# 2.4. Mutual Entropy and Conditional Entropy

Print to PDF ▸

We started our discussion about entropy talking about optimal string encodings. But most data is a little more complicated than just strings. One of the predominant data formats is the “relational model” (or tabular model) where data are organized into tables of rows and columns. Examples of tabular data models include: spreadsheets, dataframes (like Pandas),and relational databases (like MySQL).

*Tabular data*: A rectangular data structure where rows correspond to observations and each column corresponds to properties (attributes) of the observation. Each column is named.

- Fixed attribute domain. The columns names define all of the relevant attributes for each observation.
- Atomic values. Each cell (a row, column pair) is generally considered to be an atomic value—i.e., it is not readily divisible–such as, an integer, a date, a string.
- NULL values. Missing or uncertain data can only be conveyed with a NULL/empty cell.

For, example a table of two columns A and B is:

| A | B |
|---|---|
| -1 | 1 |
| 0 | 0 |
| 1 | 1 |

How should we think about entropy (and its implications about compressibility) in such a model?

## 2.4.1. Mutual Entropy

Simply put, we can consider a table as a string of tuples that correspond to rows. For example, in the data above, we can have [(-1,1), (0,0), (1,1)]. Suppose each column $C_i$ has a *domain* (set of possible value) denoted by $dom(C_i)$. We can define a symbol set for the string of tuples that is the cartesian product of each of the columns domains:

$$\Sigma = dom(C_1) \times dom(C_2) ... \times dom(C_K)$$

So a table can be considered a string drawn from this compound symbol set:

$$T \in \Sigma^*$$

Over such a set, we can define entropy as before. Assume that each tuple $X$ is drawn indepently and identically from some distribution of value. The entropy $H(X)$ is defined as:

$$H(X) = -\sum_{s \in \Sigma} \log_2 p(X = s) \cdot p(X = s)$$

Note that here $X$ is tupled-valued and not a single scalar, but the same principle applies. One goes through every possible symbol (every possible tuple combination we could have). However, it is often beneficial to decompose these relationships into their constituent parts.

Now instead of a single random variable $X$ representing the whole tuple, let’s represent it as $X = (X_1, ..., X_k)$ where there is one scalar random variable for each colum. The *mutual entropy* of each of those random variables is:

$$H(X_1, ..., X_k) = -\sum_{x_1 \in dom(C_1)} \sum_{x_2 \in dom(C_2)} ... \sum_{x_k \in dom(C_k)} \log_2 p(X_1 = x1, ..., X_k = xk) \cdot p(X_1 = x1, ..., X_k = xk)$$

Like the intuition above suggests mutual entropy captures how much information or suprise there is in a collective of random variables. We can likewise define the same concept for continous random variables by turning the summations into integrals. We can interpret it in two ways: either it is the natural extension of entropy to cover joint probabilities or it considers the compressibility of a tabular data structure (as opposed to a single string!).

To make it more concrete, let’s look at some properties.

- Let X and Y be independent random variables $H(X,Y) = H(X) + H(Y)$
- For any two random variables $X$ and $Y$, $H(X,Y) \leq H(X) + H(Y)$

## 2.4.2. Conditional Entropy

Naturally, if we have defined a concept of mutual entropy there is a corrolary definition of *conditional* entropy:

$$H(Y|X) \;= - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)}$$

Essentially, this is looking at the joint distribution and measuring what is the average value of knowing $x$. The definition above leads to the intuitive property:

- $H(Y|X) = H(X,Y) - H(X)$ where the conditional entropy is simply the joint entropy with the information of the conditioned variable subtracted out.

By Sanjay Krishnan