Introduction to Bayesian Inference.

Elena Moltchanova

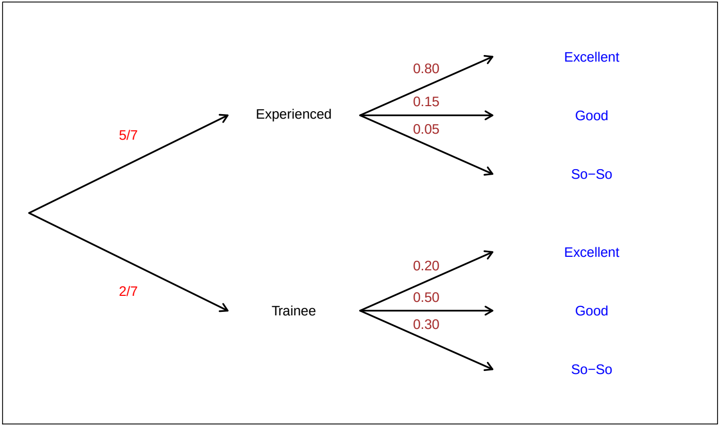
STAT314/461-2021S1

Rev.Thomas Bayes ()

An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S.

in

Philosophical Transactions of the Royal Society of London 53 (1763), 379-418



Experienced
Trainee
5/7
2/7
0.80
0.15
0.05
0.20
0.50
0.30
Excellent
Good
So−So
Excellent
Good
So−So

Pr(good coffee|trainee barista)

1. Meet barista

2. Barista makes coffee

Bayes' Theorem aka Inverse Probability Formula.

Consider a set of mutually exhaustive and mutually exclusive events $A_1, A_2, ..., A_K$ and an event B. Assume that the probabilities $Pr(A_k)$ and $Pr(B|A_k)$ are known for all $k = 1, ..., K$.Then, for some j,

$$Pr(A_j|B) = \frac{Pr(B|A_j)Pr(A_j)}{\sum_k Pr(B|A_k)Pr(A_k)}.$$

Proof.

$$\sum_k Pr(B|A_k)Pr(A_k) = \sum_k Pr(B\&A_k) = Pr(B).$$

Therefore:

$$Pr(A_j|B) = \frac{Pr(B|A_j)Pr(A_j)}{Pr(B)}.$$

Multiplying both sides by $Pr(B)$:

$$Pr(A_j|B)Pr(B) = Pr(B|A_j)Pr(A_j).$$

$$Pr(A_j\&B) = Pr(B\&A_j).$$

Thus, the equality holds.

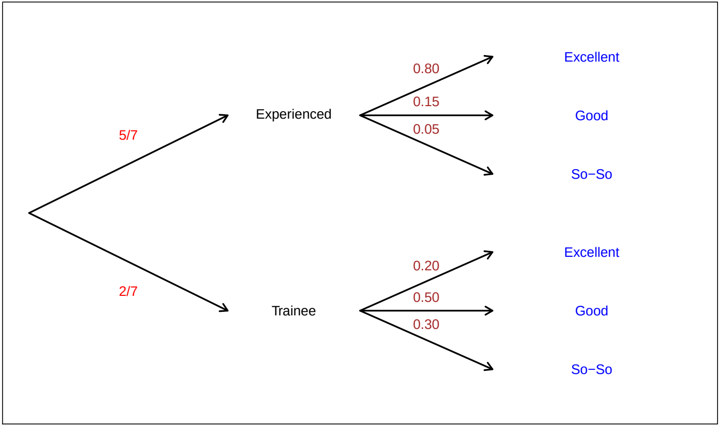
Alternatively:

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}.$$

Back to coffee:

Given that I am drinking a cup of a very good coffee, what is the probability that it was made by the trainee barista?

In other words, let's think of a set of $K = 2$ mutually exlusive and mutually exhaustive events: $A_1 =$ experienced, and $A_2 =$ trainee. And the event of interest $B =$ excellent coffee.

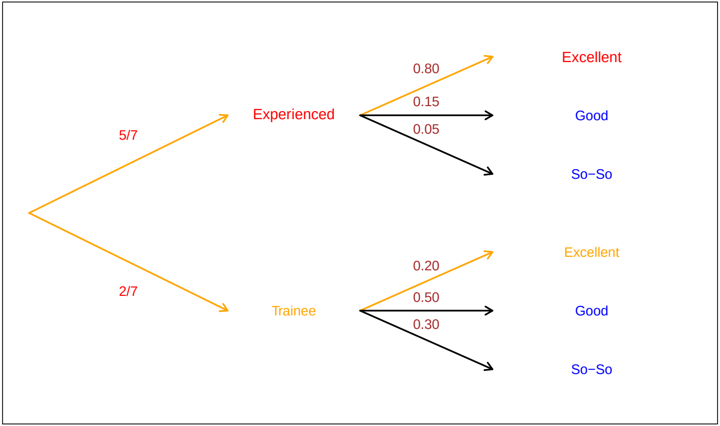

Experienced
Trainee
5/7
2/7
0.80
0.15
0.05
0.20
0.50
0.30
Excellent
Good
So−So
Excellent
Good
So−So

Put numbers into Bayes' Formula:

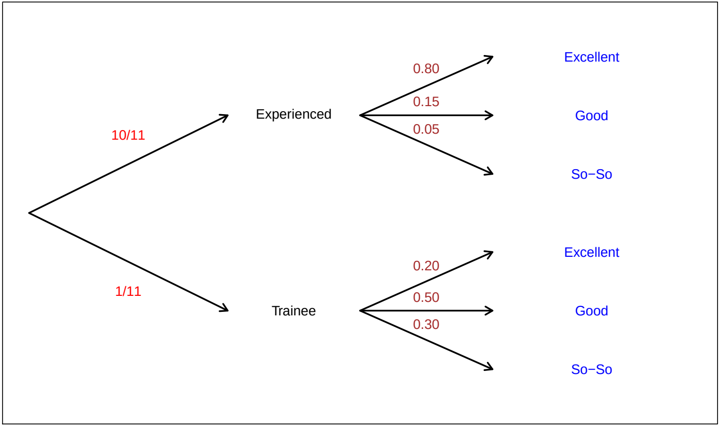
We know that $Pr(A_1) = 5/7$ and $Pr(A_2) = 2/7$. We also know that $Pr(B|A_1) = 0.80$ and $Pr(B|A_2) = 0.20$. We can use Bayes' Theorem to obtain our quantity of interest:

$$\begin{aligned} Pr(A_1|B) &= \frac{Pr(B|A_1)Pr(A_1)}{Pr(B|A_1)Pr(A_1) + Pr(B|A_2)Pr(A_2)} \\ &= \frac{5/7 * 0.80}{5/7 * 0.80 + 2/7 * .20} = 10/11 \approx 0.91. \end{aligned}$$

Using the "tree":



Another cup



Experienced
Trainee
10/11
1/11
0.80
0.15
0.05
0.20
0.50
0.30
Excellent
Good
So−So
Excellent
Good
So−So

Another cup: so-so

$$Pr(\text{Experienced}|\text{Excellent,So-so}) = \frac{10/11 * .05}{10/11 * .05 + 1/11 * .30} = 0.625.$$

Two cups at once:

The probability that you get and Excellent and a So-so cup from the experienced barista is $0.80 * 0.05 = 0.04$. The probability that you get the same from the trainee is $0.20 * 0.30 = 0.06$. Applying Bayes' Theorem we get

$$
\begin{aligned}
&Pr(\text{Experienced}|\text{Excellent,So-so}) = \\
&\frac{Pr(\text{Ex,Ss}|\text{Exp})Pr(\text{Exp})}{Pr(\text{Ex,Ss}|\text{Exp})Pr(\text{Exp}) + Pr(\text{Ex,Ss}|\text{Tr})Pr(\text{Tr})} \\
= \quad &\frac{0.04 * 5/7}{0.04 * 5/7 + 0.06 * 2/7} = 0.625.
\end{aligned}
$$

Using Bayes Formula:

- The object of inference – the probability that the particular barista is at work today – is constantly updated in light of the data.
- This natural learning process is happening sequentially - cup by cup. (Or batch by batch).
- You can easily include other people's observations into your process.

Refresher on the Probability Distributions - 1

Consider two random variables x and y with the corresponding probability density functions (p.d.f.) $f(x)$ and $f(y)$. Note, that we will use $f()$ as a generic notation for any p.d.f.

The following properties apply to any p.d.f.:

1. $f(x) \geq 0 \qquad \forall x$.
2. $\int_{-\infty}^{\infty} f(x) dx = 1$.

Refresher on the Probability Distributions - 2:

The **cumulative density function** (c.d.f.) is defined as

$$F(x) = Pr(X \leq x) = \int_{-\infty}^{x} f(x)dx.$$

Refresher on the Probability Distributions - 3:

The **joint** distribution of x and y is

$$f(x, y) = f(x|y)f(y) = f(y|x)f(x).$$

Here, $f(y|x)$ is referred to as **conditional p.d.f. of y given x**.

Note, that when x and y are independent,

$$f(x, y) = f(x)f(y).$$

Refresher on the Probability Distributions - 4:

The **marginal** p.d.f. $f(x)$ can be obtained as:

$$f(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

Refresher on the Probability Distributions - 5:

In general, for random variables $x_1, x_2, x_3, ..., x_K$ with the joint p.d.f. $f(x_1, x_2, x_3, ..., x_K)$ the following chain rule applies:

$$f(x_1, x_2, x_3, ..., x_K) = f(x_1|x_2, x_3, ..., x_K)f(x_2|x_3, ..., x_K)...f(x_K).$$

Bayes formula for distributions-1

Let $f(x|\theta)$ denote the p.d.f of data x given parameter θ, and let $f(\theta)$ denote the p.d.f. of the parameter θ. Then:

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int_\Theta f(x|\theta)f(\theta)d\theta}.$$

Bayes formula for distributions-2

Note, it is easy to check that this holds:

$$\int_{\Theta} f(x|\theta)f(\theta)d\theta = \int_{\Theta} f(x,\theta)d\theta = f(x).$$

I.e.,

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)}.$$

Classical vs. Bayesian Inference:

Classical:

- Experiments are infinitely repeatable under the same conditions (hence: 'frequentist')
- The parameter of interest (θ) is fixed and unknown
- Inference via Maximum Likelihood

Bayesian:

- Each experiment is unique (i.e., not repeatable)
- The parameter of interest has an unknown distribution
- Inference via Bayes' Theorem

What is prior distribution?

- Prior expresses our knowledge about the parameter distribution before the experiment. It may be based on some general considerations (a binomial probability has to lie between 0 and 1; average human height must lie between 140 and 190) or on previous experiments (the first diagnostic test was positive).
- If no information is available, a so-called vague or uninformative prior can be used. (BUT: what is uninformative?)
- Different statisticians may have different priors. Sensitivity analysis is important.