Bayes's Rule and Making Predictions

Subhash Suri

February 14, 2020

1 Predicting the Future

- Peter Norvig, the former director of research at Google, used to give a talk entitled "The Unreasonable Effectiveness of Data," (a play on a famous talk "The Unreasonable Effectiveness of Mathematics in Natural Science" by physicist Eugene Wigner). Norvig's message was "how billions of trivial data points can lead to understanding."
- This has indeed become a mantra of data science: *with plenty of data, no science, no theory, not even algorithms will be needed.* Everything comes down to "patterns in data" from
  - identifying cats in videos to recognizing faces (computer vision) to
  - auto-completion features on your smart phones (language learning) to
  - disease detection (medical diagnoses) to
  - sentencing and bail decisions in the courts (justice).
- Leaving those bold claims aside for the moment, most decisions in our daily life happen at the opposite extreme of the data spectrum: making decisions with "tiny data". In fact, sometimes we need to make inferences from a *single observation.*
- An interesting anecdote of making single-observation based predictions comes from Richard Gott, an astrophysics PhD student at Princeton, who in 1969 happened to be visiting the Berlin Wall in Germany, which was built just 8 years ago so he wondered "how much longer it would continue to divide the East and West."
- On the face of it, this seems like an absurd question to even attempt to answer. Even setting aside the impossibility of forecasting geopolitics, the question seems mathematically laughable: *making prediction from a single data point.*
- But humans make these types of predictions all the time:

  - arriving at a bus stop in a foreign city, and trying to predict when the next bus may arrive based purely on how many people are waiting for how long, or
  - deciding to propose to a girlfriend after dating a few weeks, or
  - deciding to work for a company after one interview and interactions with just a few people.

- The mathematical foundation for investigating these kinds of questions comes from probability theory, and Bayes's Theorem. Before discussing Bayesian approach, we briefly discuss an interesting and famous story involving Pascal and Fermat and the "unfinished game."

2 Fermat-Pascal Letters and the Unfinished Game

- In the development of probability theory, an important milestone is a famous exchange of letters between Fermat and Pascal, two of the most eminent mathematicians of all time.

- In this pre-Internet age, they wrote letters to each other and jointly developed important ideas of probability theory in the process of solving the *game of points.* The game of points is a multi-round game of chance between two players *who have equal chances of winning each round* of the game.

- Specifically, each player contributes an equal amount to a prize pot, and agrees in advance that the first player to win the certain number of rounds will collect the whole prize (pot).

- The problem posed to Fermat-Pascal was this: suppose the game gets interrupted by some unforeseen external circumstances before either player has achieved victory. *How should the pot be divided fairly between them?*

- It was tacitly assumed that the division should somehow depend on how the game has progressed so far, meaning the player who is "closer to winning" should get a larger share. What that fair division ought to be became a subject of mathematical research.

- There is also a popular book by Devlin called "The Unfinished Game," if you want to read more about it.

2.1 Early Solutions

- One of the earliest mentions of the problem dates back to 1494 when mathematician Luca Pacioli suggested that the pot should be divided *in proportion to the number of rounds won by each player.*

- As an example, suppose the first player to win 10 games was to be awarded the pot, and at the time of interruption, player I was leading 7 to 3. Then, according to Pacioli, player I should get 70% of the pot. In his calculations, the number of rounds *needed* to win did not enter.

- In he mid 16th century, another mathematician Tartaglia noticed that Pacioli's method leads to counter-intuitive results if the game is interrupted after just one game. In that case, Pacioli's method would award the entire pot to the winner of that single round, even though a one round win early in the game is hardly decisive.

- Tartaglia devised a method that avoids this problem by basing the division on the ratio between the *size of the lead* and *length of the game*.

- This solution is also not without problems because in a game of 100 rounds, it divides the stakes in the same way for a 65–55 lead as for a 99–89 lead, even though the former is still a relatively open game.

- Tartgalia himself was unsure whether the problem was *solvable* at all, and said that "in whatever way the division is made, there will be cause for litigation."

