**Entropy of Two Ordinal Variables,**

**Discretized Codings**

The Input Data: 1

3

1

2

4

5

Etc.

2024

Prof Wendy Olsen

Step 1

Decide whether the representation needed is distinct or cumulative.

If it is cumulative, we argue that a case experiences all of the conditions, up to the highest ranked condition.

If it is distinct, there is no such cumulation, so each condition is distinctive.

Step 2

Encode the single vector into multiple vectors, which are each binary.

For p input variables, there will now be q variables in the dataframe overall, q>p.

For example, one Likert scale of 5 levels including an NA option will become 5 binaries.

Education with 7 levels would become 7 levels, of which the last one is encoded all 1’s.

This means all the info is contained in 6 binaries for education [it is also the case that the information in a Likert scale is complete once 4 binaries are specified]. In an unsupervised discretization, we would not drop the constant binary after discretization. That would be a step to take later, perhaps at the statistical stage.

Figure 1: LIKERT SCALE, DISTINCT DISCRETIZATION

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Option1 | Option2 | Option3 | Option4 | Option5 |
| 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 1 |
| Etc. n  rows |  |  |  |  |

It is a sparse matrix.

Figure 2: A CUMULATIVE DISCRETIZATION

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Option1 | Option2 | Option3 | Option4 | Option5 |
| 1 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |
| Etc.  n rows |  |  |  |  |

Notice that column 1 now has 1 in every row, so it is not informative.

Step 3 Calculate and normalise entropy. The number of possible combinations of the elements in the series of events (options 1-5) is 5. The maximum entropy of both situations is 1.6094, the log of 5. This results from the Shannon entropy formula into which we insert the n cases on a uniform distribution across 5 options.