MAP estimation of continuous parameters, leading to estimating probabilities of unseen things.*

Greg Kochanski
http://kochanski.org/gpk

2004/02/23 00:36:34 UTC

1 Introduction

Today's lecture is a winding road to Good-Turing (and other) ways of estimating the probability of unseen events. We start with the realization that Bayes' Theorem (B'sT) takes your probability estimates literally, and if one of your probabilities are zero, that model is excluded, now and forever, no matter what the weight of supporting evidence. Consequently, one should not lightly declare that any probability is zero.

A related realization is that there can be lots of low-probability events that will happen eventually, even though they haven't happened *yet*[‡]. Because of that, we don't want to assign $P(E) = 0$, just because we haven't observed an event E. (Unless we have some other way of knowing that it is zero, of course.) Thus, we *don't* want to estimate probabilities in the simple and obvious way: $P(E) = N_E/N_{\text{opportunities}}$ (where N_E is the number of times you have observed E, and $N_{\text{opportunities}}$ is the number of times you have looked for E).

The solution is to treat the probability of an event as a theoretical construct, and to find the model that best fits the data we have. This is a beautiful opportunity to use (guess what?) B'sT to find the probabilities that we need to plug into B'sT to do our text classification.

However, to do *that*, we need to figure out how to choose the best value from a continuum of choices. Along the way, we can learn how to do averages, medians, linear regression, and a variety of related techniques.

2 Decision Rules for Bayes' Theorem

A topic we've managed to avoid so far is what decision rule we should use. B'sT, for all it's virtues, merely gives you the probability of each model. It does not pick the best option, because there is more than one way to define what is best. Here are three options:

2.1 Choosing the MAP model.

If you must make a decision, you can just choose the model with the largest probability. This is the **M**aximum **A**-**P**osteriori (MAP) estimate.[1]

MAP will give you the smallest probability of making an error, if you are forced to make a decision.

2.2 Cost and Risk.

If some of errors are worse than others, you might not want to use MAP. You may want to choose the option that gives you the smallest expected cost Lindgren [1976a,b]. This is also known as the minimum Bayesian Risk or the Minimum Expected Risk rule[2]. This is just a codification of the common-sense notion that a small chance of an atrocious outcome may outweigh the certainty of a minor annoyance.[††]

[1] If the prior probabilities are equal, then the MAP estimate is the same as the Maximum-Likelihood estimate.

[2] Obviously, if the models have rewards instead of costs, and you choose the one with the maximum expected reward, you're doing the same thing.

2.3 Confidence Level / Avoiding Decisions

If you have the option of *not* making a decision, choose your confidence level. If you want to be 99% certain you are not making a mistake, remain quiet and indecisive unless one of the probabilities exceeds 0.99.**

2.4 Rules for Continua of Models

All of these decision rules can apply to two, ten, three thousand, or more possible models. As we will see, however, when we go to continuum of related models, new decision rules appear that don't make much sense if one thinks of applying BsT to just a few models. These new rules are things like the expected value Lindgren [1976b] or the minimum squared error solution. They make sense only if you have a sequence of models where neighbouring models are very similar to their neighbours, and where the models are sorted into order.

The expected value doesn't pick out a particular model because it is wonderful on its own, but because it *and its neighbourhood* are important.

3 Bayes' Theorem with a continuum of choices.

We need a continuum of models that are specified by a few real numbers; the numbers are the parameters that select a single model from the continuum. Note that the terminology often undergoes a shift here. Often, the entire continuum of models is spoken of as one parameterised model.

Once we have some models, we can estimate a continuous variable by the standard calculus approach George B. Thomas [1972] of building more and more models, and taking limits as the number of models goes to infinity.

We'll number the models from $M(0)$ to $M(1)$, including fractions like $M(0.4341)$. (The parenthesised expressions mean the same as M_0, M_1, and $M_{0.4341}$.)