2.2 Enter Pascal and Fermat

- The problem resurfaced when Chevalier de Mere posed it to Blaise Pascal in 1654, who then discussed it with Pierre de Fermat in his ongoing correspondence.

- Through their discussion, they not only arrived at a convincing and self-consistent solution, but also developed important concepts that are still fundamental to probability theory.

- The starting insight of Pascal and Fermat was that the division should not depend so much on the rounds played so far as on the *possible ways the game might have continued, were it not interrupted.*

- It is intuitively clear that a player with 7–5 lead in a game of 10 has the same chance of eventually winning as a player with a 17–15 lead in a game of 20. According to Pascal and Fermat, interruption in either of these situations should lead to the same division of stakes.

- In other words, what is important is *not* the number of rounds each has won, but rather the number of rounds each player still needs to win to achieve overall victory.

- Fermat now reasoned as follows: if player I needs r more rounds to win, and player II needs s, then the game will surely have been won by someone after $r+s-1$ additional rounds.

- Therefore, imagine that the players were to play $r+s-1$ additional rounds. In total these rounds have 2^{r+s-1} different possible outcomes, because each round is equiprobable to be won by either player.

  - In some of these future rounds, the game will have actually been decided sooner, but it does no harm to imagine the players continuing to play; the final outcome still being the same.
  - Considering only equally-long futures has the advantage that one can easily convince oneself that each of the 2^{r+s-1} possibilities is equally likely.

- Fermat was then able to compute the odds for each player to win, simply by writing down a table of all 2^{r+s-1} possible continuations, and counting how many of them would lead to each player winning. Dividing the stakes in proportion to those odds seems fair, and does not suffer from inconsistencies of previous approaches.

- Fermat's solution was certainly correct, but Pascal improved in two ways First, he produced a more elaborate and convincing argument why the resulting division should be considered fair. Second, he found a way to calculate the correct division much more efficiently, since Fermat's enumeration (exponential) method was completely impractical for even most values of $r+s-1$ (without today's computers).

- Specifically, in a game where player I needs r additional points to win, and player II needs s points, the correct division of the stakes is in the ratio

  $$\sum_{k=1}^{s-1} \binom{r+s-1}{k} \qquad \text{to} \qquad \sum_{k=s}^{r+s-1} \binom{r+s-1}{k}$$

  The first term counts the number of ways player II wins $\leq\ s-1$ rounds, out of the remaining $r+s-1$, and so player I achieves victory. The second term counts number of ways player II wins at least s rounds, and therefore can claim victory.

- While today it's easy for us to evaluate this formula using computers, Pascal did not calculate this directly. Instead, he performed the calculations backward using the following recursive formula. Let $P(n, m)$ be the prob. that player I wins n points before player II wins m points, and p is the prob. that player I wins each round. Then,

$$P(n, m) \;=\; p \times P(n-1, m) \;+\; (1-p) \times P(n, m-1)$$

3 Reverend Bayes and Backwards Reasoning

- Many significant developments in probability theory, including the Fermat-Pascal exchange, have been motivated by gambling. The work of Bayes was no exception. He wanted to understand how likely it was to win a raffle *having observed only a small rounds of previous raffle drawings.*

- Suppose we buy tickets for a new and unfamiliar raffle.
  - If we buy 10 tickets, and 5 of them win prizes, then it seems reasonable to say that the prob. of winning in this raffle is 50%.
  - But what if instead if we buy a single ticket and it wins a prize?
  - Do we really expect the winning prob. for this raffle to be 100%?
  - What should be a reasonable estimate in that case?

- Bayes's critical insight: trying to use the "observed odds of winning tickets to figure out the overall ticket pool" essentially involves to "reasoning backwards."

- Bayes's method to do this estimation was
  - first reason forward from hypotheticals, and
  - then reason backwards to choose the hypothetical that best explains the observations.

- In other words, we need to first determine how probable it is that we would have drawn the tickets we did *if* various scenarios were true. This probability, which we now call *likelihood*, gives us the information we need to solve the problem.

