Applying Bayes and Beginning Classifiers.*

Greg Kochanski
http://kochanski.org/gpk

25 January 2004

1 Odds Ratios

Bayes' Theorem can be particularly simple if there are only two models. In that case, you can turn it into a nice rule for updating the "Odds ratio" of the two models whenever you get new data.

1.1 What is an Odds Ratio?

An odds ratio is just the ratio of the probabilities of two alternate outcomes. Odds ratios are traditionally used for betting: one might say that "The odds against Distant Thunder are nine to one" to mean that the horse is nine times more likely to lose than to win. In general, for event X under conditions C, the odds ratio is

$$R(X|C) = \frac{P(X|C)}{P(\neg X|C)}, \tag{1}$$

or (after a spot of algebra)

$$R(X|C) = \frac{P(X|C)}{1 - P(X|C)}. \tag{2}$$

If you have an odds ratio, you can reverse the process and solve for the probabilities:

$$P(X|C) = R(X|C)/(1 + R(X|C)). \tag{3}$$

For example, if $P(X|C) = 0.1$, then $R(X|C) = 1/9$; or if $P(X|C) = 0.9$, then $R(X|C) = 9$.

There is a useful rule for odds ratios that $R(\neg X|C) = 1/R(X|C)$, but you can derive that from Equation 1. One can see that if you swap X for $\neg X$, the right hand side simply turns upside-down.

1.2 Bayes' Theorem in Odds Ratios

If we have two alternative models (M and $\neg M$), and D is the observed data, then we can write Bayes' Theorem in its normal form as:

$$P(M|D,C) = P(D|M,C) \cdot P(M|C)/P(D|C) \quad (4)$$

$$P(\neg M|D,C) = P(D|\neg M,C) \cdot P(\neg M|C)/P(D|C). \quad (5)$$

Recall that $P(D|C)$ is the messy bit[1].

However, we come to bury $P(D|C)$, not to praise it, and the way we do it is by dividing Equation 4 by Equation 5, term by term. The result is Bayes' Theorem in odds-ratio form:

$$\frac{P(M|D,C)}{P(\neg M|D,C)} = \frac{P(D|M,C)}{P(D|\neg M,C)} \cdot \frac{P(M|C)}{P(\neg M|C)} \quad (6)$$

or (in compact form, suppressing the usual condition C)

$$R(M|D) = \frac{P(D|M)}{P(D|\neg M)} \cdot R(M). \quad (7)$$

This has a particularly clean interpretation: the odds ratio *after* seeing the data (left side) is equal to something that depends on the data *times* your odds ratio *before* seeing the data. The data-dependent part (the first term on the right side, called the "Bayes Factor" (Jackman, 2004a,b)) is also easy to understand: it tells you whether the data you actually observed (D) is more likely under the model (M) or the alternative model ($\neg M$). To the extent that the Bayes factor is far from one (either bigger or smaller), the observation provides evidence for one or the other model.

1.3 The Bayes Surprise

Bayes' Theorem, for all it's mathematical simplicity is capable of generating some highly counterintuitive results. We'll start with the medical example.

Let's imagine you go to the doctor for a routine blood test, and the test indicates that you have Acanthosis Nigricans, a rare skin disease. Suppose that it's rather a good diagnostic test that has no false negatives[2] and only 1% false positives: if you really have the disease, it will certainly indicate it, but it occasionally gives a positive indication for a healthy person who takes the test.

Now, Acanthosis Nigricans is not common, indeed only about one in a million people suffer from it. So, if model M means "you have the disease," the odds

[1] In general, if you have N models of which exactly one must be true, $P(D|C)$ will be the sum of N terms, one for each model. You compute it from a decomposition of D into elementary events $D \cap Model_1$, $D \cap Model_2$, $D \cap Model_3$, ... for each of the models. In this case, $N = 2$, the models are M and $\neg M$, and $P(D|C) = P(D \cap M|C) + P(D \cap \neg M|C)$, thus $P(D|C) = P(D|M,C) \cdot P(M|C) + P(D|\neg M,C) \cdot P(\neg M|C)$.

