Contents

(All variables are real and one-dimensional unless otherwise specified.)

1. Bayes' theorem

The conditional probability of an event B on another event A is **empirically** defined as

$$P(B \mid A) \overset{\text{def.}}{=} \frac{P(A \cap B)}{P(A)} \,.$$

Although in **science** A is clearly identified as the independent variable and B as the dependent variable because of **causality**, no reasons in **mathematics** would disallow A and B from being symmetric in the formula, so its converse must equally hold:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \,.$$

Bayes' theorem links up the **empirical definition** and the **logical argument** and results in a quite nontrivial statement: regardless of any causality,

$$P(B \mid A) \equiv \frac{P(A \mid B)\, P(B)}{P(A)} \,.$$

(About the theorem's name: although modern English suggests its name be "Bayes's theorem", i.e. with an extra "s" after the apostrophe, Thomas Bayes is an eighteenth-century Englishman, and the contemporary English language chose to favour a spelling without the "s", which has thus been followed today.)

1.1 Bayesian inference

We often want to know the how likely a hypothesis H is true after seeing some evidence E. We hope to, in some way, measure the value of $P(H \mid E)$, which is equal to $P(H \mid E) = \frac{P(E \mid H)\, P(H)}{P(E)}$ according to Bayes' theorem. We may rewrite the equation by expanding $P(E)$:

$$\begin{aligned} P(E) &= \sum_{H'} P(E \cap H') \\ &= \sum_{H'} P(E \mid H') P(H'), \end{aligned}$$

where $\{H'\}$ represents a complete set of mutually exclusive hypotheses so that $\sum_{H'} P(H') = 1$. This expansion is valid because of a simple counting argument: each time when E occurs, some hypothesis has H' to be true; hence, if we sum up the co-occurrences of E and H' over all possible H', we essentially get the number of E's occurrences.

After the expansion, the equation becomes

$$P(H \mid E) = \frac{P(E \mid H)\, P(H)}{\sum_{H'} P(E \mid H') P(H')},$$

and it builds the foundation of **Bayesian inference**.

- $P(H)$ is called the **prior** (probability) of H and describes our **original degree of belief** in H.
- $P(H \mid E)$ is called the **posterior** (probability) of H and describes our **updated** degree of belief to after seeing the evidence E.
- $P(E \mid H)$ is called the **likelihood** of E given H.
- $P(E) = \sum_{H'} P(E \mid H') P(H')$ is called the **evidence** or the **marginal likelihood**.
- $P(E \mid H)/P(E)$ is called the **Bayes factor**. If it is greater than one, the posterior is more trustworthy than the prior, so we should change our belief.

1.2 Example: a drug test

A typical example of Bayesian inference is drug tests in sport competitions: how probably has an athlete used drugs if his urine test affirms so?

Solution. Let H be "the athlete has used drugs" and E be "his urine test is positive". Since $\{H, \bar{H}\}$ is a complete set of mutually exclusive events (as the athlete has definitely either used or not used drugs), we conclude that

$$P(H \mid E) = \frac{P(E \mid H)\, P(H)}{P(E \mid H)\, P(H) + P\big(E \mid \bar{H}\big)\, P\big(\bar{H}\big)},$$

where $P\big(\bar{H}\big) = 1 - P(H)$. The value of $P(H)$ may be predetermined in **any sensible way**. For example, it **may** be the past proportion of drug-using athletes. What if we do not know this proportion? We may appeal to the **principle of indifference** and simply set $P(H) = 0.5$, which assumes the hypotheses are equally probable, so that we are as fair as possible. Different choices of prior of course lead to different values of posterior, but as long as your evidence supports your hypothesis, your posterior belief will be greater than your assigned prior.

In fact, the **subjectivity** coming from the choice of prior has led to much **criticism** against Bayesian inference (not against Bayes' theorem, though) from another school of statistics called **frequentist inference**, which would simply answer "no conclusion" if the past proportion is unknown.

1.3 A mind-blowing example: the raven paradox

Are all ravens black? The more black ravens I see, the more I believe in "all ravens are black" although I cannot check the colour of every raven.

However, "all ravens are black" is logically equivalent to "anything that is not black is not a raven", so every time I see a non-black non-raven like a red apple, a blue cow, or a green moon, I should also become more inclined to say "all ravens are black"... Right?

