Conditional Probability, Bayes' Theorem, and Unintuitive Results

Math Circle Competition Team
(motivated by 3Blue1Brown and Zach Star)

Introduction

Bayes' Theorem is one of the most famous theorems in probability. It has wide applications in medicine, economics, artificial intelligence, law, and pretty much any field one can think of. The theorem is named after the 18^{th} century English statistician Thomas Bayes, who was interested in devising a way to estimate probabilities of events given prior observations/data from experiments.

Bayes' Theorem

Let A and B be **events** of some **sample space** S. We have

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

Here, $P(A)$ and $P(B)$ are the **marginal probabilities** of observing A and B, respectively, $P(A \mid B)$ is the **conditional probability** of observing A given B is true, and $P(B \mid A)$ is the conditional probability of observing B given A is true.

Furthermore, suppose events $A_1, \ldots, A_n$ form a **partition** of the sample space S and another event B exists in S as well. For $i = 1, \ldots, n$, the **Law of Total Probability** (LoTP) gives

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots + P(B|A_n)P(A_n)}.$$

$$P(B)P(A|B) = P(A \text{ and } B) = P(A)P(B|A)$$

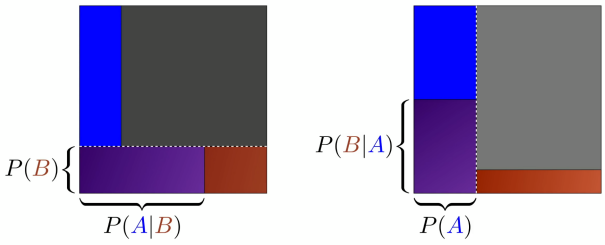

Figure 1: A visualization of Bayes' theorem, taken from 3Blue1Brown

Fundamentally, Bayes' theorem allows us a way to **update our prior beliefs** about the probability of some event happening when we are exposed to new data. To get at what this really means, let's first introduce some background on the Bayesian interpretation of probability.

Frequentist Probability versus Bayesian Probability

Many individuals' first exposure to probability is from a **frequentist perspective**. Indeed, this is how we first introduce the concept here at Math Circle. Frequentists posit that the probability of an event is its relative frequency over time. For instance, if one were to roll a six-sided dice, the probability of getting a 1 would be the ratio of the number of 1's seen to the total number of rolls as this experiment continued ad infinitum.

Frequentist Interpretation of Probability

Let A be an event and let n_a be the number of occurrences of A in n trials. If we have

$$\lim_{n\to\infty} \frac{n_a}{n} = p$$

then we say that $P(A) = p$.

While useful, the frequentist view of probability has some issues. It is impossible to actually perform infinite trials of an experiment. The measured ratio from a finite number of trials can differ if this process is repeated, but an actual probability should be constant.

An alternative interpretation of probability is that of the **Bayesian perspective**. Bayesians (also known as subjectivists) give probability a subjective status by regarding it as measuring the "degree of belief" an individual holds when assessing uncertainty. For instance, when gamblers bet on the outcome of a presidential race, there is an **odds** associated to presidential candidates based on the number of gamblers who believe the candidate will win versus the number who think the candidate will lose. These odds reflect the gamblers' subjective beliefs.

Bayesian Interpretation of Probability

Let A be an event. First establish a **prior probability** for $P(A)$ by taking into account any prior information. Then when new data becomes available, say event B occurs, find the **likelihood** $P(B \mid A)$ of B occurring if A was true (usually known ahead of time). Finally, calculate the **posterior probability** $P(A \mid B)$ using Bayes' theorem to update the prior:

$$\underbrace{P(A)}_{\text{prior}} \overbrace{\frac{P(B \mid A)}{P(B)}}^{\text{likelihood}} = \underbrace{P(A \mid B)}_{\text{posterior}}.$$

This posterior then becomes the new prior, and the process is repeated whenever new data is observed. The **evidence** $P(B)$ can be caluclated with the LoTP as

$$P(B) = \sum_{\text{all } i} \underbrace{P(B \mid X_i)P(X_i)}_{\text{all (likelihoods} \times \text{priors)}} = \sum_{\text{all } i} P(B \cap X_i).$$

One hiccup of the Bayesian approach is that a prior is usually obtained from considering a reference probability, such as from an **urn model** or **thought experiment**. However for a given problem, multiple valid thought experiments could apply. This is known as the **reference class problem**.

