What is information gain, and what does it have to do with data science math?

A brief investigation by Joy Payton

My work below relies heavily on the University of Alberta description of information gain found at http://webdocs.cs.ualberta.ca/~aixplore/learning/DecisionTrees/InterArticle/4-DecisionTree.html and on Linda Shapiro's notes on information gain found at http://homes.cs.washington.edu/~shapiro/EE596/notes/InfoGain.pdf.

What is information gain?

A document by the University of Alberta gives a good intuitive introduction to information gain. Using the example of the game "20 questions", the authors suggest that skilled players of the game ask questions that split the remaining possibilities effectively, such that as few additional questions as possible are required to narrow down the object being sought. Their thought experiment assumes that questions must be limited to yes/no.

> Think of playing "20 questions": I am thinking of an integer between 1 and 1,000 -- what is it? What is the first question you would ask?
>
> You could ask "Is it 752?" Or "Is it a prime number between 123 and 239?". Most people, however, would first ask "Is it between 1 and 500?"
>
> Why? Because this answer provides the most information: It typically makes sense to ask a question that "splits" the remaining options in half, whether the response is "Yes" or "No".

Information gain refers to the measure of the usefulness of any splitting criteria. In the example above, asking as one's first question, "is the number 23?" would be extremely unlikely to provide a significant narrowing of the possibilities that remain, and would have low information gain. The goal of machine learning is to classify data accurately and efficiently, so finding the right questions or criteria, those which have high information gain, is an important component of a good classification algorithm. This concept is not unique to machine learning, but to efficient decision making generally. An experienced physician, for example, has learned to ask a short series of questions that allow her to quickly narrow her diagnostic options and hone in on the right decisions for her patient. Increasingly, computer adaptive testing is being used in standardized tests to allow for more accurate classification of test-takers' abilities with fewer questions being required.

What are the mathematical concepts that underly information gain? First, we can imagine the concept of entropy and purity. A pure dataset contains a single class, while an impure data set has a mix of more than one class. Entropy is the measure of the degree of impurity. The mathematical model for entropy is $\sum_i -p_i \log_2 p_i$, where p_i is the proportion of the class i in

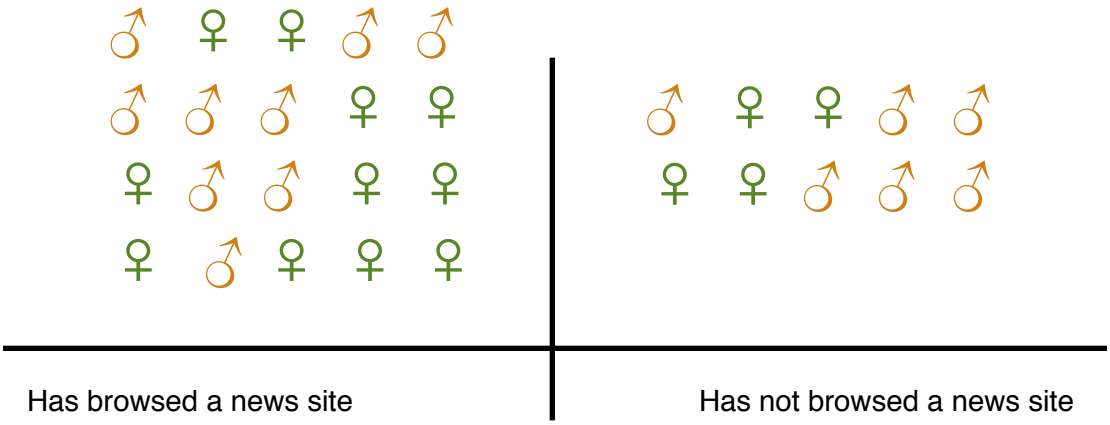
the given set (the probability of i). In order to separate classes, we want entropy to decrease after we split a set on a given criteria. (As a side note, we want entropy to be high in a training set, because it will allow an algorithm to be stressed to its highest degree, so that really good criteria emerge).

Example

As an example (I rely heavily in this section on the excellent example by Shapiro), let's take a group of internet users, which are distributed evenly among men and women. I am an advertiser who wants to determine how to easily tell from someone's browsing history if they are female, as I want to advertise my latest fashion line for women, and don't want to waste money putting ads on the screens of men. The goal is to start with my set that has high entropy (a heterogeneous mix of both sexes) and end with a set that has low entropy (mostly or all women). How do I choose the criteria by which to decide who's who?

I begin with a known set and try different questions or criteria, and measure the entropy-reduction success of each one. We'll just do two in this simple example.

Let's take the (entirely fictional, as all this example is) case of "has browsed a news site in this session." My group of 30 (50% male and 50% female) splits as shown below:



Is this a good choice for a classification criteria? Just by eyeballing the example, we can see that it doesn't seem too useful, but we can quantify this by calculating the entropy of the initial set and the resultant, split sets, and find our information gain.