Solution? The raven paradox is a notorious problem in philosophy as it ruthlessly reveals the problems of **inductive logic**, on which science heavily relies. While there is no universally accpeted answer to the paradox, some propose solving it with Bayes' theorem. Let H be "all ravens are black" and E be, for example, "I see a blue cow in a zoo".

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

There are then two cases.

- If there are no ravens in the zoo, I am not affected by whether all ravens are black or not at all, so $P(E \mid H) = P(E)$ and thus $P(H \mid E) = P(H)$. Seeing a blue cow in a zoo without ravens does not tell me anything about the colour of ravens.
- On the contrary, $P(E \mid H) > P(E)$ if there are ravens in the zoo. Informally speaking, let there be N objects in the zoo, then

$P(E) = 1/N$ but $P(E \mid H) = 1/(N-1)$ due to the disappearance of non-black ravens. As a result, $P(H \mid E) > P(H)$: seeing a blue cow in a zoo with ravens, which may make me see a blue raven instead, indeed makes me belive in "all ravens are black" more.

While the paradox may have been solved, this solution implies that how data is gathered can be just as crucial as the data itself.

2. Bayes classifier

The **Bayes classifier** is a slightly sophisticated application of Bayes' theorem.

Suppose that we have a **data matrix** $\mathbf{X} = (X_{ij})$, in which a row and a column respectively record the information about a **sample** and a **feature**. We also have a **class vector** $\vec{C} = (C_i)$, that records the **class** of each sample (and thus **classifies** them). A Bayes classifier aims to correctly guess the class of a new sample $\vec{x} = (x_j)$.

Class must be a **discrete** variable such as a person's sex, a person's state of illness (either sick or healthy), or a flower's species. If there are k classes, we may translate the classes into numbers so that we have a class variable $C \in \{1, 2, \ldots, k\}$. Then we write down

$$P(C_{\vec{x}} = a \mid \vec{x}) = \frac{P(\vec{x} \mid C_{\vec{x}} = a) P(C_{\vec{x}} = a)}{\sum_{b=1}^{k} P(\vec{x} \mid C_{\vec{x}} = b) P(C_{\vec{x}} = b)}$$

and **naively** predict that

$$C_{\vec{x}} = \operatorname*{argmax}_{a \in \{1, \ldots, k\}} P(C_{\vec{x}} = a \mid \vec{x}) \, .$$

Therefore, this algorithm is sometimes called a **naive Bayes classifier** to distinguish it with other more advanced Bayesian methods. Alternatively, this way of prediction is also called the **maximum-a-posteriori** (MAP) rule.

Since the denominator (i.e. the evidence) is common to all classes, it does not affect our decision and can be ignored. Now we proceed to the numerator, in which the prior $P(C_{\vec{x}} = a)$ may be defined as the proportion of a's in $\vec{C}$. Then for the likelihood $P(\vec{x} \mid C_{\vec{x}} = a)$, we assume that the jth feature $X_{j|C=a}$ of all samples from class a follows some distribution $f_{X_{j|C=a}}$; consequently, we can reformulate the likelihood as

$$P(\vec{x} \mid C_{\vec{x}} = a) = \prod_j f_{X_{j|C=a}}(x_j)\,.$$

Our remaining task is to estimate $f_{X_{j|C=a}}$ for each feature.

This method works as long as the features are **independent**, otherwise the product rule fails when we reformulate the likelihood. In reality, however, we rarely get completely independent features, so we often ignore this constraint and pretend that our features are independent. Sometimes this works, but sometimes this fails badly. (Big data is sometimes described to come with the "curse of dimensionality" because although there are a lot of data, much of them are highly correlated and thus bad for modelling.)

2.1 Example: spam detection

An email E is formally a set of words $\{w\}$. It is classified either as spam ($S = 1$) or not as spam ($S = 0$) according to its words. Some words such as "link", "prize", "warning", and "hacked" make the email more suspicious. Suppose that we have known $P(w \mid S = 0)$ and $P(w \mid S = 1)$ for all words in the email; how can we predict if it is spam with a Bayes classifier?

Solution. Assuming the words occur independently,

$$\begin{aligned} P(S=k \mid E) \sim & P(S=k) P(E \mid S=k) \\ = & P(S=k) \prod_i P(w_i \mid S=k) \end{aligned}$$

for $k \in \{0,1\}$. Then we compare the two posteriors and predict that E is spam if $P(S=1 \mid E) > P(S=0 \mid E)$ and not spam otherwise. Still, we must note that our assumption is unfortunately wrong because, for example, adjectives appear before nouns much more often than before verbs.

