# Bayesian learning

# Strange question

(Example token from Ian Hacking, "Introduzione alla probabilità e alla logica induttiva", Il Saggiatore)

You are on the jury in a trial.

Two taxi companies operate in your city:

- Green Cabs Ltd. whose cars are green
- Blue Taxi Inc. whose taxis are blue

Green Cabs rules the market, which a share of 85%.

In a foggy winter evening a taxi pushed a car out off the road then run away. A witness says the taxi was blue.

The witness is put to the test in similar conditions and manages to tell the right colour of a car in the 80% of cases (both for green and blue taxis).

It is more likely that the taxi was green or blue?

A useful way to cope with such problems is to reason about frequency instead of probability. We assume that frequency (past observed percentage of cases of a certain type)  is a good estimate for probability (future expected percentage). This assumption is not so innocent, yet we take it for granted.

Imagine the witness observe 100 taxis extracted in a random way from the population of taxis in your city. We expect to get 85 green taxis and 15 blue taxis. Here "we expect" means we imagine to repeat the observation a great number of times and take for granted that the percentage of green taxis in our sample will approximate asymptotically the limit 0.85. We use this limit.

The witness classifies 68 green taxis as green (rightly) and 17 as blue (wrongly), because she has a success percentage of 80%.

She classifies 12 blue taxis as blue (rightly) and 3 as green (wrongly), for the same reason.

|  | Classified colour |  |  |  |
|---|---|---|---|---|
| True colour |  | Green | Blue |  |
|  | Green | 68 | 17 | 85 |
|  | Blue | 3 | 12 | 15 |
|  |  | 71 | 29 | 100 |

The witness sees a blue taxi.
The real taxi is one out of these 29.
12 of them are blue, 17 are green.
The more likely true colour is green.
Probabilities are 17 vs. 12 in favour of green.

# Revising probability estimates

In several occasions we saw how new information enables us to improve our estimate of event probability, which in turn improves the expected value of a decision.

These improvements represent the value of information, which has a cost, too.

Bayes’ rule is a major tool for forecasting: it enables us to modify our probability estimate in a rational way.

First of all, we have to clarify what is a probability estimate.

In the Bayesian view, probability is not considered as a property of reality. We are not saying it is not: we only assume another starting point to look at probability.

“Bayesian” probability is a subjective matter. When I say “probability of tomorrow demand for sights on the fly being over 20 is 25%”, I am really saying what is my confidence in that event.

To be coherent, I have to use this number 25% systematically in my event forecasting, utility evaluating and decision making.

For example, pretend someone challenges me for a bet on that event. He/she bids on tomorrow demand being less or equal to 20. Given my claim of probability of that event being 75%, my “indifference point” is 0,33 dollars: I am available to bid 0,33 cents versus 1 dollar on demand over 20 sights.

Indeed, the odds of this bet is 1/3 from my point of view.

Bayesian view of probability is about drawing rational consequences from probability estimates *modifying them when new information is available*.

In this sense, it is a learning method.

We have an hypothesis *H* and a data item *E* (named also experience or evidence). The Bayes' rule is

$$P(H \mid E) = \frac{P(E|H)P(H)}{P(E)}$$

Once we have estimated three probabilities on the right side, we apply the rule and get an estimate for the left side.
How to make estimates for the left side is not implicit in the rule. It is a tool for modifying previous estimates, not for making the primitive ones.
We use some properties of conditional probability, like

$$P(E \wedge H) = P(E \mid H)P(H)$$

$$P(E) = P(E \mid H)P(H) + P(E \mid H^c)P(H^c)$$

where $H^c$ is the complement (negation) of *H*.
Often it is useful to formulate the rule in this way:

$$P(H \mid E) = \frac{P(E|H)P(H)}{P(E)} = \frac{P(E|H)P(H)}{P(E|H)P(H)+P(E|H^c)P(H^c)}$$

# Example: tossing a fixed coin

$$P(H \mid E) = \frac{P(E|H)P(H)}{P(E)}$$

$$P(E \wedge H) = P(E \mid H)P(H)$$

$$P(E) = P(E \mid H)P(H) + P(E \mid H^c)P(H^c)$$

We have 10 coins, 9 fair and one fixed, with two heads.
We get a coin random and, without looking at it, we toss it 6 times.
We get 6 heads.
What is the probability the tossed coin is the fixed one?

Let *H* be the hypothesis "the tossed coin is the fixed one" and *E* the evidence "we got 6 heads".
Initially, lacking any information, that is not having *E* available, we estimate $P(H) = 1/10$ and $P(H^c) = 9/10$.
It holds that $P(E|H) = 1$, because the fixed coin always gives head, while $P(E|H^c) = 1/2^6$.

Substituting those values we find the value of the *posterior* *P(H|E)*:

$$P(H \mid E) = \frac{P(E|H) \bullet P(H)}{P(E|H) \bullet P(H) + P(E|H^c) P(H^c)} = \frac{64}{73} \cong 0{,}89$$

The alternative hypothesis $H^c$, fair coin, has probability 9/73. We started with an estimate 0.1 for the *prior P(H)*, then we arrived to an estimate 0.89. We used two numbers derived from experience: the *marginal P(E)* and the *likelihood P(E|H)*. We modified our posterior estimate with a learning process.

Let us try an intuitive interpretation of the Bayes’ rule:

$$P(H \mid E) = \frac{P(E|H)P(H)}{P(E)}$$

The prior *P(H)* is our confidence in the hypothesis H before we gain new information.
Maybe this confidence was made simply by examining frequency, like in the fixed coin example. We know there are 1 fixed coin out of 10, so we intuitively state *P(H)* = 0,1
Maybe our prior is not based on empirical frequency, but on subjective intuition.
Maybe it is simply a random guess.
We are not concerned here in *how and why* the prior was chosen: we get it as primitive data.

$$P(H \mid E) = \frac{P(E|H)P(H)}{P(E)}$$

Then we observe some evidence, coming to us spontaneously or because we made an experiment. In any case, we have gained new data.