[2] A false negative is when the test falsely indicates that an ill person is healthy, and a false positive is when it falsely indicates that a healthy person is ill.

ratio before the test was 10^{-6}. The Bayes factor (the data-dependent term) in this case is

$$\frac{P(\text{positive result}|\text{disease})}{P(\text{positive result}|\neg\text{disease})}, \tag{8}$$

which equals 10^2. When you multiply to get the odds ratio *after* the test, you – surprisingly – find it to be still quite small: 10^{-4}. Thus, even after getting a positive result from this rather good test, it is still extremely unlikely that you have the disease.

The reason for this strange result is actually pretty simple. Imagine that we have a population of a million people[3]. In that group, we know that there will be about one person with the disease (although we don't know if that person is you or not). However, if you were to apply that test to a million people, the 1% false positive rate would imply the test would also give 10,000 positive results to people who were healthy.

Then, if we look at all the people who (like you) received positive results[4], there will be one with the disease and 10,000 without. Thus, the probability of having Acanthosis Nigricans, given a positive test result, is just 10^{-4} or 0.01%. So, sometimes, even with good tests with just a 1% false positive rate, you would still need several independent tests to actually provide convincing evidence that you actually have a rare disease. You would need several tests because you start out with strong evidence that you *don't* have the rare disease: the disease *is* known to be rare, after all.

Similar examples can be constructed in jurisprudence (finding one criminal out of a million suspects) or anti-terrorism (finding one terrorist out of a million air travellers). In either case, the job of the police is made very hard by the presence of huge numbers of innocent people. Even if you have evidence that would be absolutely convincing at selecting which of two people was guilty, it may still fail at picking the target out of a crowd because there are so many chances for the 999,999 innocents to accidentally display whatever feature the police are looking for.

2 Model Selection based on data

There is not any uniquely best way to pick a single model from the competing possibilities. By convention, we normally choose the most probable model. If the probability comes out of Bayes Theorem, this is known as Maximum a-Posteriori (MAP) model selection. If you don't believe in prior probabilities and would rather compute likelyhoods (thus effectively setting all the priors equal), this is the well-known and common Maximum Likelyhood (ML) model selection.

However, other choices are possible and can sometimes more appropriate. The action we take based on the probabilities may have consequences, and the results could be vastly different.

[3] About a county's worth of people.

[4] This is conditioning on D.

Consider crossing a street. We will produce some sort of an estimate of P(car coming soon) by listening and (hopefully) looking around. We could take the MAP choice, which would mean that we step out into the road if there is a 51% chance that the road is clear. Unfortunately, if we use that strategy, 49% of the time we will step out in front of a car and the consequences could be serious. So, we want to avoid that kind of mistake and bias our decision in the other direction. We want to step out only when it is 99.9% certain that the road is clear.

As a consequence, we will certainly make more errors in the other direction: sometimes we will miss an opportunity to cross the street and wait a bit longer, but the cost of that delay is usually small. This is an example of a "minimum-risk" model selection, where each model comes with a cost of mistaken selection and a cost or benefit for correct selection. You then choose the action that minimizes your expected risk, evaluating the result for each possible outcome and the probability of that outcome.

3 Repeated Bayes' Theorem

The odds-ratio form is particularly convenient for repeated application of Bayes' Theorem. Suppose we use Bayes' Theorem to give us this morning's estimate of the odds ratio of two models in terms of yesterday's. If we get another datum today, we can use this morning's estimate of the odds ratio as input (as the prior estimate), apply the Bayes factor for today's datum, and get a new estimate of the odds ratio ready for tomorrow morning. Yesterday's estimate is prior to both data, this morning's estimate incorporates one datum and prior to the other, and tomorrow morning, we will have an estimate that incorporates both data.

The flow of data looks like this: we start with $R(M|C)$, add yesterday's data and apply Bayes' Theorem to get $R(M|$ yesterday's $D,\ C)$, then add today's data and apply again to yield $R(M|$today's $D,$ yesterday's $D,\ C)$. Formally, when we apply Bayes' Theorem the second time, we include yesterday's data into condition C: our knowledge at the start of the second experiment includes the knowledge we gained yesterday.

