^{\infty} p(x) \log \frac{1}{p(x)} \mathrm{d}x.$$

Note, however, that $p(x)$ has units equal to the inverse of x, and that the log has to take a dimensionless argument. This means that the differential entropy is poorly behaved: it is not the limit as $n \to \infty$ of the (discrete) entropy. Instead, we use the relative entropy (also known as the Kullback-Leibler distance) from some base density $\mu(x)$,

$$H[X \,||\, \mu] = \int_{-\infty}^{\infty} p(x) \log \frac{\mu(x)}{p(x)} \mathrm{d}x.$$

In summary,

1. Simpson's paradox can be resolved with proper treatment of conditional dependence (that is: draw a Bayes net!).
2. Model comparison uses the model evidence, $P(\mathcal{M} \mid \mathcal{D})$, to evaluate models.
3. The evidence naturally balances complexity and model fit.
4. Entropy is a natural utility in picking experiments: it measures uncertainty.