For example, here is BsT for 100 models:

$$\begin{aligned}
P\big(M(0.005)|D,C\big) &= \frac{P\big(D|M(0.005),C\big)\cdot P\big(M(0.005)|C\big)}{P(D|C)}\\
P\big(M(0.015)|D,C\big) &= \frac{P\big(D|M(0.015),C\big)\cdot P\big(M(0.015)|C\big)}{P(D|C)}\\
P\big(M(0.025)|D,C\big) &= \frac{P\big(D|M(0.025),C\big)\cdot P\big(M(0.025)|C\big)}{P(D|C)}\\
&\cdots\\
P\big(M(0.985)|D,C\big) &= \frac{P\big(D|M(0.985),C\big)\cdot P\big(M(0.985)|C\big)}{P(D|C)}\\
P\big(M(0.995)|D,C\big) &= \frac{P\big(D|M(0.995),C\big)\cdot P\big(M(0.995)|C\big)}{P(D|C)},
\end{aligned} \tag{4}$$

where

$$\begin{aligned}
P\big(D|C\big) &= P\big(D|M(0.005),C\big)\cdot P\big(M(0.005)|C\big)\\
&\quad +P\big(D|M(0.015),C\big)\cdot P\big(M(0.015)|C\big)\\
&\quad +P\big(D|M(0.025),C\big)\cdot P\big(M(0.025)|C\big)\\
&\quad \cdots\\
&\quad +P\big(D|M(0.985),C\big)\cdot P\big(M(0.985)|C\big)\\
&\quad +P\big(D|M(0.995),C\big)\cdot P\big(M(0.995)|C\big).
\end{aligned} \tag{5}$$

Recall that $P(D|C)$ doesn't depend on the model, so we don't need it if we just want to find the model with the largest probability: $P(D|C)$ makes all the probabilities bigger or smaller together. As usual, D is the data that you have observed, and C represents all the conditions and knowledge surrounding the experiment.

We can write that more compactly like this: For all x (or $\forall x$),

$$P\big(M(x)|D,C\big) = \frac{P\big(D|M(x),C\big)\cdot P\big(M(x)|C\big)}{P(D|C)} \tag{6}$$

where

$$P(D|C) = \sum_{\text{all } x} P\big(D|M(x),C\big)\cdot P\big(M(x)|C\big), \tag{7}$$

or, when we take the limit of an infinite sequence of models,

$$P(D|C) = \int_0^1 P\big(D|M(x), C\big) \cdot P\big(M(x)|C\big) \cdot dx. \tag{8}$$

So, how can we make an infinite sequence of models, and how can we come up with all the infinite number of probabilities $P\big(D|M(x), C\big)$? How can we have an infinite number of prior opinions for $P\big(M(x)|C\big)$?

One needs an analytic model (*i.e.* a function, an algorithm, or more generally a theory) of how $P\big(D|M(x), C\big)$ depends on x. Often, this will be something like a Gaussian probability distribution: $P\big(D|M(x), C\big) = (2 \cdot \pi\sigma^2)^{-1/2} \cdot e^{-(x-D)^2/(2\cdot\sigma^2)}$.* Likewise, you need an analytic probability distribution that gives the prior probability as a function of x.

4 Linear Regression and the like.

So, let's quickly do means, medians, and linear regression.

4.1 Means

To compute a mean, we assume that the data is given by $D = \mu + \eta$, where μ is the true value of whatever we are measuring, and η is a error. The error is assumed to be a random variable that is distributed as a Gaussian with a zero mean and some positive variance (Equation 9 in the Gaussian Distribution digression). If we have several measurements, we will assume that all the errors are independent. And, finally, we will find the best estimate of the true mean, μ. Our models will be parameterised by the mean, so $M(x)$ is the model that claims that the mean is x.

Under these assumptions, B'sT becomes a product of several probabilities (here is the case of three measurements): For all x

$$P\big(M(x)|D_3, D_2, D_1, C\big) = \frac{P\big(D_3|M(x), C\big) \cdot P\big(D_2|M(x), C\big) \cdot P\big(D_1|M(x), C\big) \cdot P\big(M(x)|C\big)}{P(D|C)} \tag{10}$$

We then take the log of both sides of the equation, and substitute in the expression for a Gaussian for $P\big(D_i|M(x), C\big)$. We get the following equation:

$$\begin{aligned} \log\big(P\big(M(x)|D, C\big)\big) \quad &= -(x - D_3)^2/(2 \cdot \sigma^2) - (x - D_2)^2/(2 \cdot \sigma^2) \\ &\quad -(x - D_1)^2/(2 \cdot \sigma^2) + \log(P\big(M(x)|C\big)) \\ &\quad +\text{terms that don't involve } x. \end{aligned} \tag{11}$$

If you look at it closely, you can see that it consists of a sum of terms, each of which is a squared error between the model (x) and the data (D_i), except for the last one, which is our prior.

