Understanding Logistic Regression Coefficients

towardsdatascience.com/understanding-logistic-regression-coefficients-7a719ebebd35

December 15, 2019

Ravi Charan
Dec 14, 2019

11 min read

Or a better way to think about probability in terms of evidence

Logistic Regression suffers from a common frustration: the coefficients are hard to interpret. If you've fit a Logistic Regression model, you might try to say something like "if variable X goes up by 1, then the probability of the dependent variable happening goes up by ???" but the "???" is a little hard to fill in.

The trick lies in changing the word "probability" to "**evidence**." In this post, we'll understand how to quantify evidence. Using that, we'll talk about how to **interpret Logistic Regression coefficients**.

Finally, we will briefly discuss **multi-class** Logistic Regression in this context and make the connection to Information Theory.

This post assumes you have some experience interpreting *Linear* Regression coefficients and have seen Logistic Regression at least once before.

Part 1: Two More Ways to Think about Probability

Odds and Evidence

We are used to thinking about probability as a number between 0 and 1 (or equivalently, 0 to 100%). But this is just a particular mathematical representation of the "degree of plausibility."

There is a second representation of "degree of plausibility" with which you are familiar: odds ratios. For example, if I tell you that "the odds that an observation is correctly classified is 2:1", you can check that the probability of correct classification is two thirds. Similarly, "even odds" means 50%.

My goal is convince you to adopt a third: the **log-odds**, or the logarithm of the odds. For interpretation, we we will call the log-odds the **evidence.** This follows E.T. Jaynes in his post-humous 2003 magnum opus Probability Theory: The Logic of Science.

In general, there are two considerations when using a mathematical representation. First, it should be interpretable. Second, the mathematical properties should be convenient.

Interpreting Evidence: Measuring in Hartleys

In order to convince you that evidence is interpretable, I am going to give you some numerical scales to calibrate your intuition.

First, evidence can be measured in a number of different units. We'll start with just one, the Hartley. The Hartley has many names: Alan Turing called it a "ban" after the name of a town near Bletchley Park, where the English decoded Nazi communications during World War II. It is also called a "dit" which is short for "decimal digit."

The formula to find the evidence of an event with probability p in Hartleys is quite simple:

$$\text{Evidence} = \log_{10} \text{odds}$$

Computing the evidence, in Hartleys

Where the odds are p/(1-p). This is much easier to explain with the table below. Note that judicious use of rounding has been made to make the probability look nice. With this careful rounding, it is clear that 1 Hartley is approximately "1 nine."

Evidence (Hartleys)	Odds For:Against	Probability (rounded)
$\log_{10} \frac{p}{1-p}$	$\frac{p}{1-p}$	p
-3	1:1000	0.0001
-2	1:100	0.001
-1	1:10	0.09
0	1:1	0.5
1	10:1	0.9
2	100:1	0.99
3	1000:1	0.999

Table of Evidence, Odds, and Probability

Notice that 1 Hartley is quite a bit of evidence for an event. A more useful measure could be a tenth of a Hartley. A "deci-Hartley" sounds terrible, so more common names are "**deciban**" or a **decibel**. Here is another table so that you can get a sense of how much information a deciban is. Hopefully you can see this is a decent scale on which to measure evidence: not too large and not too small.

Evidence (decibans)	Odds For:Against	Probability (rounded)
$10\log_{10} \frac{p}{1-p}$	$\frac{p}{1-p}$	p
0	1:1	0.5
1	1.3:1	0.56
3	2.0:1	0.67
5	3.2:1	0.76
7	5.0:1	0.83
9	7.9:1	0.89
10	10:1	0.91

Using Evidence: Bayes' Rule

I also said that evidence should have convenient mathematical properties. It turns out that evidence appears naturally in Bayesian statistics.

Suppose we wish to classify an observation as either True or False. We can write:

$$P(\text{True}|\text{Data}) = \frac{P(\text{Data}|\text{True})P(\text{True})}{P(\text{Data})}$$

$$P(\text{False}|\text{Data}) = \frac{P(\text{Data}|\text{False})P(\text{False})}{P(\text{Data})}$$

Bayes' Law for Binary Classification

In Bayesian statistics the left hand side of each equation is called the "posterior probability" and is the assigned probability **after** seeing the data. The P(True) and P(False) on the right hand side are each the "prior probability" from **before** we saw the data. We think of these probabilities as states of belief and of Bayes' law as telling us how to go from the prior state of belief to the posterior state. If you don't like fancy Latinate words, you could also call this "**after** ← **before**" beliefs.