- Consider the following example. Suppose we bought three tickets, and they all won.
  1. If the raffle was a particularly generous one, where all tickets are winners, then our win situation would occur with prob. 1.
  2. If, instead, only half the tickets were winners, then we would have 1 in 8 chance to win all three.

3. If the winning tickets were only 1 in 1000, then our lucky draw would have had a 1 in one billion chance.

- Bayes argued that in the light of our lucky draw it is 8 times more likely that all tickets are winners than only half are, and 125 million times likelier that winning tickets are one per thousand.

- This is the crux of Bayes's argument: reasoning forward from hypothetical pasts is the foundation for working backward to the most probable one.

- It was an ingenious argument although we should point out that it did not completely solve the raffle problem.

  – He did not manage to distill all the (infinitely many) possible hypotheses into a single specific expectation.

  – However, in presenting his work to Royal Society, his friend Richard Price did manage to prove that if you buy a single ticket and it is a winner, then there is *a 75% chance that at least half the tickets are winners.*

3.1 Laplace's Rule of Succession

- Bayes had found a way to compare the relative probability of one hypothesis to another, but he did not manage to distill all the possible hypotheses into a single specific expectation. In the case of raffle, for instance, there are literally infinitely many hypotheses: one for every conceivable proportion of winning tickets.

- Some years later, and completely unaware of Bayes's work, Laplace solved the inference problem from few observations, in an ambitious paper called "Treatise on the Probability of the Causes of Events."

- Laplace was able to prove that this vast spectrum of possibilities could be distilled down to a single estimate, and *surprisingly concise one* at that.

- **Laplace Rule:** Suppose in any drawing of n raffles, if r are observed to be winners, then the (raffle's winning) expectation is

$$\frac{r+1}{n+2}$$

- More precisely, Laplace claimed the following:

> *Suppose an event occurs with some unknown probability p, and before we observed any outcomes our prior is that all values of p are equally likely. Then, after observing r wins in n trials, a good estimate of p is $(r+1)/(n+2)$.*

- For instance, if we observe 5 wins in 10 draws, then Laplace rule gives us winning expectation of $(5+1)/(10+2) = 1/2$, consistent with our intuition.

- If only draw once, and it wins, Laplace's estimate is 2/3, which is quite reasonable; this is also more actionable that Bayes/Price estimate that there is a 75% prob. of a better than 50% chance of success.

- The beauty of Laplace's rule is that it works equally well whether we have a single data point or millions of them. Suppose we have seen the sun rise for one day. What is the prob. we should assign that it will rise again tomorrow? According to Laplace, our best estimate is 2/3. But if it has risen 1.6 trillion days in a row, then the chance of another sun rise is virtually 100%.

3.2 Derivation of Laplace's Rule of Succession

- Let us first consider the simple case in which all n trials result in success, namely, $r = n$. The prob. of this sequence is p^n.

- We need to turn this into a distribution for p that represents our belief. If we believe that all values of p are equally likely, then the distribution for p is just proportional to p^n.

- Since the prob. distributions must integrate to 1, and $\int_0^1 p^n dp = 1/(n+1)$, the full prob. distribution for p must be of the form

$$f(p) = (n+1)p^n$$

- The expectation of this distribution is

$$\int_0^1 p\, f(p) dp = (n+1) \int_0^1 p^{n+1} dp = \frac{n+1}{n+2}$$

- **General Case.** Let us now consider the general case of r successes in n trials. The prob. of this particular event is $p^r(1-p)^{n-r}$.

- Once again, because of our prior belief that all values of p are equally likely, the distribution of p is proportional to $p^r(1-p)^{n-r}$.

- Prob. distributions must integrate to 1, and so use the standard result that

$$\int_0^1 p^r(1-p)^{n-r}dp \quad = \quad \frac{r!(n-r)!}{(n+1)!}$$

- Thus, the full prob. distribution for p must be of the form

$$f(p) \quad = \quad \frac{(n+1)!}{r!(n-r)!}p^r(1-p)^{n-r}$$

- We can now find its expectation

$$\int_0^1 pf(p)dp \quad = \quad \frac{(n+1)!}{r!(n-r)!}\int_0^1 p^{r+1}(1-p)^{n-r}dp \quad = \quad \frac{(n+1)!}{r!(n-r)!}\cdot\frac{(r+1)!(n-r)!}{(n+2)!} \quad = \quad \frac{(r+1)}{(n+2)}$$

3.3 Why is Laplace's Rule useful?

- Suppose we perform n independent trials, and observe that s of them result in success. In the absence of any other information, what prob. of success we should assign?

