Bayesian probabilities: why and how

1. the Cox axioms
2. Bayesianism
3. terminology
4. frequency counts
5. frequentist probabilities
6. MAP with a prior
7. conjugate priors and pseudo counts
8. how the prior affects MAP predictions
9. fully Bayesian inference

‘A wise man proportions his belief to the evidence’

Since we never know the “real” state of things with absolute certainty, knowledge always comes down to *beliefs*: assertions about the real state of the world.
How should beliefs change in the light of evidence?

degrees of belief

Denote the degree of belief in[1] A being true, given that $\mathcal{I}$ is assumed true, by $Bel(A|\mathcal{I})$. Cox ~ 1950, posed three axioms (’MORE’,’OR’,’NOT’) about Bel, and showed they were isomorphic to probabilities.

axiom 1: MORE

(a.k.a. transitivity)

$$\begin{aligned} \text{if} \quad & Bel(A|\mathcal{I}) > Bel(B|\mathcal{I}) \\ \text{and} \quad & Bel(B|\mathcal{I}) > Bel(C|\mathcal{I}), \\ \text{then} \quad & Bel(A|\mathcal{I}) > Bel(C|\mathcal{I}) \end{aligned}$$

Implies plausibilities can be mapped onto real numbers. We can arbitrarily squash the reals into a given finite range, so map “totally positive it is true” to 1, and “absolutely ruled out” to 0.

[1] “plausibility of”, “preference for”, “confidence in”, “degree of belief in”...

Incredibly, just two more axioms are needed to calculate a numerical value on $Bel(A|\mathcal{I})$.

axiom 2: NOT

Our confidence in a proposition being true is directly related to our confidence that it’s false, and this *relationship* should be independent of the details of the actual proposition. That is:

$$Bel(\neg A|I) = F\left[Bel(A|I)\right]$$

axiom 3: OR

If there are two ways of arriving at the plausibility of some proposition, both these ways should arrive at the same result, provided they use the same background information. Specifically, the truth of some compound proposition AB (meaning "both A and B are true") can be found be establishing A first and then B (given A), *or* by finding B first and then A given B.

$$Bel(AB|I) = G\left[Bel(A|I),\ Bel(B|AI)\right]$$

Note this is invariant to interchanging A and B.

beliefs should be updated using the probability rules

These three axioms, together with the scaling choice of $Bel \in [0, 1]$, entirely determine how to calculate degrees of belief, for they imply that $F(x) = 1 - x$ and $G(x, y) = xy$. This leads to the *product rule*:

$$Bel(AB|I) = Bel(A|I)\,.\,Bel(B|A, I)$$

and the *sum rule*:

$$Bel(A|I) + Bel(\neg A|I) = 1$$

- if all uncertainty is removed, these reduce to the two basic rules of Boolean algebra for the negation and conjunction of propositions.

The language of inference is probability theory

Degrees of belief are isomorphic to (obey exactly the same rules as) probabilities. Thus all coherent beliefs and predictions can be mapped to probabilities, and the 2 laws of probability describe how to update these in the light of data.

Cox showed that Bayesian inference is the only inductive inference that is logically consistent. Any inference scheme *other* than using these two laws of probability is inconsistent with at least one of the Cox axioms.

> *"In engineering we respect the laws of physics or court disaster, and in mathematics we respect those of arithmetic: in inference we should respect the laws of probability" (Skilling)*

terminology

We know how to interpret $P(\mathcal{D}|\mathcal{H})$: that's the *likelihood* of the data, given that the hypothesis is true.

And Bayes theorem gives:

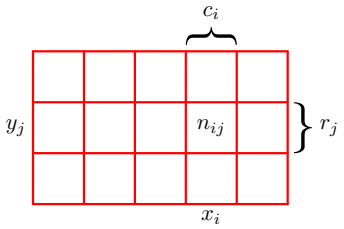
$$P(\mathcal{H}|\mathcal{D}) \quad = \quad \frac{P(\mathcal{D}|\mathcal{H})\ P(\mathcal{H})}{P(\mathcal{D})}$$

- $P(\mathcal{H}|\mathcal{D})$ is the **posterior**
- $P(\mathcal{D}|\mathcal{H})$ is the **likelihood**
- $P(\mathcal{H})$ is the **prior**
- $P(\mathcal{D})$ is "just normalisation": it's the numerator integrated over hypotheses $\mathcal{H}$. (Sometimes called the *evidence*).