3. Decision theory

Decision theory provides a general framework for deciding among hypotheses in a delicate way. It does not merely consider probability of hypotheses but also the **cost** of believing them: c_{ij} amounts to the cost of believing a hypothesis H_i when a hypothesis H_j is true.

Given some evidence E, a hypothesis H is true with a probability

$$P(H \mid E) \sim P(E \mid H) \, P\left(H\right).$$

For simplicity, we assume our evidence $E = Z \in \mathbb{R}$ is a one-dimensional real signal, so

$$P(H \mid Z) \sim P\left(H\right) f_Z(z \mid H) \mathrm{d}z.$$

There are in general k hypotheses $\{H\}$. Decision theory does not ask us to compare their posteriors $\{P(H \mid Z)\}$ like the Bayes classifier; instead, it asks us to partition the real number line into k regions $\{R\}$ so that we believe the ith hypothesis H_i when the evidence Z falls into the ith region R_i.

Now let us consider a decision between only two complementary hypotheses $\{H_0, H_1\}$, which are labelled to satisfy $\langle Z \mid H_0\rangle < \langle Z \mid H_1\rangle$ for $\langle Z \mid H\rangle \equiv \int z f_Z(z \mid H)\mathrm{d}z$, i.e. H_0 is expected to produce a smaller signal than H_1. Rewriting their priors as $P(H_0) = p$ and $P(H_1) = q \equiv 1 - p$, the expected cost of our decision becomes

$$\begin{aligned} C =& c_{00} p \int_{R_0} f_Z(z \mid H_0)\mathrm{d}z + c_{01} q \int_{R_0} f_Z(z \mid H_1)\mathrm{d}z \\ &+ c_{10} p \int_{R_1} f_Z(z \mid H_0)\mathrm{d}z + c_{11} q \int_{R_1} f_Z(z \mid H_1)\mathrm{d}z, \end{aligned}$$

and our goal is to find out the regions $\{R_0, R_1\}$ that **minimizes the cost**.

In the simplest situation, the real number line is best bisected with a **single cutoff** Z^* to yield $R_0 = (-\infty, Z^*]$ and $R_1 = [Z^*, +\infty)$, then we may proceed to solve $C'(Z^*) = 0$ for the optimal Z^*. However, we may be not lucky enough to encounter this simple situation: the real number line may in fact need to be bisected with **multiple cutoffs**, e.g. $\{Z_1^*, Z_2^*\}$ for obtaining $R_0 = [Z_1^*, Z_2^*]$ and $R_0 = (-\infty, Z_1^*] \cup [Z_2^*, +\infty)$. Not only does the calculus get cumbersome in solving $C'(\{Z^*\}) = 0$ as the number of cutoffs rise, but we do not even know in advance how many cutoffs are required, so it is impractical to directly obtain $\{Z^*\}$ from the equation.

On the other hand, a careful analysis can brilliantly lead us to a **Bayes detector** without performing actual calculus: H_0 should be chosen if

$$\frac{f_Z(z \mid H_1)}{f_Z(z \mid H_0)} < \frac{c_{10} - c_{00}}{c_{01} - c_{11}} \frac{p}{q},$$

or H_1 should be chosen if the opposite inequality holds; more importantly, the two sides of the inequality become equal at $z = Z^*$, so we can now get all cutoffs by merely solving $\frac{f_Z(Z^* \mid H_1)}{f_Z(Z^* \mid H_0)} = \frac{c_{10} - c_{00}}{c_{01} - c_{11}} \frac{p}{q}$ instead of doing the calculus.

Frequentist statistics name things differently. They regard a Bayes detector as a **likelihood-ratio test**, for which the distribution ratio on the L.H.S. becomes the test's **likelihood ratio** $\Lambda(z)$ and the constant ratio on the R.H.S. becomes its corresponding **threshold** η. As a result, one can apply theorems proved for likelihood-ratio tests, the most important of which is the Neyman-Pearson lemma.

3.1 Example: symmetric triangular distribution

What is Z^* if $f_Z(z \mid H_0)$ and $f_Z(z \mid H_1)$ are both $2a$-unit wide symmetric triangular distributions but respectively centred at k_0 and $k_1 \in [k_0, k_0 + a]$?