Medical Testing

A Classic Example

A 50-year-old woman showing no symptoms participates in a routine mammography screening. She tests positive for breast cancer, and is alarmed. Understandably, she wants to know from her doctor what her probability of having breast cancer is given the positive test. Apart from the screening result nothing else is known about this woman. The doctor has the following information:

- The prevalence of the disease in the population is 1%.
- The **sensitivity** of the test is 90%. That is, the probability of the test giving someone who actually has breast cancer a positive result is 0.9. Sensitivity is also known as the **true positive rate** (TPR).
- The **specificity** of the test is 91%. That is, the probability of the test giving someone who does not have breast cancer a negative result is 0.91. Specificity is also known as the **true negative rate** (TNR).

What is the woman's approximate probability of having breast cancer?

$$\textbf{(A)}\ \frac{9}{10} \qquad \textbf{(B)}\ \frac{8}{10} \qquad \textbf{(C)}\ \frac{1}{10} \qquad \textbf{(D)}\ \frac{1}{100}$$

Between 2006-2007, the psychologist Gerd Gigerenzer gave a series of statistics seminars to practicing gynecologists where he presented this example. In one of the sessions, *over half of the doctors* picked the incorrect answer choice **(A)**. This is a surprising result, as we can calculate with Bayes' theorem:

Solution. Let D be the event that the woman has the disease and $+$ the event that she tests positive. Similarly, let D^C be the event that she doesn't have the disease and $-$ the event that she tests negative. We wish to calculate $P(D \mid +)$. We are given $P(D) = 0.01$, $P(+ \mid D) = 0.9$, and $P(- \mid D^C) = 0.91$. We can then calculate $P(D^C) = 1 - P(D) = 0.99$ and $P(+ \mid D^C) = 1 - P(- \mid D^C) = 0.09$. Thus,

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^C)P(D^C)} = \frac{(0.9)(0.01)}{(0.9)(0.01) + (0.09)(0.99)} = \frac{10}{109}.$$

Answer choice $\boxed{\textbf{(C)}}$ gives the closest approximation.

A visualization that one can do for these kinds of problems is to imagine a population of 1,000 women, 10 of which (1%) have breast cancer and 990 of which do not. With a sensitivity of 90%, the test gives positive results to $(0.9)(10) = 9$ of those with breast cancer (9 true positives, 1 false negative). With a specificity of 91%, the test gives negative results to $(0.91)(990) \approx 901$ of those without cancer (901 true negatives, 89 false positives). Thus, out of those who have tested positive (9 truly, 89 falsely), the ratio of those who have cancer to the total is $9/(9 + 89) = 9/98$, which is in the ballpark of $1/10$. Note there is a slight difference between $9/98$ and the exact answer $10/109$ from rounding $(0.91)(990) = 900.9$ to 901 but this is negligible for an on-the-fly calculation.

The takeaway from Gigerenzer's findings isn't necessarily that practicing doctors have no conception of conditional probability, as statistics training is generally a large portion of many medical school curricula. Rather, Bayesian calculations as they are normally taught are *hard to do on-the-fly*. Unlike us, these doctors were not primed to first visualize a population of 1,000 and break down the numbers from there, but most likely saw the sensitivity and specificity both hovering around 9/10 and went with their gut feeling. It is decently unintuitive that the woman should have a cancer probability as low as 1/10 even though she tested positive. However, recall that before we knew the result of her test, our prior knowledge of her cancer probability was the prevalence of the disease, or merely $1\% = 1/100$. The test actually *updated our prior* by a factor of 10.

The previous example can be made more intuitive, and use Bayes' theorem more efficiently, if we restructure the numbers to be in terms of **odds** rather than probability.