That tells us how our belief in some hypothesis $\mathcal{H}$ should change in the light of data $\mathcal{D}$.

frequency counts

Suppose we have a data set of events, each of which consists of a pair of values, (x, y). We could form a table of *frequency counts* documenting how many events occured at certain values:



- total number of x_i events:

$$c_i = \sum_j n_{ij}$$

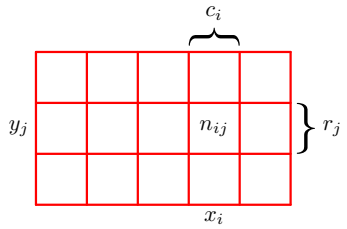
- total number of y_j events:

$$r_j = \sum_i n_{ij}$$

total number of events:

$$N = \sum_i \sum_j n_{ij} = \sum_i c_i = \sum_j r_j$$

trivially true



a sum:

$$\frac{c_i}{N} = \sum_j \frac{n_{ij}}{N}$$

"the proportion of (x_i) events is sum of the proportions of possible (x_i, y_j) events"

a product:

$$\frac{n_{ij}}{N} = \frac{c_i}{N}\frac{n_{ij}}{c_i}$$

"the proportion of (x_i, y_j) events is the proportion of (x_i) events times the proportion, *among those* (x_i), of (x_i, y_j) events"

probabilites as limiting frequencies

Provided they come from a consistent source, as more and more events build up these ratios will tend towards certain values, which we call *probabilities*.

$$\lim_{N\to\infty} \frac{n_{ij}}{N} = P(x_i, y_j)$$

$$\lim_{N\to\infty} \frac{c_i}{N} = P(x_i)$$

the sum rule

$$P(x_i) = \sum_j P(x_i, y_j)$$

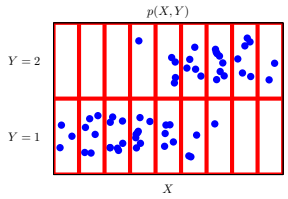
The probability of x_i is the sum of the probabilities of all possible (x_i, y_j) joint events

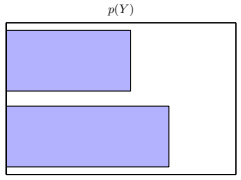
the product rule

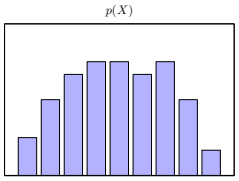
$$P(x_i, y_j) = P(x_i)\, P(y_j|x_i)$$

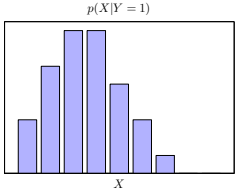
The probability of the joint event (x_i, y_j) is the probability of x_i times the probability of y_j given that x_i has occured

in pictures









Bayes theorem

Notice that $P(x, y)$ can be written in either of two ways: $P(x)\, P(y|x) = P(y)\, P(x|y)$ and so

$$P(y|x) = \frac{P(y)\, P(x|y)}{P(x)}$$

dependence and independence

If I throw a coin and then a die, the probability I get a head and a 6 is $P(\text{head}, six) = P(\text{head})\ P(six)$.

Two ways to say "Events x and y are independent":

- $P(x, y) = P(x)P(y)$
- $P(x|y) = P(x)$ (by the product rule)

Loosely, "if two events are independent, learning about one of them tells you nothing about the other".

