Computer Science
University of Warwick

Exercise 1: Calculating Information Gains

Module CS909 March 8, 2019

1 Introduction

In this exercise you will get some practice at calculating measures of attribute importance. The work is based on the material in Lecture 5: Classification - Decision Trees.

At the end of this exercise you should be able to:

- compute class probabilities for decision tree branches;
- calculate the entropy of a discrete probability distribution;
- calculate information gain.

These measures are used in decision tree classifiers to choose the best attribute to split on and also for attribute selection (e.g. choosing the input attributes with the highest information gain).

2 Theory

The theory was covered in Lecture 5, but here is a quick recap of the basic ideas. Note that you are expected to memorise these formulae for the exam.

Entropy Consider the probability distribution $(p_1, p_2, \ldots, p_n)$, where p_i represents the probability of the ith outcome so $0 \leq p_i \leq 1$ and $\sum_{i=1}^{n} p_i = 1$. The *entropy* (or information content) of this distribution (measured in bits) is defined to be

$$\text{entropy}(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i = -p_1 \log_2 p_1 - p_2 \log_2 p_2 \cdots - p_n \log_2 p_n, \tag{1}$$

Note that $\log_2 1 = 0$, $\log_2 0$ is undefined, but $0 \log_2 0 = 0$. Entropy must always be non-negative; if your calculation ever gives a result that is negative it is *wrong*!

In general, the information content of a set of positive integers $a_1, a_2, \ldots, a_n$ with sum A is given by

$$\text{info}([a_1, a_2, \ldots, a_n]) = \text{entropy}\left(\frac{a_1}{A}, \frac{a_2}{A}, \ldots, \frac{a_n}{A}\right). \tag{2}$$

This means that we first turn the set of numbers into a set of probabilities by dividing each value by the sum of all of them and then compute the entropy of the distribution. There is a neat trick to reduce the amount of work that depends on the properties of the log function.

$$\text{info}([a_1, a_2, \ldots, a_n]) = \frac{[-a_1 \log_2 a_1 - a_2 \log_2 a_2 - \cdots - a_n \log_2 a_n + A \log_2 A]}{A}. \tag{3}$$

Information Gain of an attribute is the difference between the information before and after splitting the data on that attribute. Formally

$$\text{Information Gain} = (\text{Information before split}) - (\text{Information after split}). \tag{4}$$

Suppose we split at a node p into k partitions, the entropy after split is:

$$\text{entropy}_{\text{after split}} = \left(\sum_{i=1}^{k} \frac{n_i}{n} \text{entropy}(i) \right), \tag{5}$$

where n_i is the number of examples in child i, and n is the number of examples at parent node p.

That is, we first calculate the entropy value of each child node and then compute the weighted sum of these entropy values, with the weights given by the number of examples in the corresponding branch divided by the total number of examples at the parent node.

Finally the information gain is calculated as the difference of entropy at the parent node p and the weighted sum of the entropy values of its child nodes:

$$Gain = \text{entropy}(p) - \text{entropy}_{\text{after split}}. \tag{6}$$

3 Worked Example

These formulae look more daunting than they are to work with. A few examples should clarify matters. We are going to use the following dataset which represents information about a cricket team (form, number of key players present, ground where the match was played, whether they batted first) for predicting the result.

Form	Key Players	Ground	Bat First	Match Result
Variable	one	B	yes	Won
Good	one	B	yes	Won
Variable	neither	B	no	Won
Poor	one	B	no	Lost
Poor	neither	B	yes	Lost
Good	both	B	yes	Lost
Poor	one	A	yes	Won
Poor	one	A	no	Won
Poor	neither	A	no	Won
Good	neither	A	no	Won
Variable	both	A	yes	Won
Variable	both	A	yes	Won
Good	one	A	yes	Lost
Good	both	A	yes	Lost

To perform the calculations we can use the built-in calculator in Windows. Launch the calculator from the Accessories group of Programs and select ‘scientific’ mode.

In the exam you can bring your own calculator provided that it is in the CASIO FX82, FX83 and FX85 series.

1. We shall start by calculating the intrinsic information of the dataset before splitting, which is the entropy of the 'Result' attribute.