To get our MAP estimate for the best model, we maximise the left side, which, because of the minus signs, means that we are minimising the sum of squared errors plus the log of our prior:

$$\begin{aligned} -\log\big(P\big(M(x)|D, C\big)\big) \quad &= \left[(x - D_3)^2 + (x - D_2)^2 + (x - D_1)^2\right]/(2 \cdot \sigma^2) \\ &\quad -\log(P\big(M(x)|C\big)) + \text{terms that don't involve } x. \end{aligned} \tag{12}$$

(If we are just finding the MAP estimate, we don't care about the terms that don't involve x: they are constant, and don't shift the minimum to a different value of x.)

Application of some calculus allows us to find the minimum[3]. If $P\big(M(x)|C\big)$ is constant, then we get the familiar formula for an average: $x = (D_3 + D_2 + D_1)/2$. As you might expect, if the prior is not constant, it will affect the average.

With some effort, one can derive χ^2 tests, F-tests, and t-tests by these techniques.

4.2 Medians

The derivation of the median exactly follows the mean, except that we start with a two-sided exponential distribution, instead of a Gaussian[4]. We take $P\big(D|M(x), C\big) = \frac{1}{2\cdot w} \cdot e^{-|x-D|/w}$, and find that x is obtained by minimising the sum of terms in the form $|x - D_i|$: the absolute value of the errors, plus a term involving the prior.

[3] You take the derivative of the equation, and find the point where it is zero. That is the minimum. The derivative of the terms not involving x is zero.

[4] The two-sided exponential has longer tails than a Gaussian: it describes a distribution where there is a significant chance of making a large error. The assumption of a longer-tailed distribution is the reason why a median is less sensitive to occasional outlying data than the mean.

One can see that (as one expects for a median) x will generally be equal to one of the data, because the minimum will normally be at one of the sharp corners that result from the absolute value operation.

Using this formalism, one can compute interesting and useful statistics like a weighted median. Weighted medians are like medians, except that some of the data are more important than others. It's very analogous to a weighted average.

4.3 Linear Regression

Linear regression follows the same approach used to calculate the mean, except that there is more than one parameter needed to select a particular model. By a suitable choice of a (boring) prior, one can derive all the normal formulae for linear regression.

5 Reprise: What is a probability?

Back in the first lecture, we had the following list of reasonable interpretations:

- ...
- It is an estimate of how often something happens, under specified conditions.
- It counts how big a specified subset of possible results is relative to the complete set of possible results.

We're going to combine the last two ideas. The first says that a probability is something you can derive objectively: from counting, from actual data. The second describes a probability as more of an abstract thing: you derive it from *possible* results. It is a ratio of two Platonic ideals: the number of ways your particular event could possibly happen, divided by the number of ways your conditions could be met[5]. Despite their apparent incompatibility, we can both have our cake and eat it, with a little care.

To accomplish the merge, we need to think of a probability as a numeric parameter of a random process that generates our real-world observations. The random process is a theoretical construct, but fortunately there are (mathematical) random processes that beautifully mirror some interesting bits of the real world[†].

An example of a useful random process is the Binomial process as a model for a series of coin flips. The Binomial process assumes that all the coin flips are independent, and it assumes that each flip has the same probability of coming up heads. This probability, p, is an adjustable parameter that controls the process. When p is near zero, the Binomial process produces sequences like TTTTTTTTTTTHTTTTTTTHTT, for p near 0.5, the sequences look like TTHTHHHTHHTTTH, and for p near one, they look like HHTHHHHHHHHHHTH.

For long sequences generated by the Binomial process, the fraction of heads will be close to p. [7]

6 Estimating probabilities from counts.

What we will do is to use B'sT in an almost recursive fashion to estimate how probable are different values of p, given that we've seen the data: $P(p|D)$. The Binomial random process gives us the input we need, $P(D|p)$. Depending on the prior we choose, and the decision rule we use, we can produce a variety of different estimates of p, including Good-Turing estimation Gale [1994].