Now we transform our probability estimation.

The likelihood *P(E|H)* is the probability of observing *E in a world where our hypothesis is true*. Note that *E* is just what we really observed in our world (where we do not know whether *H* is true or false).

The likelihood is divided by *P(E)* i.e. the probability of observing E in the collection of all possible worlds, both with *H* true or with *H* false.

This ratio is less than 1 if observing *E* when *H* is true is less probable than in a generic world. This means *H* hinders the happening of *E*.

In this case, it is rational to decrease confidence in *H*. The lower the ratio, the stronger the decrement in confidence.

$$P(H \mid E) = \frac{P(E|H)P(H)}{P(E)}$$

The ratio is greater than 1 if observing *E* is more probable when *H* is true. This means *H* favors *E*. So, we increase our confidence in *H*.
The ratio likelihood / marginal is 1 when probability of observing *E* is left unchanged by *H*. Really, *H* and *E* are independent events.
You can reason in terms of "how much *H* favors/hinders *E*" or in terms of "how much knowledge of *H* being happened makes us more confident in *H*".

The ratio likelihood / marginal can be interpreted as a learning factor, which multiplies our prior giving us a posterior.

This learning process is recursive: we start with an initial prior, then at each step we apply the rule and get a posterior, which is the prior for the next step.

# Example: forecasting the winner

$$P(H \mid E) = \frac{P(E|H)P(H)}{P(E)}$$

We are forecasting the winner of the next match between teams $T_0$ and $T_1$.
Evidence is the history of their previous matches:

- $T_0$ won 65% times, $T_1$ 35%
- 30% of $T_0$ ‘s wins were playing at $T_1$’s home
- 75% of $T_1$ are at home

The next match will be at $T_1$’s home.
We are going to estimate probability of winning for the teams.

We use two random variables *W* (winner) and *H* (host, playing at home). They can get values 0 and 1, the names of the teams.

Historical data gives us frequencies we can use to set up the priors:

- P(W = 0) = 0.65
- P(W = 1) = 0.35
- P(H = 1 | W = 0) = 0.30
- P(H = 1 | W = 1) = 0.75

We compare P(W = 1 | H = 1) and P(W = 0 | H = 1).

$$P(W=1 \mid H=1) = \frac{P(H=1 \mid W=1)P(W=1)}{P(H=1)} =$$

$$= \frac{P(H=1 \mid W=1)P(W=1)}{P(H=1 \wedge W=1) + P(H=1 \wedge W=0)} = \frac{P(H=1 \mid W=1)P(W=1)}{P(H=1 \mid W=1)P(W=1) + P(H=1 \mid W=0)P(W=0)} =$$

$$= \frac{0.75 \bullet 0.35}{0.75 \bullet 0.35 + 0.30 \bullet 65} = 0.5738$$

Of course, a similar computation gives P(V = 0 | C = 1) = 0.4262.

# Example: clinical tests

A test for a certain illness can give outcome *Pos* or *Neg*.

If a patient is really ill, the test answers

- *Pos* in 98% cases (true positive)
- *Neg* in 2% cases (false negative)

If a patient is not ill, the test answers

- *Neg* in 97% cases (true negative)
- *Pos* in 3% cases (false positive)

Let us summarize:

| | |
|---|---|
| $prob(Mal) = 0{,}008$ | $prob(non\ Mal) = 0{,}992$ |
| $prob(Pos \mid Mal) = 0{,}98$ | $prob(Neg \mid Mal) = 0{,}02$ |
| $prob(Pos \mid non\ Mal) = 0{,}03$ | $prob(Neg \mid non\ Mal) = 0{,}97$ |

If I do the test and the outcome is *Pos*, what is the probability I am ill?

The answer is neither 98% (true positive) nor 3% (false positive).

None of the previous numbers is the answer in itself.

Indeed, we have to take into account probabilities of an event that did not happen: outcome negative. But first of all, we need a prior: the probability to be ill before doing the test.

Missing data is just the prior. Once it is known, we are able to correctly estimate the probability of being ill.

Let 0.8% the percentage of ill people in the whole population, independently on any evidence from the test.

Out of 10.000 people, 80 are ill and 9920 sane: P(Maladie) = 0,008.

Out of 9920 sane people, those getting a positive outcome are
9920 * 3% = 298.

Out of 80 ill people, those getting a positive outcome are 80 * 98% = 78.

Therefore, in the whole population of 10.000 people, the outcome will be *Pos*
For 298 + 78 = 376 people.

But only 78 are really ill. Then the probability of being ill once the test is positive is 78 / 376 = 21%. Less than intuition.

The key point is that ill people are very rare in the population, *independently on test outcomes.*

# Method of maximum a posteriori

$$P(H_k \mid E) = P(H_k)\frac{P(E|Hk)}{\sum_{i=1}^{N} P(H_i)P(E|H_i)}$$

In this form of Bayes' Rule, H$_1$, …, H$_N$ are hypotheses composing a *partition of the event space*, i.e. one and only one of them is true.
Once we have observed *E*, we can select an hypothesis to bid on simply computing the posterior for each of them and choosing which one has the maximum posterior.

$$\forall H_i : P(H_i \mid E) = \frac{P(E \mid H_i)P(H_i)}{P(E)}$$

$$\hat{H} = \arg\max_{H_i} P(H_i \mid E)$$