   The first point to note is that some calculator may not have a button to calculate $\log_2$. Instead, it can calculate $\log_{10}$. This problem can be solved using the following identity:

   $$\log_2(x) = \frac{\log_{10}(x)}{\log_{10}(2)}. \tag{7}$$

   Now, to calculate the entropy of the 'Result' attribute we first have to calculate class probabilities. Class 1 is 'Won', and there are 9 examples in this class. Class 2 is 'Lost' and there are 5 examples in class 2. We now calculate the entropy using Equation 1

   $$\begin{aligned} \text{entropy}\left(\frac{9}{14}, \frac{5}{14}\right) &= -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} \\ &= \frac{-9\log_2 9 - 5\log_2 5 + 14\log_2 14}{14} \qquad \text{using Equation (3)} \\ &= \frac{-9\log_{10} 9 - 5\log_{10} 5 + 14\log_{10} 14}{14\log_{10} 2} \\ &\approx \frac{-8.5881 - 3.4949 + 16.0458}{4.2144} \\ &\approx 0.9403. \end{aligned}$$

   You need to write down a reasonable number of decimal places for intermediate results to reduce rounding error; four seems sensible.

2. The next step is to compute the entropy value of a split. Let us pick the attribute 'Ground'. This has two values, so there are two branches.

   (a) Consider first branch 'B'. The information content of this branch is given by the distribution of matches won and lost given that the ground is 'B'. There are three examples from each of the won and lost classes, so the entropy of this branch is

   $$\begin{aligned} \text{entropy}(3/6, 3/6) &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} \\ &= \frac{-\log_{10} 1 - \log_{10} 1 + 2\log_{10} 2}{2\log_{10} 2} \\ &= \frac{0 + 0 + 2\log_{10} 2}{2\log_{10} 2} \\ &= 1. \end{aligned}$$

   (b) Now you have a go at the 'A' branch. There are ... examples of the won class and ... examples of the lost class. The entropy of the branch is:

   $$\begin{aligned} \text{entropy}(\ldots \quad , \ldots \quad) &= \frac{-\ldots \ \log_{10}\ldots \ - \ldots \ \log_{10}\ldots \ + \ldots \ \log_{10}\ldots}{\ldots} \\ &\approx \frac{-\ldots \qquad -\ldots \qquad +\ldots}{\ldots} \\ &\approx \ldots \end{aligned}$$

(c) The next step is to combine these two entropy values in a weighted sum using Equation 5. The weights are given by the number of examples in the corresponding branches:

$$\begin{aligned}\text{entropy}_{\text{after split}} &= \frac{6}{14}\,\text{entropy}(3/6, 3/6) + \frac{8}{14}\,\text{entropy}(6/8, 2/8) \\ &\approx \frac{6}{14} \times 1 + \frac{8}{14} \times \ldots \\ &\approx 0.8922.\end{aligned}$$

(d) The final step in the calculation of the information gain is to subtract the information value of the split data from the information value of the unsplit data:

$$\text{gain} \;\approx 0.9403 - 0.8922 = 0.0481.$$

4 Exercises

1. Compute the information gain of the 'form' attribute. I get the following values.

   - Entropy of 'Good' branch: entropy$(2/5, 3/5) \approx 0.9710$.
   - Entropy of 'Variable' branch: entropy$(\ldots \quad , \ldots \quad) \approx \ldots$.
   - Entropy of 'Poor' branch: entropy$(\ldots \quad , \ldots \quad) \approx \ldots$. (You shouldn't need to do a long calculation here; just look at an earlier answer).
   - Entropy after split:

   $$\frac{5}{14} \times 0.9710 + \frac{\ldots}{\ldots} \times \ldots \quad + \frac{\ldots}{\ldots} \times \ldots \qquad\qquad \approx 0.6936.$$

   - Information gain: $\ldots$.

   Is 'ground' or 'form' the better attribute for predicting match results?

2. Compute the information gain of the 'key players' attribute. I get the answer 0.9111 for the entropy of the split and hence 0.0292 for the information gain.

3. For further practice, try the 'bat first' attribute.