We can write out Bayes' Theorem for two sequential observations as follows:[5]

$$R(M|D_t, D_y) = \frac{P(D_t|M, D_y)}{P(D_t|\neg M, D_y)} \cdot \frac{P(D_y|M)}{P(D_y|\neg M)} \cdot R(M). \tag{9}$$

Note that the left-most Bayes Factor (containing D_t) is conditioned on the first observation. That is irrelevant if D_t and D_y are independent events, but sometimes it can make quite a bit of difference.

Here are some examples of what might happen:

[5] C is suppressed.

3.1 Strongly correlated events:

Assume D_y and D_t stand for the events "I am less then one meter tall" on two successive days. In that case, if D_y is true, D_t will very likely be true also. Model M stands for the statement "I am less than five years old."[6] Then, $P(D_y|M)$ is fairly large, and $P(D_y|\neg M)$ is fairly small, so yesterday's Bayes factor might be (*e.g.*) 5. Since it is substantially larger than one, it shows that being short is good evidence of youth.

Today's Bayes factor is different. Since it is conditioned on yesterday's data and the height of most people changes slowly, if you were less than a meter tall yesterday, you are almost certain to be less than a meter tall today, too. That means $P(D_t|M, D_y)$ is very close to 100%: it's only false for those few children who were 99.9 cm tall yesterday and ate their vegetables. Likewise, $P(D_t|\neg M, D_y)$ is also very close to 100%; it is only false for a few short five year olds who grew past the threshold. Consequently, the ratio of those probabilities (which is today's Bayes factor) will be very close to one. It shows that *remaining* short is not strong evidence either for or against youth. We didn't learn much from today's height measurement because yesterday's revealed all the information that is available.

In general, with strongly correlated measurements, later Bayes factors will tend to be close to unity because they are predictable from earlier measurements.

3.2 Independent Events:

Discuss: what's wrong with this example?

Buying a newspaper last week is evidence for literacy. Buying a newspaper today is also evidence for literacy[7], and we assume that there is no cause that would either force you to buy newspapers on those two days. In general, for two independent events D_t and D_y, $P(D_t|M, D_y) = P(D_t|M, \neg D_y) = P(D_t|M)$, and we can write the odds-ratio form of Bayes' Theorem as follows:

$$R(M|D_t, D_y) = \frac{P(D_t|M)}{P(D_t|\neg M)} \cdot \frac{P(D_y|M)}{P(D_y|\neg M)} \cdot R(M). \tag{10}$$

Now, if 95% of people are literate, the odds ratio for literacy started at 19:1 last week. Let's assume that literate people are 30% likely to buy a newspaper on a given day and illiterate people are 5% likely to buy[8], then the Bayes factor $\frac{P(\text{newspaper}|\text{literate})}{P(\text{newspaper}|\neg\text{literate})} = \frac{30\%}{5\%} = 6$. So observing person X buying a newspaper boosts his/her odds ratio up to 114:1, making it 99% likely he or she is literate.

[6] To make everything precise, we should really say something like "I was less than five years old on the date of the first height measurement," or "I was born after 1/1/2000." Since we are using a single symbol for the model, we need to make sure that the model means the same thing today as it did yesterday. Tying it to an absolute date does that. Tying it to your age is not correct because today might be your birthday.

[7] Neglecting people who buy newspapers to paper-train new puppies.

[8] Some may want to appear literate, and others may want the phone numbers in the adverts or buy the paper to wrap fish and chips.

A second observation (on a different date) of a newspaper purchase will boost his odds ratio even higher, to 684:1, so it will be 99.8% likely that he or she can read[9].

3.3 Repeated Measurements, General Case

In the case of multiple observations, we have one Bayes factor for each observation. These factors are multiplied together and with the prior odds ratio. If the events are independent, each Bayes factor is conditioned only on the model (and C, of course, as an extension of Equation 10). If the events are dependent, the Bayes factors are conditioned on all the previous data (extending Equation 9).

Repeated Bayes Theorem in the probability representation (as opposed to the odds-ratio representation) is very useful, but too messy to write out. Unless you are using a computer algebra system, you simply calculate it numerically, one step at a time (*c.f.* Equation **??** or **??**), being careful to keep track of what is conditioned on what.

3.4 Conditioning and Text Classifiers

Conditioning matters because there are lots of correlations in language. Neighbouring letters in a word are highly correlated because they are part of the same word. Nearby words are correlated because they are in the same sentence, and even distant words are somewhat correlated because they are (presumably) all related to the topic of the document. We'd like to assume things are independent, but we know that they (mostly) aren't.