Odds and the Bayes Factor

The odds of an event A is the ratio of the probability of the event to the probability of the event's **complement** A^C. We write

$$O(A) = \frac{P(A)}{P(A^C)}.$$

Odds are often written with a colon. For example, a probability of 1/2 would correspond to an odds of $1:1$, and a probability of 1/10 would correspond to an odds of $1:9$.

The woman's prior (before the test) probability of having cancer was the prevalence 1%, so we can say that her prior odds were $1:99$. Another way to write this would be to return to the visualization of 1,000 people. There are 10 people with the disease and 990 people without, so our prior odds are also $10:990$. After the test, the numerator of the odds is updated by $P(+ \mid D) = 0.9$ from 10 to 9 and the denominator is updated by $P(+ \mid D^C) = 1 - P(- \mid D^C) = 0.09$ from 990 to 89. Overall, our prior odds $10:990$ get updated to our posterior odds $9:89$.

$$\underbrace{O(D)}_{\text{prior odds}} \overbrace{\left(\frac{P(+ \mid D)}{P(+ \mid D^C)}\right)}^{\text{Bayes factor}} = \underbrace{O(D \mid +)}_{\text{posterior odds}},$$

$$\left(\frac{10}{990}\right)\left(\frac{0.9}{0.09}\right) = \frac{9}{89.1}.$$

From here, one can easily convert the odds to a probability by taking the numerator over the (numerator + denominator), or $9/(9+89) = 9/98$. The ratio of sensitivity, $P(+ \mid D)$, to 1 - specificity, $P(+ \mid D^C)$, is known as the **Bayes factor**. It is a **likelihood ratio**.

Note. The media often does not directly report the sensitivity and specificity of a test, but rather the **accuracy**. Usually, the sensitivity and specificity are around the same ballpark as the accuracy, but occasionally they can differ significantly. Keep this in mind when analyzing any media coverage of medical tests. See the next page for a chart of how these metrics are related.

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$

Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Figure 2: Various metrics for statistical and medical testing, taken from Wikipedia

Let P be the number of real positive cases in the data and N the number of real negative cases in the data. Then let TP be the number of true positive cases, TN the number of true negative cases, FP the number of false positive cases, and FN the number of false negative cases all found after a test. We have

- Sensitivity, or true positive rate (TPR):

$$\text{TPR} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

- Specificity, or true negative rate (TNR):

$$\text{TNR} = \frac{\text{TN}}{\text{N}} = \frac{\text{TN}}{\text{TN} + \text{FP}}.$$

- Accuracy (ACC):

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}.$$

Conditional Probability in the Courtroom

Just before noon on June 18, 1964, an elderly lady in Los Angeles by the name of Juanita Brooks was walking home from grocery shopping. As she made her way down an alley, she stooped to pick up an empty carton, at which point she suddenly felt herself being pushed to the ground. When she looked up, she saw a young woman with a blond ponytail running away down the alley with her purse. Near the end of the alley, a man named John Bass saw a woman run out of the alley and jump into a yellow car. The car took off and passed close by him. Bass subsequently described the driver as black, with a beard and a mustache. He described the young woman as Caucasian, slightly over five feet tall, with dark blonde hair in a ponytail.

Several days later, the L.A. Police arrested Janet Louise Collins and her husband Malcolm Ricardo Collins and charged them with the crime. Unfortunately for the prosecutor, neither Mrs. Brooks nor Mr. Bass could make a positive identification of either of the defendants. Instead, the prosecutor called as an expert witness a mathematics instructor at a nearby state college to testify on the probabilities that, he claimed, were relevant to the case. The prosecutor asked the mathematician to consider several ballpark figures pertaining to the features of the two perpetrators:

Description	Probability
Black man with a beard and mustache	1/40
White woman with blonde hair and ponytail	1/30
Interracial couple in a yellow car	1/10,000

Table 1: Ballpark estimates for probabilities of randomly seeing described features

Based on these figures, the mathematician then used the product rule of independence to calculate the overall probability that a random couple would satisfy all of the above criteria, which he worked out to be 1 in 12 million. Impressed by those long odds, the jury found Mr. and Mrs. Collins guilty as charged. But did they make the right decision?