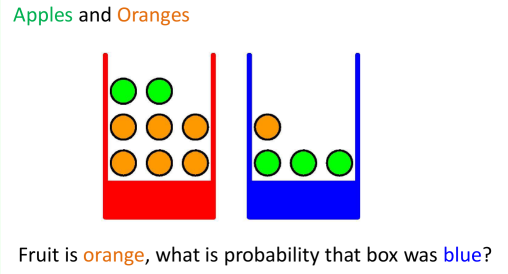
And if x and y are independent given z, then

- $P(x, y|z) = P(x|z)P(y|z)$
- $P(x|y, z) = P(x|z)$

exercise (from a Chris Bishop video lecture)

this is adapted from the video at
`http://videolectures.net/mlss09uk_bishop_ibi/`

In the dark, you grabbed a fruit from one of two boxes. The *prior* prob it was the blue box is $P(Box = blue) = 3/5$. The fruit was an orange. What do you think now?



Various dichotomies in machine learning

predicting	$\leftrightarrow$	acting
inference	$\leftrightarrow$	learning
optimize over	$\leftrightarrow$	integrate over
supervised	$\leftrightarrow$	unsupervised ($\leftrightarrow$ reinforcement)
exact calculation	$\leftrightarrow$	sample approximation
continuous	$\leftrightarrow$	discrete
marginal dependence	$\leftrightarrow$	conditional dependence

recap: Bayes

We know how to interpret $P(\mathcal{D}|\mathcal{H})$: that's the *likelihood* of the data, given that the hypothesis is true.

And Bayes theorem gives:

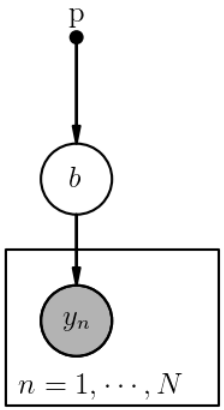
$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H})\ P(\mathcal{H})}{P(\mathcal{D})}$$

- $P(\mathcal{H}|\mathcal{D})$ is the **posterior**
- $P(\mathcal{D}|\mathcal{H})$ is the **likelihood**
- $P(\mathcal{H})$ is the **prior**
- $P(\mathcal{D})$ is "just normalisation": it's the numerator integrated over hypotheses $\mathcal{H}$. (Sometimes called the *evidence*).

That tells us how our belief in some hypothesis $\mathcal{H}$ should change in the light of data $\mathcal{D}$.

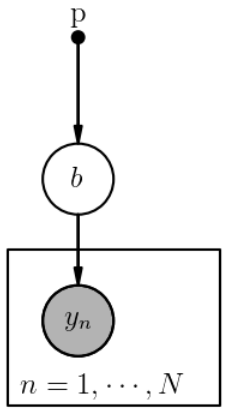
PGM: probabilistic graphical model

conventions



- "filled dots" (?!) are parameters we're prepared to assume fixed values for
- white nodes is unknown (we say "latent").
- shading indicates the value of a node is known to us ("observed").

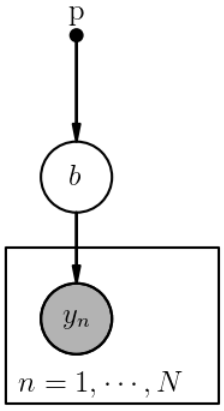
example: one coin, 2 possibilities



Eg: two coins,  versus  :

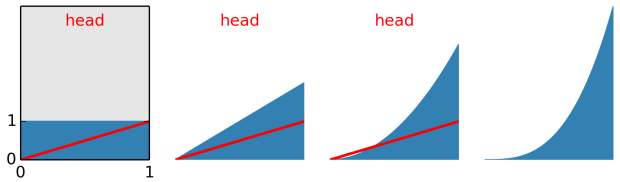
- $\mathbf{p}$ is the prior prob of choosing each coin: $\mathbf{p} = (\frac{1}{2}, \frac{1}{2})$
- b is unknown "bentness" of the coin: $b \in \{\frac{1}{2}, \frac{3}{4}\}$
- y_n is the outcome of the n^{th} toss: $y_n \in \{\mathtt{H}, \mathtt{T}\}$
- the box is shorthand, known as a "plate" – see next slide.