- The usual rule of thumb is to use $p = s/n$. Laplace's rule of succession suggests that in some circumstances $p = (s+1)/(n+2)$ is more useful.

- Of course, for large n, the two estimates approach the same value, but for small n the latter is more meaningful.

  1. For instance, suppose n is small and none of the trials resulted in success, namely, $s = 0$. In this case, the non-Laplacian method gives $p = 0$, which may be too pessimistic.
  2. Similarly, if the very first trial leads to success, $s = n = 1$, assigning $p = 1$ is also imprudent.

3.4 Prior Beliefs

- Laplace also considered another modification to Bayes's argument: how to handle situations where one hypothesis is more likely than another. For instance, while it *is* possible that a lottery might give away prizes to 99% of the people who play, it is much more likely that only 1% of the tickets are winning ones.

- Consider another toy problem to make this distinction more concrete. A person shows you two different coins, a normal fair coin with a 50-50 chance of head or tails and a two-headed coin. He then drops them into a bag; and pulls one out at random. Without looking at it, he flips it once, and it comes out heads. Which coin do you think is it?

- A fair coin comes up heads 50% of the time, while a two-headed coin comes up heads 100% of the time. Thus, we can say that it is 100/50, or twice, as likely that the coin is a two-headed coin.

- Now consider a twist: the person puts *nine* fair coins and one two-headed coins into the bag, and repeats the same experiment: pulls one coin out of the bag at random; flips it once, and it comes up heads. What are the odds of it being a fair coin now?

- Laplace's method provides an impressively simple solution in this case too. As before, a fair coin is exactly *half as likely* to come up head as a two-headed coin. But now we are also nine times more likely to draw a fair coin out of the bag. Thus, to take these two considerations together, we just need to multiply their odds together, giving us the result that it is 4.5 times *more likely that a fair coin* is pulled than a two-headed one.

- **Stigler's Law of Eponymy.** Ironically, while most of the heavy lifting in developing these ideas was done by Laplace, this mathematical formula is now known as *Bayes's Rule.*

- In fact, there is even a law, called *Stigler's law of eponymy*, proposed by statistician Stephen Stigler, which states that no scientific discovery is named after its original discoverer. Examples include Hubble's law which was derived by Georges Lemaitre two years before Edwin Hubble, the Pythagorean theorem which was known to Babylonian mathematicians before Pythagoras, and Halley's comet which was observed by astronomers since at least 240 BC. Stigler himself named the sociologist Robert K. Merton as the discoverer of "Stigler's law" to show that it follows its own decree.

3.5 Uniform Priors and Copernican Principle

- One crucial element in applying Bayes-Laplace rule is the *preexisting beliefs*: we need to know what kinds of coins are the bag. If we had absolutely no knowledge of what is in the bag, we won't be able to carry out this reasoning.

- This sense of what is in the bag before the coin flip—the chances for each hypothesis to have been true before we saw any data—is known as the *prior probabilities* or "priors" for short.

- Bayes's rule needs some prior, even if it's only a guess. In fact, this dependence on priors has been a source of controversy for Bayesian analysis. But in reality it is quite rare to *not have any prior beliefs.*

- Let us return to Gott's story about the Berlin wall. He asked himself the following question: "in the total life span of this artifact, where have I happened to arrive?"

- In other words, imagining the (future-included) life span of the wall as the *unit interval* $[0, 1]$, what can we say about the present point x at which Gott visited it?

- Gott made the assumption that there was nothing special about it, so it should be treated as a *random* point on this interval. If every moment was equally likely to be x, then its *average* value should be 0.5—that is, he was equally likely to have arrived before the halfway or to have arrived after the halfway.