No, the jury forgot how conditional probability works.

1 in 12 million is the conditional probability that a couple would match the descriptions *given they were innocent* (that is, there are a lot of couples in L.A. and randomly selecting a couple probably gives you an innocent one considering there's only one guilty couple). However, the jury conflated this with the not-at-all-the-same probability that a couple is innocent *given they matched the descriptions*. They essentially reversed the conditionals but maintained equality, which we know pretty much never holds! This case became a classic example of the **prosecutor's fallacy**.

Luckily, the Supreme Court would eventually overturn this decision, but the case is a clear example of how a seemingly convincing use of probability can egregiously mislead the unwary. The prior for a random couple being guilty was 1/(# of couples in L.A.). The descriptions then update that prior to the posterior of 1/(# of couples in L.A. that match the descriptions). However, there was almost surely more than a single couple in L.A. that matched those descriptions, L.A. is a huge city! In fact, even if there were only 5 couples in the entire city who matched that description, a reasonable jury should still not have convicted on a mere 1/5 probability (way less than beyond reasonable doubt).

Naive Bayes Classification

In statistics and machine learning, we are often interested in **classification**: the process of separating data into **classes**. For instance, a doctor may be interested in classifying patients into an "at-risk" group or "not-at-risk" group (classes) for a disease using the patients' medical info (data). The doctor may work with a dataset that has some number of rows, which would correspond to the number of patients, and some number of columns, which would correspond to the number of **features**. For instance, age, weight, height, and smoking status could all be relevant features.

Naive Bayes classifiers are classifiers that work by making strong assumptions about the data. Unlike many other classifiers which assume that, for a given class, there will be some correlation between features, naive Bayes explicitly models the features as conditionally independent given the class. This may seem like an overly simplistic (naive) restriction on the data. In our doctor example for instance, individuals in certain age ranges may be more likely to smoke, which would violate the independence assumption between features. In practice however, naive Bayes is competitive with more sophisticated techniques and enjoys some theoretical support for its efficacy.

Because of the independence assumption, naive Bayes classifiers are highly scalable and can quickly learn to use high dimensional features with limited training data. This is useful for many real world datasets where the amount of data is small in comparison with the number of features for each individual piece of data (such as speech, text, and image data). Examples of modern applications include email spam filtering, automatic medical diagnoses, medical image processing, and vocal emotion recognition.

The Naive Bayes Model

Given a data row vector $\vec{x} = \{x_1, \ldots, x_p\}$ with p features (columns), naive Bayes predicts the class C_k out of K distinct classes for $\vec{x}$ according to the probability

$$P(C_k \mid \vec{x}) = P(C_k \mid x_1, \ldots, x_p) \text{ for } k = 1, \ldots, K.$$

Using Bayes' Theorem, this can be factored as

$$P(C_k \mid \vec{x}) = \frac{P(\vec{x}|C_k)P(C_k)}{P(\vec{x})} = \frac{P(x_1, \ldots, x_p \mid C_k)P(C_k)}{P(x_1, \ldots, x_p)}.$$

Using the **probabilistic chain rule** $P(A \cap B) = P(B \mid A)P(A)$, the term $P(x_1, \ldots, x_p \mid C_k)$ in the numerator can be further decomposed as

$$P(x_1, \ldots, x_p \mid C_k) = P(x_1 \mid x_2, \ldots, x_p, C_k)P(x_2 \mid x_3, \ldots, x_p, C_k) \cdots P(x_{p-1} \mid x_p, C_k)P(x_p \mid C_k)$$

At this point the naive conditional independence assumption is put into play. Specifically, naive Bayes models assume that feature x_i is independent of feature x_j for $i \neq j$ given the class C_k. Using the previous decomposition, this can be formulated as

$$P(x_i \mid x_{i+1}, \ldots, x_p, C_k) = P(x_i \mid C_k) \implies P(x_1, \ldots, x_p \mid C_k) = \prod_{i=1}^{p} P(x_i \mid C_k).$$

The Naive Bayes Model (cont.)