example: 1 coin, ∞ possibilities



Say I have a coin, which might be "bent" (biased). I (the world) start tossing it, and you (the agent) start trying to predict...

example continued

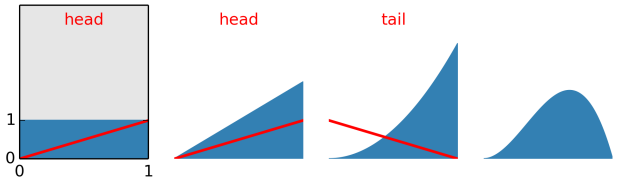


- Bayes theorem requires a prior distribution.
  (e.g. uniform distribution is the one with "maximum entropy")

Notice:

- the prior times the likelihood becomes the posterior
- that posterior is the "new prior", in effect

example



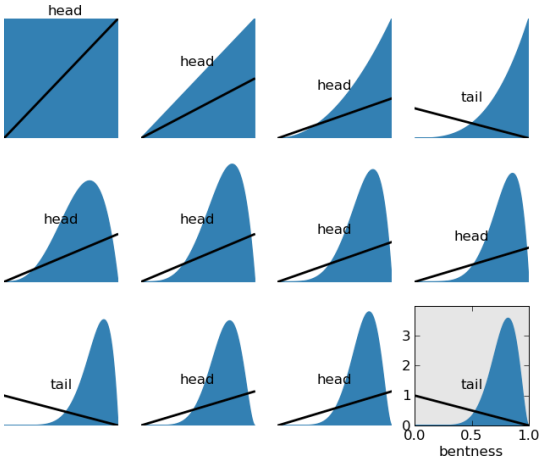
Note:

- we can think of the prior for the new data as being the posterior resulting from all previous data. In that case the likelihood is just b or $1-b$. OR...
- we can think of the prior being 1 (flat), and the likelihood being the likelihood of ALL the data to date: $b^{\#\text{ heads}}\,(1-b)^{\#\text{ tails}}$

These views are equivalent here - this only happens if trials are "i.i.d" (independent, identically distributed).

You can try it yourself with `bayes-coin.py` on the website.

and so on...



Posterior is the "Beta Distribution": $P(b|\mathcal{D}) = \frac{1}{Z} b^{\#\text{heads}}(1-b)^{\#\text{tails}}$

prior or posterior?

Consider 2 tosses of the coin:

- we can consider a prior, and find the posterior

$$P(b \mid \mathcal{D}_{1\ldots n}) = P(\mathcal{D}_{1\ldots n} \mid b)\;\; P(b) \qquad \text{(and normalise...)}$$

- OR we can think of iteratively ("yesterday's posterior is today's prior") and consider just the likelihood of the latest event, knowing that successive tosses are independent events:

$$P(b \mid \mathcal{D}_{1\ldots n}) = P(\mathcal{D}_n \mid b)\;\; P(b \mid \mathcal{D}_{1\ldots n-1}) \quad \text{(and normalise...)}$$

If you're dealing with just a "bag" of data that has no natural ordering to it, the first way is best.

If you're dealing with sequential data, you're more likely to do the second.

recap: Probabilistic Graphical Models

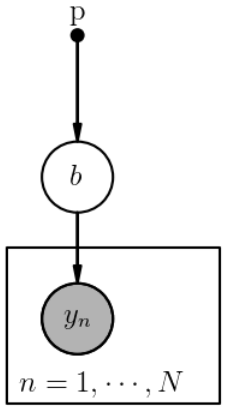
The joint distribution

refers to the distribution over *all variables of interest*, in this case the unknown b and the outcome y: the joint is $P(b, y)$.