- If we make this assumption, then there is a simple guess for how long the wall's life span is: *twice the value of x!* That is, it is expected to last exactly as long in the future as it has already lasted.

- Gott had arrived in 1969, which was 8 years after its erection, so he predicted it will last another 8 years, namely, 1977. (It ended up lasting until 1989, so he was off by 12 years.)

- Gott's reasoning is a *temporal* application of Copernican principle, because Copernicus made the case that *earth's* position in the universe wasn't special.

- Using this principle, one would predict that USA will last as a nation until approx. 2264, and Google will last until about 2040, etc.

- Is the Copernican Principle right? After Gott's publication in Nature, a lot of criticism was raised. For instance, according to this principle, if you encounter a 90 year old man, it would predict him to live 180 years, while we will predict a rather early death of a 6 year old.

- To understand why the Copernican principle works, and why it sometimes doesn't, we need to return to Bayes's Rule because Copernican principle is really an instance of the Bayes's rule.

- When predicting the future, such as the longevity of Berlin Wall, the hypotheses we need to evaluate are all possible durations: will it last a week, a month, a year, a decade and so on. To apply Bayes's Rule, we first need to assign *prior probabilities* to each of these durations.

- Turns out that Copernican principle is exactly what results if we apply Bayes's Rule using *uninformative* prior, namely, all priors are equally likely.

- At first glance this may seem like a contradiction: if Bayes requires us to specify our prior beliefs, what happens when we don't have any. In the case of the raffle, for instance, one way to plead ignorance would be to *assume* the "uniform prior," which considers every possible proporation of winning tickets to be equally likely. In the case of Berlin Wall, what this prior says is that we don't have any reason to suspect that one time span is likelier than another.

- However, we have arrived at the Wall after 8 years of existence, so any hypothesis that precludes a 8-year life span should be ruled out, but all other hypotheses are fair game. When Bayes's Rule combines all these probabilities, the Copernical principle emerges: the best guess for future span equals its past span.

- In fact, mathematical reasoning like this had already been applied quite successfully to make predictions from small data even before Gotts.

  1. Statistician Harold Jeffreys tried to determine the number of tramcars in a city given the serial number on *just one tramcar:* just double the serial number.
  2. More famously, during World War II, Allies wanted to estimate the number of tanks being produced by Germany.
     - Statisticians, using serial numbers on captured tanks, came up with an estimate of 246 per month.
     - Military estimates based on risky and extensive reconnaissance suggested a a figure of 1400 per month.
     - After the war, German records revealed the true figure: 245!!

- Thus, the Copernican principle is a reasonable tool exactly in those situations when we know nothing, such as Berlin Wall in 1969, but it is completely wrong when we have strong prior knowledge such as the life span of humans.

- The richer the prior information we have for Bayes's Rule, the better our predictions.

4 Bayesian Learning

- The use of prior data to make predictions about future (unseen) data is the heart of modern machine learning. In this lecture, we briefly describe the simplest of these methods, called Naive Bayes, applied to the problem of "classification."

- In classification, the task is to build a function g that takes as input a vector of features $\mathcal{X}$, and returns a label $\hat{Y}$, which is the algorithm's prediction for the type or class of $\mathcal{X}$. We use $\hat{Y}$ to denote the label computed by the algorithm, and Y to denote the *true* label. We hope to that in most cases our algorithm achieves $Y = \hat{Y}$.

- For instance, suppose we want an algorithm to decide if an animal is a cat or a dog. Our "features" might be length of the animal's whiskers, its weight, a binary variable indicating whether the animal's ears stick up or are droopy etc. Or, deciding whether a fruit is an apple, using features such as its color, size, roundness, etc.

- Each vector of these features corresponds to a particular instance of the animal, and the algorithm needs to decide its output ($\hat{Y}$) on an *previously unseen vector* using *only* these features.

- A natural approach is to predict the label with the highest conditional probability: that is, we choose

$$g(\mathcal{X}) \quad = \quad \text{argmax}_y \; P(Y = y \mid \mathcal{X})$$

  That is, of all the classification types, choose y so that the *conditional prob.* $P(Y = y|\mathcal{X})$ is the largest.