More on what our prior ("before") state of belief was later. The standard approach here is to compute each probability. This is a bit of a slog that you may have been made to do once. The slick way is to start by considering the **odds**. If we divide the two previous equations, we get an equation for the "posterior odds."

$$\frac{P(\text{True}|\text{Data})}{P(\text{False}|\text{Data})} = \frac{P(\text{Data}|\text{True})}{P(\text{Data}|\text{False})}\frac{P(\text{True})}{P(\text{False})}$$

The Posterior Odds

And then we will consider the **evidence** which we will denote Ev. So Ev(True) is the prior ("before") evidence for the True classification. And Ev(True|Data) is the posterior ("after"). We get this in units of Hartleys by taking the log in base 10:

$$\log_{10}\frac{P(\text{True}|\text{Data})}{P(\text{False}|\text{Data})} = \log_{10}\frac{P(\text{Data}|\text{True})}{P(\text{Data}|\text{False})} + \log_{10}\frac{P(\text{True})}{P(\text{False})}$$

$$\underbrace{\text{Ev}(\text{True}|\text{Data})}_{\text{posterior evidence}} = \underbrace{\log_{10}\frac{P(\text{Data}|\text{True})}{P(\text{Data}|\text{False})}}_{\text{new evidence}} + \underbrace{\text{Ev}(\text{True})}_{\text{prior evidence}}$$

The Data Science process, quantified

In the context of binary classification, this tells us that we can interpret the Data Science process as: collect data, then add or subtract to the evidence you already have for the hypothesis. By quantifying evidence, we can make this quite literal: you add or subtract the amount!

Other Unit Systems

There are three common unit conventions for measuring evidence. We have met one, which uses Hartleys/bans/dits (or decibans etc.). This choice of unit arises when we take the logarithm in base 10.

The next unit is "nat" and is also sometimes called the "nit." It can be computed simply by taking the logarithm in base *e*. Recall that $e \approx 2.718$ is Euler's Number.

The final common unit is the "bit" and is computed by taking the logarithm in base 2. It is also sometimes called a Shannon after the legendary contributor to Information Theory, Claude Shannon.

Until the invention of computers, the Hartley was the most commonly used unit of evidence and information because it was substantially easier to compute than the other two. (Note that information is slightly different than evidence; more below.)

With the advent computers, it made sense to move to the bit, because information theory was often concerned with transmitting and storing information on computers, which use physical bits.

Finally, the natural log is the most "natural" according to the mathematicians. For this reason, this is the default choice for many software packages. It is also common in physics.

I believe, and I encourage you to believe:

- The Hartley or deciban (base 10) is the most interpretable and should be used by Data Scientists interested in quantifying evidence.
- The nat should be used by physicists, for example in computing the entropy of a physical system.

Note, for data scientists, this involves converting model outputs from the default option, which is the nat.

Finally, here is a unit conversion table. I have empirically found that a number of people know the first row off the top of their head. The 0.69 is the basis of the Rule of 72, common in finance. The $3.01 \approx 3.0$ is well known to many electrical engineers ("3 decibels is a doubling of power").

bit	nat	deciban
1	0.69	3.0
1.44	1	4.3
0.33	0.23	1

Unit Conversion Table for Evidence

Converting Evidence to Odds and Probability

Having just said that we should use decibans instead of nats, I am going to do this section in nats so that you recognize the equations if you have seen them before. Let's denote the evidence (in nats) as *S*. The formula is:

$$S = \underbrace{\mathrm{Ev}(\mathrm{True})}_{\text{in nats}} = \ln \frac{P(\mathrm{True})}{P(\mathrm{False})}$$

Formula for the Evidence

Let's say that the evidence for True is S. Then the odds and probability can be computed as follows:

$$\text{odds} = e^S : 1$$

$$P(\mathrm{True}) = \frac{e^S}{1+e^S} \qquad P(\mathrm{False}) = \frac{1}{1+e^S}$$

Converting evidence S to odds or a probability

If the last two formulas seem confusing, just work out the probability that your horse wins if the odds are 2:3 against. You will first add 2 and 3, then divide 2 by their sum.

Part 2: Understanding Logistic Regression

If you believe me that evidence is a nice way to think about things, then hopefully you are starting to see a very clean way to interpret logistic regression. First, remember the logistic sigmoid function:

$$\sigma(S) = \frac{1}{1+e^{-S}} = \frac{e^S}{1+e^S}$$

Hopefully instead of a complicated jumble of symbols you see this as the function that converts information to probability. It's exactly the same as the one above!

Let's treat our dependent variable as a 0/1 valued indicator. So 0 = False and 1 = True in the language above. The logistic regression model is

$$\mathbb{E}[y_i] = P(y_i = 1) = \sigma(\beta^T X_i)$$

Where X is the vector of observed values for an observation (including a constant), β is the vector of coefficients, and σ is the sigmoid function above.

This immediately tells us that we can interpret a coefficient as the amount of evidence provided per change in the associated predictor.

For example, suppose we are classifying "**will it go viral or not**" for online videos and one of our predictors is the number minutes of the video that have a cat in it ("cats").

- If the coefficient of this "cats" variable comes out to 3.7, that tells us that, for each increase by one minute of cat presence, we have 3.7 more nats (16.1 decibans) of evidence towards the proposition that the video will go viral.
- Add up all the evidence from all the predictors (and the prior evidence — see below) and you get a total score.
- Classify to "True" or 1 with positive total evidence and to "False" or 0 with negative total evidence. But more to the point, just look at how much evidence you have!