One *factorisation* of this is

$$P(b)P(y|b)$$

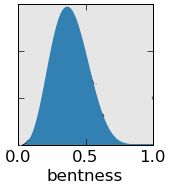
Probabilistic Graphical Model (of that factorisation)



3 ways of predicting

- the ***maximum likelihood*** approach is to make predictions using the parameter value most likely to have generated the data:

$$P(x^{\text{new}}|\mathcal{D}) = P(x^{\text{new}}|b^{\text{ML}})$$
$$b^{\text{ML}} = \underset{b}{\operatorname{argmax}} P(\mathcal{D}|b)$$



- the ***maximum a posteriori*** (MAP) approach is use the most plausible parameter value ("the one you believe in the most"):

$$P(x^{\text{new}}|\mathcal{D}) = P(x^{\text{new}}|b^{\text{MAP}}) \quad \text{where } b^{\text{MAP}} = \underset{b}{\operatorname{argmax}} P(b|\mathcal{D})$$

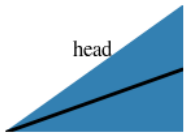
- the ***Bayesian*** approach is *integrate out* the unknown:

$$P(x^{\text{new}}|\mathcal{D}) = \int P(x^{\text{new}}, b|\mathcal{D})\, db$$

max likelihood and over-fitting

After you toss one HEAD, likelihood $P(\mathcal{D}|b) = b$.

- $b^{\text{ML}} = 1$, so the prediction is 100% that a HEAD will follow... This is the simplest possible case of *overfitting*.

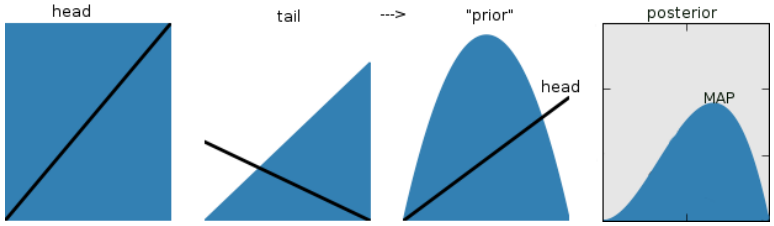


NOTE: MAP seems "more Bayesian", so is it better?

- The maximum entropy prior (most non-committal) is flat.
- With this, posterior $\propto$ likelihood and so MAP inference is going to be just as bad!
- Maybe we should we choose a "better" prior then?

MAP inference with a simple prior

- We could, for example, pretend we've seen one head and one tail, and treat the resulting posterior as a prior.
- Now the MAP point is more sensible. *Ex: what is that point?*
- A solution for overfitting?



conjugate priors

Priors typically have some functional form. If, after multiplication by the likelihood, they leave a posterior distribution of the *same* form (same family), we say they're the "conjugate prior" for that likelihood.

If n HEADs and m TAILs are thrown, the likelihood is

$$P(\mathcal{D}|b) \;\propto\; b^n(1-b)^m$$

Suppose the prior is a Beta distribution:

$$P(b) = \frac{1}{Z} b^u (1-b)^v$$

where Z is normalisation. u,v are called hyperparameters[2]. The posterior will be

$$P(b|\mathcal{D}) = \frac{1}{Z'} b^{u+n} (1-b)^{v+m}$$

(we say the Beta distribution is the conjugate prior for the Binomial distribution)

The u and v hyperparameters are *just like events*, except they never happened... Often called "pseudo counts".
$u = v = 0$ is equivalent to having a flat prior.

[2] *ie.* parameters of a prior over parameter b

MAP prediction for the coin

Using pseudocounts of $a = b = 1$, and tossing $\mathcal{D} = \{1 \text{ HEAD}\}$

$$\begin{aligned} P(b|\mathcal{D}) &= b^2(1-b)^1 \ / Z \qquad \text{and taking logs.....} \\ \log P(b|\mathcal{D}) &= 2\log b + \log(1-b) \ - \log Z \end{aligned}$$

The gradient w.r.t. b of that is:

$$\frac{d}{db} \log P(b|\mathcal{D}) = \frac{2}{b} + \frac{1}{1-b}$$

This function has one value of b where its slope is zero, and has negative curvature everywhere (its second derivative $\frac{d^2}{db^2} < 0$ everywhere), so this has to be the maximum:

$$b^{\text{MAP}} = \frac{2}{3}$$

And so our prediction is $P(\text{HEAD}|b^{\text{MAP}}) = \frac{2}{3}$: crisis averted?

priors and regularisation

- You've met the idea of complexity control by regularisation or "smoothing".
- The prior, in the MAP approach, seems to be playing a similar role.