References

Alexander Bogomolny. *Zipf's Law, Benford's Law.* http://www.cut-the-knot.org/do_you_know/zipfLaw.shtml, 2004. URL `http://www.cut-the-knot.org/do_you_know/zipfLaw.shtml`.

W. Gale. Good-turing smoothing without tears, 1994. URL `citeseer.nj.nec.com/161518.html`.

Jr. George B. Thomas. *Elements of Calculus and Analytic Geometry.* Addison-Wesley, New York, Menlo Park, London, 1972. ISBN 0-07-053917-0, 0-07-113637-1.

Aaron Krowne. *Zipf's Law.* http://PlanetMath.org, http://planetmath.org/encyclopedia/ZipfsLaw.html, January 2003. URL `http://planetmath.org/encyclopedia/ZipfsLaw.html`. Object ID 3422, Canonical name ZipfsLaw.

[5] This is yet another reason why you should think of all probabilities as conditional probabilities. The condition specifies the set of possible results we are considering.

[7] Strictly speaking, as the length of the sequence goes to infinity, the fraction of heads will become arbitrarily close to p.

Wentian Li. *References on Zipf's Law*. North Shore-Long Island Jewish (LIJ) Research Institute, http://linkage.rockefeller.edu/wli/zipf/, 2003. URL `http://linkage.rockefeller.edu/wli/zipf/`.

Bernard W. Lindgren. *Statistical Theory*. In , Lindgren [1976d], third edition, 1976a. ISBN 0-02-370830-1.

Bernard W. Lindgren. *Statistical Theory*. In , Lindgren [1976d], third edition, 1976b. ISBN 0-02-370830-1.

Bernard W. Lindgren. *Statistical Theory*. In , Lindgren [1976d], third edition, 1976c. ISBN 0-02-370830-1.

Bernard W. Lindgren. *Statistical Theory*. MacMillan, third edition, 1976d. ISBN 0-02-370830-1.

Sheldon Ross. *A First Course in Probability*. In , Ross [1984c], 1984a. ISBN 0-02-403910-1.

Sheldon Ross. *A First Course in Probability*. In , Ross [1984c], 1984b. ISBN 0-02-403910-1.

Sheldon Ross. *A First Course in Probability*. Macmillan, New York, London, 1984c. ISBN 0-02-403910-1.

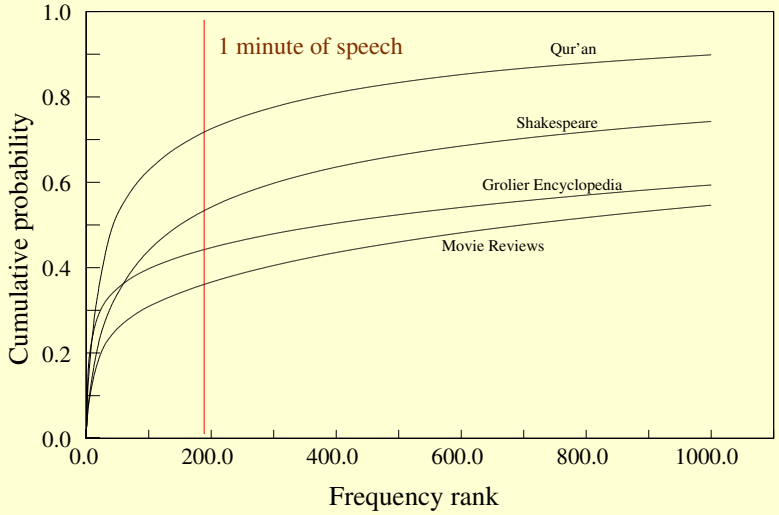
Eric W. Weisstein. *Central Limit Theorem*. Wolfram Research, Inc., http://mathworld.wolfram.com/CentralLimitTheorem.html, 1999-2004. URL `http://mathworld.wolfram.com/CentralLimitTheorem.html`. MathWorld, also copyright 1999 CRC Press LLC.