Miscellany

A few brief points I've chosen not to go into depth on.

1. The inverse to the logistic sigmoid function is the given above. Many authors define logistic regression in terms of the logit. Where the logistic function converts evidence into probabilities, its inverse converts probabilities into evidence.Also — as usual, mathematics is done in units of nats but you are of course free to use a different base for the logarithm if you want a different unit.

2. The is nothing to be afraid of. By default, you chose the prior of “no evidence either way” in other words, 0 evidence. Hopefully this seems reasonable. Changing your prior is equivalent to changing the threshold for classification. This is a good way to think about how an is constructed.

$$\text{logit}(p) = \ln \frac{p}{1-p}$$

The logit function is the inverse of the logistic function

3. You can check that the (also called the log-loss or deviance) may be described as follows. Let evidence be given by the model in favor of the wrong prediction. Then, in the limit as is large, the loss is . Conversely, if is the evidence given in favor of the correct prediction, then, in the limit as is large, the cross-entropy loss is .

Part 3: Multi-class logistic regression

Given the discussion above, the intuitive thing to do in the multi-class case is to **quantify the information** in favor of each class and then (a) **classify** to the class with the most information in favor; and/or (b) **predict probabilities** for each class such that the log odds ratio between any two classes is the difference in evidence between them.

We can achieve (b) by the softmax function. The probability of observing class k out of n total classes is:

$$P(y_i = k) = \frac{e^{S_k}}{e^{S_1} + e^{S_2} + \cdots + e^{S_n}}$$

Softmax: Probability of observing class k out of n possibilities given the information in favor of each

Dividing any two of these (say for k and ℓ) gives the appropriate log odds.

How do we estimate the information in favor of each class? There are two apparent options:

1. () Notice that, mathematically, shifting the whole list of information in favor of each class of by some constant number of Hartleys doesn’t change the probability distribution. This is because we only care about the differences in information between classes.So, we might as well pick a class, say class ⋆, and set its information to 0. Then estimate the evidence for each other class relative to class ⋆.
2. () For each class, say class k, run a simple logistic regression (binary classification) for “is the observation class k or not.”

In the case of n = 2, approach 1 most obviously reproduces the logistic sigmoid function from above. Approach 2 turns out to be equivalent as well.

Warning: for n > 2, these approaches are . (The good news is that the choice of class $\star$ in option 1 does not change the results of the regression.)

I am not going to go into much depth about this here, because I don't have many good references for it. If you want to read more, consider starting with the scikit-learn documentation (which also talks about 1v1 multi-class classification). If you have/find a good reference, please let me know! The point here is more to see how the evidence perspective extends to the multi-class case.

Part 4: Information Theory

This will be very brief, but I want to point towards how this fits towards the classic theory of Information. Information Theory got its start in studying how many bits are required to write down a message as well as properties of sending messages. In 1948, Claude Shannon was able to derive that the information (or entropy or surprisal) of an event with probability p occurring is:

$$I(p) = -\log(p)$$

Given a probability distribution, we can compute the expected amount of information per sample and obtain the entropy S:

$$S = -\Sigma p_i \log p_i$$

where I have chosen to omit the base of the logarithm, which sets the units (in bits, nats, or bans). Physically, the information is realized in the fact that it is impossible to losslessly compress a message below its information content.

The connection for us is somewhat loose, but we have that in the binary case, the evidence for True is

$$\mathrm{Ev}(\mathrm{True}) = -\left[I(\mathrm{True}) - I(\mathrm{False})\right]$$

The negative sign is quite necessary because, in the analysis of signals, something that always happens has no surprisal or information content; for us, something that always happens has quite a bit of evidence for it.

Conclusion

> Information is the resolution of uncertainty– *Claude Shannon*

Probability is a common language shared by most humans and the easiest to communicate in. But it is not the best for every context. In this post:

- We saw that evidence is simple to compute with: you just add it;

- we calibrated your sense for “a lot” of evidence (10–20+ decibels), “some” evidence (3–9 decibels), or “not much” evidence (0–3 decibels);
- we saw how evidence arises naturally in interpreting logistic regression coefficients and in the Bayesian context; and
- we saw how it leads us to the correct considerations for the multi-class case

I hope that you will get in the habit of converting your coefficients to decibels/decibans and thinking in terms of evidence, not probability.

–Ravi

Reference/Recommendation

I highly recommend E.T. Jaynes’ book mentioned above.

For context, E.T. Jaynes is what you might call a militant Bayesian.

- The perspective of “evidence” I am advancing here is attributable to him and, as discussed, arises naturally in the Bayesian context.
- Another great feature of the book is that it derives (!!) the laws of probability from qualitative considerations about the “degree of plausibility.” I find this quite interesting philosophically.

Also: there seem to be a number of pdfs of the book floating around on Google if you don’t want to get a hard copy.