One obvious example that we will look at soon in detail is a text classifier. To make our classifier work as well as possible, we'd like to use more than one word or letter; ideally we'd use all the text: every letter and every word. Unfortunately, nearby places in texts are highly correlated, so we should use Equation 9. However, nearly everyone uses Equation 10 anyway, because there is no good way of estimating all the probabilities in all the different conditions that you'd need for the exact calculation.

This is a deep and serious problem, not just an inconvenience. Let's see why.

Suppose we are classifying text into English or Spanish by looking at a single character at a time, computing whether that character is more likely to be part of an English text or a Spanish text, then using that Bayes factor to update the odds ratio between the two models for the text (English and Spanish).

Were the characters (data) independent of each other, we would need to compute only about 150 probabilities, corresponding to observations of all the letters, uppercase, spaces, and punctuation marks in the two different languages:

[9] Astute readers will recognize that this is only true if the act of buying a newspaper today is independent of the act of buying it yesterday. In other words, I'm making the bad assumption that there is no single event that might make you buy a newspaper for two days in a row. Obvious conterexamples are getting a new puppy, opening a business that needs newspapers to wrap fish, or having a week-long visit from someone who can read.

P(letter="a"|English Text), P(letter="a"|Spanish Text),
P(letter="b"|English Text), P(letter="b"|Spanish Text),
...,
P(letter="Z"|English Text), P(letter="Z"|Spanish Text).

One can estimate all the needed probabilities if one has 10^3–10^4 characters of text available. So, if adjacent letters were independent, we could train a Bayes unigram classifier on a few pages of text, and it would then give the best possible classification. If that were true, computational linguistics would have happily come to the end of the road around 1960, terminated by its own success.

In such a world, a classifier is a simple thing. One picks a character from the document, computes the Bayes factor, multiplies the odds ratio, and repeats. Eventually, the odds ratio will become either so big or so small that one of the hypotheses is effectively excluded. At that point, stop and announce the decision.

4 Text Classifiers in the Real World

But, fortunately for our expressive power, and fortunately for the employment prospects of computational linguists, adjacent letters *are* correlated. What does that imply? Horrible things, in principle.

If we look at two characters from the file, we find that we need all the above probabilities for the Bayes factor of the first character we pick from the file, and then a much bigger set for the second. We need the probabilities of all the letters, conditioned on the two models, and also conditioned on all the values of the first character:

$P(a|\text{English, first="a"})$, $P(\text{Spanish}|S, \text{first="a"})$,
$P(b|\text{English, first="a"})$, $P(b|\text{Spanish, first="a"})$,
...,
$P(w|\text{English, first="a"})$, $P(w|\text{Spanish, first="a"})$,
...,
$P(Z|\text{English, first="a"})$, $P(Z|\text{Spanish, first="a"})$,
$P(a|\text{English, first="b"})$, $P(a|\text{Spanish, first="b"})$,
$P(b|\text{English, first="b"})$, $P(b|\text{Spanish, first="b"})$,
...,
$P(Z|\text{English, first="b"})$, $P(Z|\text{Spanish, first="b"})$,
$P(a|\text{English, first="c"})$, $P(a|\text{Spanish, first="c"})$,
...,
$P(Y|\text{English, first="Z"})$, $P(Y|\text{Spanish, first="Z"})$,
$P(Z|\text{English, first="Z"})$, $P(Z|\text{Spanish, first="Z"})$.

To compute our odds ratios, we need to be able to estimate any of those 10,000 probabilities which we can do if we have 10^6 characters of training text, more or less.

But, it gets worse. Should we want to use four characters, we need to be able to estimate any one of the multiple millions of probabilities that look like $P(m|E,$ first="v", second="a", third="t"), and training text starts get-

ting scarce. Note that these are needed, in principle, even if we are using four isolated unigrams from the same document, let alone N-gram models.

So, we are frustrated. Bayes' Theorem is the exact and optimal solution to classifying texts, but we can't use it without a training set the size of the entire universe. Much of the field of computational linguistics is devoted to finding a way around this problem.