G. K. Zipf. *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press, Cambridge, MA, 1932.

‡Zipf's Law: Especially in language, where Zipf's Law Zipf [1932], Ross [1984a], Krowne [2003], Li [2003], Bogomolny [2004] often applies, if you keep accumulating data, you will keep observing new, previously unseen events. Many aspects of language follow this behaviour, where there are a relatively few high-probability events (*e.g.* words like "and" and "the" and "it"), and vast numbers of improbable events (*e.g.* words like "interdepartmental" and "geological"). Characteristically, the sum total of all the improbable events is comparable to the sum of all the high-probability events.
Formally, Zipf's law is

$$P(\text{i}^{\text{th}} \text{ most common word}) = P(\text{the most common word})/i, \tag{1}$$

although many variants of it have been proposed that adjust the behavior for the most common few words, for very rare words, or the slope in between.
The figure below shows some examples of word frequencies in different documents illustrating this general principle. The vertical axis is the cumulative probability for all the words more common than a particular word. The rapid initial rise implies that a few very common words account for much of text, and the slow increases thereafter You can see that the hundred most common words account for much of a text: from 30% for movie reviews up to 70% for the Qur'an. You can also see that the rare words, such as words with rank greater than 1000, also account for much of the text: 10% for the Qur'an up to 50% for movie reviews. (Movie reviews and Encyclopedias have an unusually large number of rare words because they contain many names of people and things.)
As you can see, the curves are not identical, so Zipf's law is an approximation, but it captures an important observation.



The horizontal axis is the frequency rank of a word (*i.e.* rank zero is most common, rank 1000 is the thousandth most common). The fifty most common words account for a large amount of a text, but many low-probability words, out beyond rank 1000 are needed to fill in in between. To give a sense of scale, the vertical line marks how many words one can comfortably say in a minute.

††Example of Minimum Risk decision rule.: You get the cost or risk associated with a choice by adding up the cost of each possible outcome times the probability that it happens. Often, you have a different cost for each model, and a different cost if the model turns out to be true or false.
Let's clarify that with an example. Imagine it is summer, and you have a house with an old central heating system. Based on various data (gurgles, leaks, buzzes and thumps), you think that it might not make it through the winter. So, M_1 is "The heating system will fail before spring," and you have settled on $P(M_1|Data) = 0.1$. M_2 then means that it will work all winter, and $P(M_2|Data) = 0.9$. (These probabilities could well be the outcome of B'sT.)
Under MAP, you'd simply chose M_2, and act as if it were true: you would put off repairs until next year. Under a minimum risk decision rule, though, you consider the costs of all options, and they are:

Decision (*i.e.* which model do you choose to accept), α.	Eventual Outcome (*i.e.* which model would have turned out to be true), β.	Cost ($\mathbb{C}_{\alpha,\beta}$)
M_1 (fail)	M_1 (fail)	£3,000 paid now, while the contractor is short on business.
M_1 (fail)	M_2 (survive)	£3,000 paid now.
M_2 (survive)	M_1 (fail)	£4,000 paid in midwinter for emergency repairs, plus the risk of water damage. All told, £5,000.
M_2 (survive)	M_2 (survive)	£3,000, paid next year, so you can accumulate the interest on the money for a year. The Net Present Value of that money is £2,850.

Now, if we act upon M_1, the risk is

$$\begin{aligned}\mathbb{R}_1 &= \textstyle\sum_{\beta\in\text{possibilities}} P(M_\beta|D)\cdot\mathbb{C}_{1,\beta}\\ &= P(M_1|D)\cdot\mathbb{C}_{1,1} + P(M_2|D)\cdot\mathbb{C}_{1,2}\\ &= 0.1\cdot£3000 + 0.9\cdot£3000\\ &= £3000.\end{aligned} \tag{2}$$

Likewise, if we act upon M_2,

$$\begin{aligned}\mathbb{R}_2 &= \textstyle\sum_{\beta\in\text{possibilities}} P(M_\beta|D)\cdot\mathbb{C}_{2,\beta}\\ &= P(M_1|D)\cdot\mathbb{C}_{2,1} + P(M_2|D)\cdot\mathbb{C}_{2,2}\\ &= 0.1\cdot£5000 + 0.9\cdot£2850\\ &= £3065.\end{aligned} \tag{3}$$

So, we see that the expected cost of fixing it now is slightly lower than the expected cost of hoping it will make it through the winter. We should then fix the heating system to avoid the 10% chance of getting stuck with a large midwinter bill.
Minimum expected risk solutions are quite sensible for many purposes, and are used for all sorts of economic decisions, from nuclear reactor design on down. The do require you to be able to quantify your costs into a single number, which is not always possible.

