

Some minimal background to Bayes, Noisy-channel and the like for
Bever & Poeppel (2010)

The key idea of Bayesian statistics is to calculate certain hard-to-come-by (unavailable) conditional probabilities on the basis of other (available) probabilities.

A probability is a number between 0 and 1 that is assigned to some (possibly complex) event – often, probabilities are thought of as **relative frequency**, as when we say that there is a 1 in 6 chance that you throw a specific number, say 3, with a dice. The event in this case is “throwing a 3” and its probability is 1/6.

In Bayesian statistics, an alternative interpretation of probability is used: the probability assigned to some event does not represent the relative frequency of its occurrence; rather, it expresses a **degree of confidence we have with regard to the event's occurrence**. This interpretation fits somewhat nicely with the use made of probabilities in current cognitive science (e.g. the paper at hand), but let's get down to the real stuff, i.e. the math.

Conditional probabilities express the probability of an event, given that another event has definitely occurred. While $P(\text{“throwing a 3”})$ is 1/6, the probability of throwing a 3, given that one throws a number smaller than 4, is 1/3 which is written as $P(\text{“throwing a 3”} | \text{“throwing a number } < 4\text{”}) = 1/3$. Let's make the example a little bit more relevant to our discussion and consider the problem of understanding speech.

We have two persons **P1** and **P2**, a **message** and a **medium**. The message is some '**mental content**' **M** and the **medium is an acoustic signal AS**. The problem P2 faces is to extract **M** from **AS** – as we all know, not a trivial task. One possible way to think about this in probabilistic terms is to say that P2 is trying to find the maximally likely message **M** given **AS**, that is he tries to solve

$$1. \quad \underset{M}{\operatorname{argmax}} P(M|AS)$$

This might look intimidating at first but actually it is quite intuitive. We have an unknown message we wish to recover, a known signal which we perceive and we want to find the message that, in some sense, constitutes the best match for the given signal. (1) is the same put in mathematical and probabilistic notation. The problem is that (1) does not help us at all, for we don't know how to calculate the probability of a message given a signal – $P(M|AS)$ is a very hard-to-come-by probability, and in a way, finding it constitutes the problem we are trying to solve. Luckily, we can use Bayes' rule to turn (1) into something more tractable. Bayes' rule says that

$$2. \quad P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

and that therefore (1) is mathematically equivalent to

$$3. \quad \underset{M}{\operatorname{argmax}} P(AS|M) \times P(M) \quad 1$$

1 We can drop the denominator $P(AS)$ for the acoustic signal is given and does not vary with the different Ms.

The interesting and useful thing about this is that $P(AS|M)$ can be interpreted as **“The probability that a person who wanted to express M produced AS”**², and this is a probability that can be computed more easily, especially for a speaker of the language in question.

Why so? Well, he can simply run his own 'generative process' with different Ms and see in which case the result matches the perceived AS best. It is, of course, of central importance that he uses the same 'generative process' the sender has used – in the case of languages a justified assumption. This is the general and core idea of so-called noisy-channel models where we assume that we cannot directly observe the true message because it has been sent by some channel that distorts it in some systematic way. If we know the exact nature of the noisy-channel, we can 'decode' the true message indirectly by sending our possible guesses through our own copy and picking the one which creates the result that fits the observed signal best. I know, I am starting to repeat myself but this is one of the most important ideas in current natural language processing and computer science and, as we see from the paper at hand, cognitive science.

The second probability, $P(M)$, is intended to account for the fact that not all messages are overall equally likely. Without going into the details, assume that you observe the signal 'ay s kr I m' – the probabilities $P('ay s kr I m' | \text{Ice-cream})$ and $P('ay s kr I m' | \text{I scream})$ are likely to be (near to) equivalent; in such a case, however, you can still decide that one of the two matches the input better for the context might be such that $P(\text{I scream})$ is greater (smaller) than $P(\text{Ice-cream})$.

2 Whereas $P(M|AS)$ is the probability that a person which pronounced AS wanted to utter M”