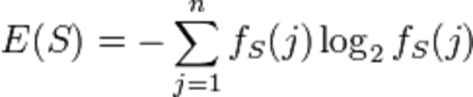
How to decide which data to split on, calculating information gain and entropy

|  |  |  |  |
| --- | --- | --- | --- |
| W1 | W2 | Remainder | Class |
| We | are | all about the city | N |
| What | a | vulgar, self-publicising woman she is | N |
| What | a | wonderful end to the day: the hunter was home from the hill | N |
| What | we | want is Watney’s | N |
| What | are | some of the false and real causes of male hair loss | Q |
| What | this | one does is fantastic | N |
| What | did | you do when your card was stolen | Q |
| What | did | you achieve in your last role | Q |
| When | are | the quarterly estimated tax returns due | Q |
| When | I | get a spare moment all I can think about is my warm, cosy bed | N |
| When | I | went to college, he was still in high school | N |
| Where | a | parent might not be sometimes a teacher will be listened to | N |
| Where | are | the enlarged group's headquarters | Q |
| Where | can | I find an answer to a query I have about a cbbc programme | Q |
| Where | I | would really like to go is Urbino in Italy | N |
| You | did | Run faster when you were younger | N |

Calculate entropy of original data set



*fs*(Q) = 6/16 = 0.375 *, fs*(N) = 10/16 = 0.625 (probabilities)

Entropy of original dataset = -(6/16)xlog2(6/16) + -(10/16)log2(10/16)

= 0.954434

Calculating for W1

|  |  |  |  |
| --- | --- | --- | --- |
| W1 | Occurrences | Class q | Class n |
| We | 1 | 0 | 1 |
| What | 7 | 3 | 4 |
| When | 3 | 1 | 2 |
| Where | 4 | 2 | 2 |
| you | 1 | 0 | 1 |

Calculate entropy of we

**We**

0 questions – contributes -0/1 x log2(0/1) to entropy

1 non-question – contributes -1/1 x log2(1/1) to entropy

Questions 0/1 = 0, log2(0) – not calculable, therefore this term set to 0.

Non-questions 1/1=1, log2(1/1) = 0

Therefore **We** contributes 0 entropy

Calculate entropy of what

7 occurrences

3 questions – contributes -3/7 x log2(3/7) to entropy

4 non-question – contributes -4/7 x log2(4/7) to entropy

Questions -3/7 x log2(3/7) = 0. 52388246628

Non-questions -4/7 x log2(4/7)=0. 46134566974

Therefore **what** contributes 7/16(0.52388246628 + 0.46134566974) = 0.4310373095

Calculate entropy of when

3 occurrences

1 questions – contributes -1/3 x log2(1/3) to entropy

2 non-question – contributes -2/3 x log2(2/3) to entropy

Questions -1/3 x log2(1/3) = 0.52832083357

Non-questions -2/3 x log2(2/3) = 0. 38997500048

Therefore **when** contributes 3/16(0.52832083357 + 0. 38997500048) = 0.17218046888

Calculate entropy of where

4 occurrences

2 questions – contributes -2/4 log2(2/4)

2 non-question – contributes -2/4 log(2/4)

Questions: -2/4 log2(2/4) = 0.5

Non-questions -2/4 log2(2/4) = 0.5

Therefore **where** contributes 4/16(1) = 4/16 = 0.25

Calculate entropy of you

1 occurrence

0 questions – contributes – not calculable

1 non-question – contributes -1/1log2(1/1)

Questions:

Non-questions 1/1log2(1/1) = 0

Therefore **you** contribute: 1/16 \* 0 = 0

Not to calculate information gain from splitting on w1 is…

Add them all together

0.4310373095 + 0 + 0.17218046888 + 0.25 + 0 = 0.85321777838

And then minus that answer from the original entropy of the data

0.954434 - 0.85321777838 = 0.10121622162

And that is the information gain on splitting on W1 = 0.10121622162

So we now need to work out W2 in the same way, and see which is best

**Calculate W2(we are doing it for you)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| W2 | Occurances | Class q | Class n | Entropy |
| Are | 4 | 3 | 1 | 0.202819351 |
| A | 3 | 0 | 3 | 0 |
| We | 1 | 0 | 1 | 0 |
| This | 1 | 0 | 1 | 0 |
| did | 3 | 2 | 1 | 0.172180469 |
| I | 3 | 0 | 3 | 0 |
| can | 1 | 1 | 0 | 0 |

Original entropy 0.954434

Information gain = 0. 954434 – (0.202819351 + 0.172180469) =

0.954434 – 0.375 = 0.579434

So we have calculated both W1 and W2, information gain is as follow:

W1: 0.101

W2: 0.579

Therefore split on W2 first and then W1 if you need to

e.g.



The problem with using this calculation is that it favours splitting on nodes with many options

We can use Gain Ratio to get round this

**Calculate Splitin(w1)**



Splitin(w1) = -1/16x log2(1/16) - 6/16 x log2(6/16) . . . . etc.

Calculate gain ratio (w1)



Now calculate the gain ratio for w2

Which gain ratio is higher (w1 or w2)? Which information gain is higher (w1 or w2)?