Finally, our desired probability is

$$P(C_k \mid x_1, \ldots, x_p) = \frac{P(x_1, \ldots, x_p \mid C_k)P(C_k)}{P(x_1, \ldots, x_p)} = \frac{P(x_1 \mid C_k)P(x_2 \mid C_k) \cdots P(x_p \mid C_k)P(C_k)}{P(x_1, \ldots, x_p)}$$
$$= \frac{P(C_k)\prod_{i=1}^{p} P(x_i \mid C_k)}{P(x_1, \ldots, x_p)} \propto P(C_k)\prod_{i=1}^{p} P(x_i \mid C_k).$$

We use $\propto$ as the "proportional to" symbol. Practically speaking, the class conditional feature probabilities $P(x_i \mid C_k)$ are usually modeled using the same probability distribution, such as the binomial distribution or Gaussian distribution.

Naive Bayes gives the probability of a data point $\vec{x}$ belonging to class C_k as proportional to a simple product of $p+1$ factors (the class prior $P(C_k)$ plus p conditional feature probabilities $P(x_i \mid C_k)$). One classification rule that utilizes this probability is **maximum a posteriori estimation** (or MAP estimation). That is, we assign the class C_k to the data point for which the value $P(C_k \mid \vec{x})$ is greatest. The proportional product can then be used to determine the most likely class assignment.

Maximum a Posteriori Estimation for Naive Bayes Classification

For classes C_a and C_b we have

$$P(C_a)\prod_{i=1}^{p} P(x_i \mid C_a) > P(C_b)\prod_{i=1}^{p} P(x_i \mid C_b) \implies P(C_a \mid x_1, \ldots, x_p) > P(C_b \mid x_1, \ldots, x_p).$$

Thus, the MAP class assignment for a data point $\vec{x} = \{x_1, \ldots, x_p\}$ can be found by calculating $P(C_k)\prod_{i=1}^{p} P(x_i \mid C_k)$ for $k = 1, \ldots, K$ and assigning $\vec{x}$ the class C_k for which this value is largest. In mathematical notation, this is defined

$$\widehat{C} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} P(C_k)\prod_{i=1}^{p} P(x_i \mid C_k)$$

where $\widehat{C}$ is the **estimated** class for $\vec{x}$ given its features $x_1, \ldots, x_p$.

Since determining the most likely class for a data point $\vec{x} = x_1, \ldots, x_p$ consists of calculating the product of $p+1$ factors K times, the **big O notation** for the runtime complexity of classification is $O(pK)$. This is very computationally efficient (see embarrassingly parallel) and gives naive Bayes its high scalability since the runtime scales linearly in the number of features p and number of classes K. This is especially useful in the domain of very high dimensional data, such as bag-of-words classifiers for large vocabulary corpora or high resolution image data, such as MRI scans.

Naive Bayes learns conditional probability features for each feature separately, so it is very efficient for learning to classify small datasets. This is especially the case when $p > n$, where n is the number of training samples (rows) in the entire dataset $X = [\vec{x}_1, \ldots, \vec{x}_n]^T$. This is often the case for medical images, where a single MRI scan can have millions of features (lots of pixels, high p), but can be very costly to obtain (few scans, low n). As long as n is sufficiently large to accurately estimate the individual factors, the number of features p can be any size.

Problems

1. Repeat to the breast cancer example, but now suppose that the prevalence was instead 0.1%. Do the same for a prevalence of 10%. Why can we always multiply the prior odds by the likelihood ratio/Bayes factor to get the posterior odds, but can not do the same with the prior probability?

2. An individual has been described by a neighbor as follows: "Steve is very shy and withdrawn, invariably helpful but with very little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail." Is Steve more likely to be a librarian or a farmer?

