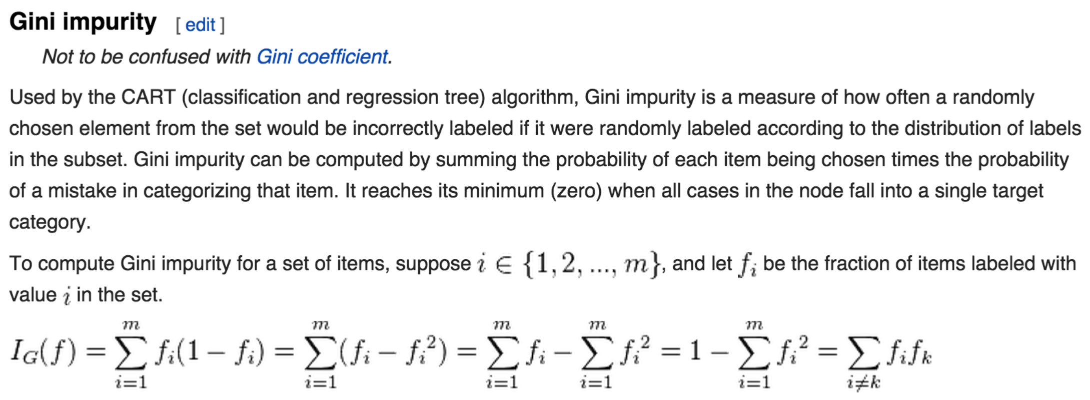
From wikipedia: https://en.wikipedia.org/wiki/Decision_tree_learning



Gini impurity [edit]

Not to be confused with Gini coefficient.

Used by the CART (classification and regression tree) algorithm, Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. Gini impurity can be computed by summing the probability of each item being chosen times the probability of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

To compute Gini impurity for a set of items, suppose $i \in \{1, 2, ..., m\}$, and let f_i be the fraction of items labeled with value i in the set.

$$I_G(f) = \sum_{i=1}^{m} f_i(1-f_i) = \sum_{i=1}^{m}(f_i - {f_i}^2) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} {f_i}^2 = 1 - \sum_{i=1}^{m} {f_i}^2 = \sum_{i \neq k} f_i f_k$$

I am unable to get my head around two of the steps:

1. The first equation: $f_i(1-f_i)$. This does not immediately become apparent as the "probability of being chosen times the probability of miscategorization". Instead it looks to me like "probability of being chosen times the probability of *others* being chosen" (but not necessarily incorrectly)
2. The arithmetic of the last simplification eludes me: how to get from $1-\sum(f_i^2)$ to $\sum(f_i f_k)$

Tips appreciated.

gini

share cite improve this question

edited Oct 1 '15 at 19:47
user30490

asked Oct 1 '15 at 18:59
javadba
232 2 10

add a comment

3 Answers

active oldest votes

4

I don't know about the algebra, but you can prove the identity with a probabilistic argument. If I roll two dice with m sides and the probability of side i is f_i, then the probability of a double is $\sum f_i^2$. Thus $1-\sum f_i^2$ is the probability that I roll distinct values. But arguing differently, the probability, say, that I get i followed by j is $f_i f_j$. Summing over all possibilities, with $i \neq j$, I get the probability of rolling distinct outcomes: $\sum f_i f_j$, and the identity is proven.

As for the first point, If you role the m sided die, there is a probability f_i that side $i$comes up. Suppose I have to guess the value, and I do this by rolling a die of my own with the same weights. The probability that I guess wrong, conditional on value i being true, is $1-f_i$. The probability that I get it wrong, summing over the possible values, is $\sum f_i(1-f_i)$.

share cite

edited Oct 1 '15 at 20:45

answered Oct 1 '15 at 20:40

improve this answer

Placidia
11.2k 5 26 51

nice explanations. I feel closer to an understanding. – javadba Oct 1 '15 at 22:05

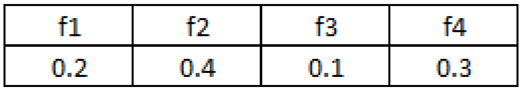
add a comment

4

I think it's best to answer your question in reverse order as we'll back into your first question by answering your second.

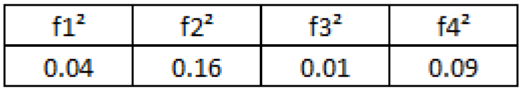
Question 2

Imagine you have a probability distribution function ($f_i$fi) that distributes its probabilities as such:



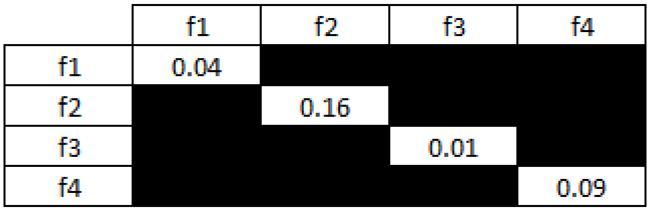
f1	f2	f3	f4
0.2	0.4	0.1	0.3

I can then square the probabilities (f_i^2fi2) and get:



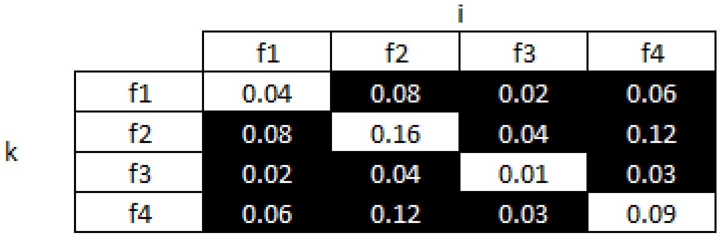
$f1^2$	$f2^2$	$f3^2$	$f4^2$
0.04	0.16	0.01	0.09

Another way of looking at it is putting each probability distribution along the axis of a grid. Each cell now represents the product of the function along the respective axes.



	f1	f2	f3	f4
f1	0.04	■	■	■
f2	■	0.16	■	■
f3	■	■	0.01	■
f4	■	■	■	0.09

The grid itself sums to 1, just like you'd see in a two dice roll probability table. It should be clear that 1 minus the sum of the diagonal probabilities is the same as the non-highlighted squares below.

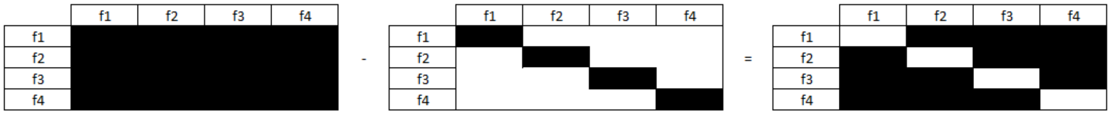


i

k

	f1	f2	f3	f4
f1	0.04	0.08	0.02	0.06
f2	0.08	0.16	0.04	0.12
f3	0.02	0.04	0.01	0.03
f4	0.06	0.12	0.03	0.09

If we call one of the axis k to differentiate it, but still have it render the same function, we can then make the statement.



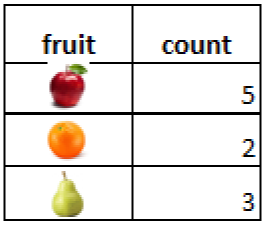
$1-\sum f_i^2$ 1−∑fi2 = $\sum_{i\neq k} f_i f_k$ ∑i≠kfifk

Question 1

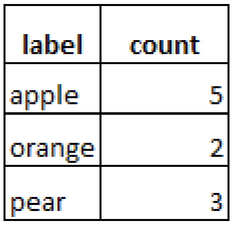
We can now use some of the intuition from answering question 2 to drive the intuition for question 1.

Let's take our same table from question 2, but change what the two axes mean. Across one axis we'll have labels for objects, while on the other we'll have the actual object.

For a concrete example, let's assume we have a bowl of fruit: apples, oranges and pears. In another bowl we'll have labels corresponding to apples, oranges and pears in the same proportion as the actual objects.

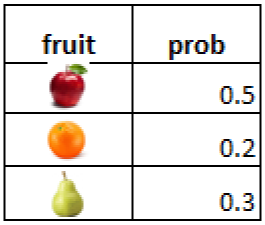


fruit	count
🍎	5
🍊	2
🍐	3

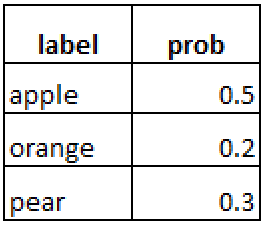


label	count
apple	5
orange	2
pear	3

If we then look at the probability of choosing each at random we get the following distribution.

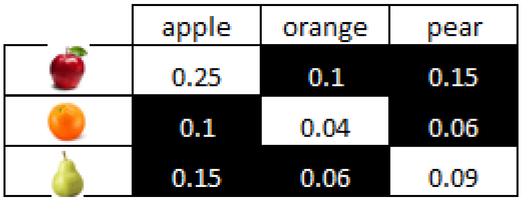


fruit	prob
🍎	0.5
🍊	0.2
🍐	0.3



label	prob
apple	0.5
orange	0.2
pear	0.3

Now we want to look at the joint distribution. The Geni impurity tells us the probability that we select an object at random and a label at random and it is an incorrect match. The Geni impurity is the sum of the probabilities in the black shaded areas. These are where the label does not match the object, thus the impurity.



	apple	orange	pear
🍎	0.25	0.1	0.15
🍊	0.1	0.04	0.06
🍐	0.15	0.06	0.09

This should look very familiar to the answer to question 2. If the explanation for question 2 convinced you that $1-\sum f_i^2$ 1−∑fi2, you should be able to work backwards through the algebra you provided to see that also equals $\sum f_i(1-f_i)$