What we need is a theory. Specifically, a theory that predicts these conditional probabilities.

4.1 Invariants and Approximations

Normally, we make some common-sense assumptions about documents that are fairly safe. These assumptions allow us to predict many of the probabilities that we need in terms of others, and drastically reduce the number of probabilities that we need to estimate.

For instance, we know that characters far apart in documents are indeed pretty nearly uncorrelated[10], so if we were to randomly sample ten or even a hundred unigrams from a page of text, the probability distribution for each sample[11] would be nearly independent, and we could use Bayes' Theorem with confidence that our calculated probability was nearly correct.

We often assume that distant words are independent. This is not a bad assumption for many, commonly used words, though it breaks down for the rarer, subject-specific words. For instance, this document uses the word "probability" far more than a typical document in the British National Corpus (BNC). Consequently, if you included this text in the BNC, and happened to pick "probability" once, you should be less surprised if you pick the same word again from the same document. From the first instance, you can deduce that you've probably found a math or computational linguistics paper, which will likely be enriched in words like "probability," "sample," "event," and "corpus."

This kind of subject-specific probability is what allows us to classify documents by subject and to search on the web. It's also critical to spam filters. In addition to being subject-specific, word choices are also author-specific. The fact that different authors prefer to use different words to express the same ideas sometimes allows us to attribute an unknown document to a known author, if we have a collection of works to which we can compare it.

We also know that documents are approximately translationally invariant; this means that the probability distributions don't depend strongly on the absolute position inside the document. The probability that the 100^{th} character is "t" is about the same as that probability for the 200^{th} character.

[10] Except, perhaps, the rarest characters. If you see an "x", it's possible that the document is about xylophones or King Xerxes, in which case you might expect extra "x"s to exist in an exaggerated number of places, as the exact same nouns will recur. This is an extreme example, admittedly. Most of the time the admixture of rare letters in a text remains small, and it is acceptable to make the approximation that King Xerxes, even if he is mentioned once, will not dominate the extracted text.

[11] *I.e.* the probability that the letter we choose is an "a", a "b", *etc.*

4.2 Engineering Practise

Real systems will grossly violate the assumptions of independence. Systems will often use overlapping 3-grams or 4-grams, and treat them as independent events. When they proceed to use Bayes' Theorem and perform poorly, no one should be surprised. This is known in the trade as a "naive Bayes classifier," and is often used as a comparison point for different algorithms.

As an example, in the Bell Labs speech recognition system, overlapping 30 millisecond samples of the speech waveform were processed, then treated as independent events to determine the phoneme sequence. Even non-overlapping 30 ms sequences would be highly correlated if they were within 100 ms of each other. The output of the ASR system was a list of the most likely phoneme sequences, along with their probabilities, obtained from Bayes' Theorem. In this list, it wasn't rare for the correct phoneme sequence to be the second or third best item on the list, with a calculated probability of 10^{-20}.

Now, if the probability of such a mistake were *really* 10^{-20}, one would never see it happen, not even if every computer on Earth were an ASR system. The answer was that the strong correlations in the data made many of the Bayes factors close to one in reality, while the software didn't take that into account. If, hypothetically, only every 10^{th} frame were effectively independent, then the real probability would be about the 10^{th} root of the reported probability. One would indeed expect every 10^{th} frame to contain new information, because that spacing corresponds roughly to the length of a phoneme. The frames in between are more-or-less just duplicates. And if we make that (huge) correction, we get a much more plausible answer, that the error probability is on the order of a percent, rather than one in a hundred million million million.[12]

Doubtless, this bad assumption causes problems, but no one knows how to fix it properly while keeping the CPU and memory consumption of the ASR system within reasonable bounds.

Thanks to Peet Morris for comments and suggestions.

References

Jackman, S. (pre-2004a). *Bayes Factor*. Stanford University, http://jackman.stanford.edu/papers/bayesfactor.pdf. Apparently written for the *Encyclopedia of Research Methods for the Social Sciences*.

Jackman, S. (pre-2004b). *Bayes Factor*. Stanford University department of Political Science, http://jackman.stanford.edu/papers/bayestheorem.pdf. Apparently written for the *Encyclopedia of Research Methods for the Social Sciences*.

[12] The actual error probability was more like 30%.