- (We use the standard prob. notation where capital letters are used for names of variables, and lower case for actual values assigned to those variables.)

- In order to help with computing the probabilities $P(Y = y \mid \mathcal{X})$, we are given a bunch of *training data*, consisting of features-label pairs $(\mathbf{x}^1, y^1), (\mathbf{x}^2, y^2), \ldots, (\mathbf{x}^N, y^N)$, where $\mathbf{x}^i$ is a vector of m discrete features for the ith training example, and y^i is the discrete (true) label for this example.

- In classification, Y takes on discrete values, such as cats, dogs, horse, etc. In contrast, in *regression* we try to predict a *continuous value*, such as temperature, movie gross.

- In order to keep our discussion simple, we will assume that our features (as well as the classifications) are all binary: that is, $y^i \in \{0, 1\}$ and $x^i_j \in \{0, 1\}$.

4.1 The Naive Bayes Algorithm

- Recall Bayes's Theorem:

$$\Pr(Y \mid X) \quad = \quad \frac{\Pr(X \mid Y) \; \Pr(Y)}{\Pr(X)}$$

- We wish to return the label Y that maximizes the probability in LHS. We can ignore the denominator prob. in RHS, since it only depends on the inoput vector X, and thus it is sufficient to maximize the numerator. The numerator requires us to compute the probabilities of all possible input vectors conditioned on all possible labels.

- As the number of features or their range of values get large, we will need to estimate the probability for every unique combination of vectors. (Even if we have $|\mathcal{X}| = n$ *binary* features, we need to assign prob. for the $\Theta(2^n)$ combinations, and most likely we will not even have training data for many of those combinations.)

- The naive Bayes gets around this computational bottleneck by making a strong simplifying assumption: rather than estimate prob. of all feature combinations, it *estimates the prob. of each feature independently*, and then simply sums them over the features. This reduces the exponentially many calculations to *linearly* many.

- More specifically, during the training phase, we estimate the probabilities of each label class $P(Y)$ and each feature (conditioned to each label), namely,$P(X_j|Y)$, where X_j is the jth feature.

- Using the MLE (most likely estimator) we compute these priors as follows:

$$\hat{p}(X_j = x_j \mid Y = y) \quad = \quad \frac{(\#\text{ training examples where } X_j = x_j \text{ and } Y = y)}{(\text{ training examples where } Y = y)}$$

- Or, by using the Laplace MAP estimator

$$\hat{p}(X_j = x_j \mid Y = y) \quad = \quad \frac{(\#\text{ training examples where } X_j = x_j \text{ and } Y = y) + 1}{(\text{ training examples where } Y = y) + 2}$$

- Once these priors have been established using our training examples, we can now make predictions for new objects. On an input $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ we estimate the value of y as

$$\begin{aligned} \hat{y} \;=\; g(\mathcal{X}) \quad &= \quad \operatorname{argmax}_y \hat{P}(Y)\hat{P}(\mathcal{X}|Y) \\ &= \quad \operatorname{argmax}_y \hat{p}(Y = y) \prod_{j=1}^{m} \hat{p}(X_j = x_j | Y = y) \end{aligned}$$

where the second equality follows from the (naive) Bayes's assumption that individual feature prob. are independent.

- Multiplying probabilities can lead to numerical stability issues, since we are multiplying many small numbers. In order to get around that problem, it's often more convenient and helpful to convert the product into sum, by using logarithms. That is, we use

$$\text{argmax}_y \left(\log \hat{p}(Y = y) \ + \ \sum_{j=1}^{m} \log \hat{p}(X_j = x_j | Y = y) \right)$$

- This method is called *naive Bayes* because instead of estimating the full joint distribution $\hat{P}(Y, \mathcal{X})$ over all possible feature vectors, it performs naive summation over feature distributions. This assumption is clearly too simplistic and wrong, but practically useful. It allows us to make predictions using space and data that is *linear* with respect to the size of the features.