3. Linda is thirty-one years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in various protests. Rank the following scenarios from highest (1) to lowest (5) in probability. Linda is:

   ___ an elementary school teacher

   ___ active in the feminist movement

   ___ a bank teller

   ___ an insurance salesperson

   ___ a bank teller also active in the feminist movement

4. A couple has two children, the older of which is a boy. What is the probability that they have two boys?

5. A couple has two children, at least one of which is a boy. What is the probability that they have two boys? Is this the same as Problem 4?

6. A couple has two children, at least one of which is a boy, and that boy was born on a Tuesday. What is the probability that they have two boys? Is this the same as Problem 5?

7. Balls numbered 1 through 20 are placed in a bag. Three balls are drawn out of the bag without replacement. What is the probability that all the balls have odd numbers on them?

8. Zeb's coin box contains 8 fair, standard coins (heads and tails) and 1 coin which has heads on both sides. He selects a coin randomly and flips it 4 times, getting all heads. If he flips this coin again, what is the probability it will be heads?

9. There are 10 boxes containing blue and red balls. The number of blue balls in the n^{th} box is given by $B(n) = 2^n$. The number of red balls in the n^{th} box is given by $R(n) = 1024 - B(n)$. A box is picked at random, and a ball is chosen randomly from that box. If the ball is blue, what is the probability that the 10^{th} box was picked?

10. Suppose one has the training dataset given below:

Shape	Color	Size	Class
circle	blue	large	+
circle	red	medium	−
circle	red	large	−
square	blue	small	−
square	red	small	−
square	red	medium	+
square	blue	medium	+
square	blue	large	−
triangle	red	small	+
triangle	red	large	+
triangle	blue	medium	+

In our training set, our classes are not equiprobable in frequency. We have $P(+) = 6/11$ and $P(-) = 5/11$. We use this as our prior probability distribution. Classify the following sample as either + or − using MAP estimation and the naive Bayes assumption:

Shape	Color	Size	Class
circle	red	small	?

That is, determine which posterior is greater out of + or − for the above sample.

Hint: For the classification as + the posterior is given by

$$\text{posterior}(+) = \frac{P(+)\, P(\overbrace{\text{Shape}}^{\text{circle}} \mid +)\, P(\overbrace{\text{Color}}^{\text{red}} \mid +)\, P(\overbrace{\text{Size}}^{\text{small}} \mid +)}{\text{evidence}}.$$

For the classification as − the posterior is given by

$$\text{posterior}(-) = \frac{P(-)\, P(\overbrace{\text{Shape}}^{\text{circle}} \mid -)\, P(\overbrace{\text{Color}}^{\text{red}} \mid -)\, P(\overbrace{\text{Size}}^{\text{small}} \mid -)}{\text{evidence}}.$$

The evidence (also termed normalizing constant) may be calculated:

$$\begin{aligned}\text{evidence} = {} & P(+)\, P(\text{Shape} \mid +)\, P(\text{Color} \mid +)\, P(\text{Size} \mid +) \\ & + P(-)\, P(\text{Shape} \mid -)\, P(\text{Color} \mid -)\, P(\text{Size} \mid -).\end{aligned}$$

However, given the sample, the evidence is a constant and thus scales both posteriors equally. It therefore does not affect classification and can be ignored. One only needs to compare the numerators of the posteriors.

★ 11. Suppose one has the training dataset given below:

Person	height (feet)	weight (lbs)	foot size (inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

Let's say we have equiprobable classes, so $P(\text{male}) = P(\text{female}) = 0.5$. This prior probability distribution might be based on our knowledge of frequencies in the larger population, or on frequency in the training set. Assume feature distributions conditioned on the classes are Gaussian (that is, $P(\text{height} \mid \text{male})$ follows a normal distribution, and similarly for other feature-class pairs). Classify the following sample as either male or female using MAP estimation and the naive Bayes assumption:

Person	height (feet)	weight (lbs)	foot size (inches)
?	6	130	8

That is, determine which posterior is greater out of male or female for the above sample.