Our initial (parent) set contains 15 men and 15 women, so the we calculate its entropy:

entropy for men: $-0.5(\log_2 0.5) = -0.5(-1) = 0.5$
entropy for women: $-0.5(\log_2 0.5) = -0.5(-1) = 0.5$
total entropy = 1

Now let's calculate the entropy of the two sets that result from our "browsed a news site" criterion:

Has browsed a news site entropy (men):

$$-\frac{9}{20}(\log_2 \frac{9}{20}) = -0.45(\log_2 0.45) = -0.45(-1.152) = 0.70$$

Has browsed a news site entropy (women):

$$-\frac{11}{20}(\log_2 \frac{11}{20}) = -0.55(\log_2 0.55) = -0.55(-0.862) = 0.47$$

Total entropy for "has browsed a news site": 1.17

Has not browsed a news site entropy (men):

$$-\frac{6}{10}(\log_2 \frac{6}{10}) = -0.6(\log_2 0.6) = -0.6(-0.737) = 0.44$$

Has not browsed a news site entropy (women):

$$-\frac{4}{10}(\log_2 \frac{4}{10}) = -0.4(\log_2 0.4) = -0.4(-1.32) = 0.53$$

Total entropy for "has browsed a news site": 0.97

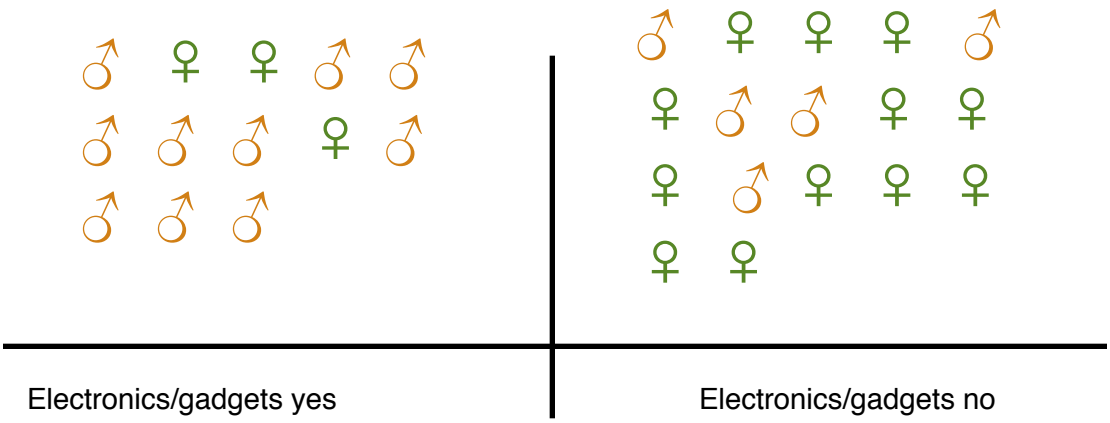
To combine these two values for two "child" sets, we weigh them according to their relative population. The "has" set contains 20 members, while the "has not" contains 10, so our calculation for total entropy for both sets is:

$$\left(\frac{20}{30} \bullet 1.17\right) + \left(\frac{10}{30} \bullet 0.97\right) = 0.78 + 0.32 = 1.10$$

Our information gain is simply how much more or less entropy was introduced by our split, so we subtract the total entropy of the child sets from the entropy of the parent set:

1 - 1.10 = -0.10 . Our information gain is actually negative… we have increased entropy!

Let's consider instead a more helpful criterion: "has browsed an electronics or gadgets online store in this browsing session". We continue to use the same 30 person parent set, composed of half men and half women.



Let's calculate the entropy of the two sets that result:

Has browsed an electronics / gadgets online store entropy (men):

$$-\frac{10}{13}(\log_2 \frac{10}{13}) = -0.77(\log_2 0.77) = -0.77(-0.37707) = 0.29$$

Has browsed an electronics / gadgets online store entropy (women):

$$-\frac{3}{13}(\log_2 \frac{3}{13}) = -0.23(\log_2 0.23) = -0.23(-2.120294) = 0.49$$

Total entropy for "has browsed a news site": 0.78

Has not browsed a news site entropy (men):

$$-\frac{5}{17}(\log_2 \frac{5}{17}) = -0.29(\log_2 0.29) = -0.29(-1.785875) = 0.52$$

Has not browsed a news site entropy (women):

$$-\frac{12}{17}(\log_2 \frac{12}{17}) = -0.71(\log_2 0.71) = -0.71(-0.494109) = 0.35$$

Total entropy for "has browsed a news site": 0.87

To combine these two values for two "child" sets, we weigh them according to their relative population. The "has" set contains 20 members, while the "has not" contains 10, so our calculation for total entropy for both sets is:

$$\left(\frac{13}{30} \bullet 0.78\right) + \left(\frac{17}{30} \bullet 0.87\right) = 0.338 + 0.493 = 0.831$$

Our information gain is 1 - 0.831 = .169 . Our information gain is greater than our last attempt, and therefore we have a mathematical basis for supporting the use of this criteria instead of the “news site” criteria above.

A question I am left with at this point is whether the logarithm is binary because of the number of child sets resulting from a criteria. Would it be a log base 3 if I were to split a parent set into three child sets (one thinks of the traditional “animal, vegetable, or mineral” question)? Or is a binary split simply better and more efficient in machine learning?