**** Expressing Confidence Levels in terms of Risk:** A Confidence Level decision rule can also be expressed as a minimum-risk decision rule. You just imagine you have one extra choice: add the option of "Neither" to Model 1 and Model 2. Each option has a cost if true or false. Only Models 1 and 2 are possible, so that $P(\text{Neither}|\text{Data}) = 0$, bit if the cost of sitting quiet and choosing "Neither" is small, and the cost of being noisily wrong is large, then it may pay to choose neither Model 1 nor Model 2.

*** Gaussian Distributions:** Gaussian probability distributions are a popular assumption for two reasons. First, they are easy to calculate, but second, because the Central Limit Theorem Ross [1984b], Weisstein [1999-2004], Lindgren [1976c] implies that they must be very common.

Theorem 1 *The sum of arbitrary random variables that have equal variance approaches a Gaussian distribution, as the number of random variables goes to infinity. The mean of the Gaussian is the sum of the individual means, and the variance of the Gaussian is the sum of the individual variances.*

The above is true even if the variances are not equal, so long as all the variances are within some finite range that does not include zero. (However, it may take a long time to approach a Gaussian if the largest variance is far larger than the smallest variance.) That theorem means that almost any measurement that is an average or sum of more than a handful of numbers will have a probability distribution close to Gaussian.

The general form of a Gaussian probability distribution is

$$P(D|M) = (2 \cdot \pi\sigma^2)^{-1/2} \cdot e^{-(D-\mu)^2/(2\cdot\sigma^2)}, \tag{9}$$

where μ is the mean of model M, σ is it's standard deviation, and σ^2 is the model's variance. A Gaussian drops off rather rapidly when D gets more than about $3 \cdot \sigma$ from μ. When you are $6 \cdot \sigma$ out, the probability is about a million times less than at the centre, and it drops even faster as you go out further.

†Limitations of purely Data-Driven Approaches: An example of a good match between a mathematical random process and part of the real world is a series of coin flips. The math assumes that each coin flip is independent of all other flips, and it assumes that the probability of getting heads and tails doesn't change from flip to flip. It also assumes that there are only two possible outcomes: heads and tails.

In the real world, these assumptions hold good. Based on our knowledge of solid state physics, we know that the coin has no memory, and won't wear out or change its properties in the course of a few hundred flips. From knowledge of human anatomy and the way nerves and muscles work, we know that humans cannot control the spin and trajectory of the coin precisely enough[6]

So, we deduce from our models of humans and of coins that the flips should be independent, and that the probability distribution is *stationary* (*i.e.* it doesn't change from flip to flip). We can then experimentally check that those properties hold for our actual coin.

As an aside, note that the theoretical understanding is crucial, because you can never exclude the possibility that the coin is non-stationary based on just the data alone. After all, if the coin produces HTHHTTH, perhaps it had P(H)=1 for the first flip, which changed to P(H)=0.01 for the second flip, to P(H)=0.98 for the third flip, *et cetera*. This could be accomplished by hiring a magician and asking him to produce $HTH\ldots$, or by a properly designed collection of computers and servomechanisms. If so, the sequence that you observed would be a quite normal and highly probable result from that particular non-stationary random process, and you have no way of telling by just observing the result. One has to understand the mechanism that produces the coin flips to know.

If we assumed that all coins behaved the same way, then perhaps we could flip 100 coins fresh from the mint, and show the assumption to be false, but other assumptions are equally plausible.

Similar arguments show that data alone cannot exclude all possible ways the coin flips could be mutually dependent. One needs at least enough theoretical understanding to know that (for example) any possible dependence would be short-ranged, affected by only the last few flips, before you can construct a statistical test to look for correlations between coin flips.