But first, let's get all fundamentalist for a moment...

maximization is for decisions, not for inference

- the *maximum likelihood* approach is to make predictions using the parameter value you believe in the most:

$$b^{\text{ML}} = \text{argmax}_b P(\mathcal{D}|b)$$

- the *maximum a posteriori* (MAP) approach is use the most plausible parameter value (the one you believe in the most):

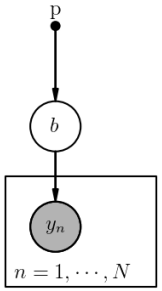
$$b^{\text{MAP}} = \text{argmax}_b P(b|\mathcal{D})$$

- **But note both involve maximizing something, which isn't mentioned in the probability calculus and is thus Against The (Cox Axiom) Rules!**

Bayesian prediction

Let's try to be Good Bayesians. After one toss, $\mathcal{D} = \{\text{HEAD}\}$, and starting from a flat prior:

$$P(y^{\text{new}} = \text{HEAD}|\mathcal{D}) = \int_0^1 db \;\; P(y^{\text{new}} = \text{HEAD}, b|\mathcal{D})$$



$$\begin{aligned}
&= \int_0^1 P(y^{\text{new}} = \text{HEAD}|b, \mathcal{D}) P(b|\mathcal{D}) \;\; db \\
&= \int_0^1 P(y^{\text{new}} = \text{HEAD}|b) P(b|\mathcal{D}) \;\; db \\
&= \int_0^1 2b^2 \;\; db \\
&= \left[\frac{2b^3}{3}\right]_0^1 = \frac{2}{3}
\end{aligned}$$

Integrating over the unknown does not overfit, even with a flat prior!

Optimization loses information about the posterior

Picking one "best" explanation from many possible ones is a computational shortcut with obvious computational benefits.

But maximizing something means throwing all the other options away, and this can be a bad thing.

- "loss of doubt" $\rightarrow$ false certainty $\rightarrow$ over-confidence (as in over-fitting)
- Beliefs so obtained may become inconsistent with the Cox Axioms.
- We didn't discuss this, but apparently using such beliefs in betting against a Bayesian will lose you money (*eg.* google "de Finetti, coherence").
- Worse, the belief distribution might not be unimodal (although it's bound to be for the coin example). Without care, optimization throws away all knowledge of these other modes.

---

the bottom line : optimization is for decisions, not for inference

so why do we do it so much?

- **historically**, that's how machine learning came about (eg. neural networks, "backpropagation", coupled the dominance of frequentist thinking in stats in the 20th Century). This isn't a good reason to favour optimization.
- humanimals like their world-descriptors succinct, pithy. We like summaries that capture what we think are the most important aspects.
- exact Bayesian integration is **often intractable**. Sampling methods like MCMC help, but are "unsatisfying" to many. Intractability is a good reason to optimize, provided you're aware of the pitfalls.

Good advice from Tony Vignaux (emeritus prof in OPRE here): "Avoid optimizing things prematurely".

practical consequences

- if you have oodles of data, it hardly matters what prior you had: the effect of the likelihood will be to swamp it.
- you can solve the over-fitting problem by using MAP inference with a "sensible" prior.
- For the coin example, big values for pseudo counts u and v amount to strong assumptions: these ameliorate the undesirable effects of using the MAP solution for prediction.
- The fully Bayesian solution won't overfit even with a loose prior. BUT that integral can get to be a pain in many interesting cases.

This is discussed, for example, in *Barber*, Section 8.6. He uses the Gaussian as a worked example instead of my bent coin. It's also in *Bishop*, Section 2.1 (and 2.2).