Solution. We need to solve $\frac{f_Z(Z^* \mid H_1)}{f_Z(Z^* \mid H_0)} = \eta$ regardless of the actual value of the threshold η. Because a $2a$-unit wide symmetric triangular symmetric distribution centred at k can be symbolically expressed as

$$f_X^{\Delta}(x; a, k) \sim \begin{cases} a + x - k & (k - a \leq x \leq k) \\ a - x + k & (k \leq x \leq k + a) \\ 0 & (\text{otherwise}) \end{cases}$$

up to some normalization constant, the equation becomes $\frac{f_Z(z \mid H_1)}{f_Z(z \mid H_0)} = \frac{f_Z^{\Delta}(z; a, k_1)}{f_Z^{\Delta}(z; a, k_0)} = \eta$ and splits into a system

$$\begin{cases} \dfrac{a+Z^*-k_1}{a+Z^*-k_0} = \eta & (Z^* \in [k_1-a, k_0]) \\ \dfrac{a+Z^*-k_1}{a-Z^*+k_0} = \eta & (Z^* \in [k_0, k_1]) \\ \dfrac{a-Z^*+k_1}{a-Z^*+k_0} = \eta & (Z^* \in [k_1, k_0+a]) \end{cases}$$

based on the range of Z^*. Although there are three roots in this system, two of them will contradict with their assumed ranges and thus get rejected, leaving the self-consistent one as the only valid answer.

3.2 Terminology

Usually H_0 denotes a **negative** hypothesis that suggests the **absence** of whatever we are interested in; correspondingly, H_1 is a **positive** hypothesis that suggests **presence**. Statistics commonly refer to them as a **null hypothesis** and an **alternative hypothesis**.

The scenarios S_{ij}, defined as "believing H_i when H_j is true", may be called in various ways.

S_{11}	S_{10}	S_{01}	S_{00}
true positive	false positive	false negative	true negative
	false alarm	miss	
	type I error	type II error	

The names "**type I error**" and "**type II error**" are extremely confusing but somehow very common in the literature of statistics. The probability of each scenario may also be denoted in several nontrivial ways.

$P(S_{11})$	$P(S_{10})$	$P(S_{01})$	$P(S_{00})$
sensitivity			specificity
power	significance		
$1-\beta$	α	β	$1-\alpha$

You may learn more about the conventions on Wikipedia.

3.3 Alternative detectors

We may resort to alternatives to the Bayes detector when we do not know the exact cost c_{ij}.

Maximum-a-posteriori detector. Usually there is no reward for a correct decision, while the cost of a false alarm sometimes equals that of a miss. Hence we may assume $c_{00} = c_{11} = 0$ and $c_{01} = c_{10} = 1$ and reduce a Bayes detector to a **maximum-a-posteriori (MAP) detector**:

$$\frac{f_Z(z \mid H_1)}{f_Z(z \mid H_0)} < \frac{p}{q},$$

which helps minimize the probability of any wrong decision. It is in fact identical to the MAP rule that we have seen for a naive Bayes classifier because it can be rearranged to become $q f_Z(z \mid H_1) < p f_Z(z \mid H_0) \Leftrightarrow P(H_1 \mid z) < P(H_0 \mid z)$. (In this regard, the MAP detector also helps you remember which hypothesis a less-than or a greater-than sign in a Bayes detector suggests.)

Neyman-Pearson detector. A miss is normally much more costly than a false alarm, though. (While a false alarm of a missile attack merely creates panic, a corresponding miss kills people.) Therefore, we occasionally set up a **tolerable** probability α of false alarm to prevent misses as much as possible. The more tolerant of a false alarm we are, the less likely we will miss. Practically, we let $R_1 = [Z^*, +\infty)$ and solve

$$\int_{R_1} f_Z(z \mid H_0)\mathrm{d}z = \alpha$$

for Z^*. This is called the **Neyman-Pearson detector**. If we observe $Z < Z^*$, we believe H_0; otherwise, we believe H_1 even though we wrongly believe it with a probability α.

The frequentist framework may reformulate this formula in a more abstract way: after transforming $f_Z(z \mid H_0)$ to the distribution $f_\Lambda(\lambda \mid H_0)$ of the likelihood ratio $\Lambda(z)$, we need to solve

$$\int_{\eta^*}^{+\infty} f_\Lambda(\lambda \mid H_0)\mathrm{d}\lambda = \alpha$$

for a critical threshold η^* and believe H_0 if $\Lambda(Z) < \eta^*$ but H_1 otherwise.