Entropy, Conditional Entropy, and Mutual Information

An Introduction with Examples

Christoph Lohrmann, PhD

15th of September 2020

Empirical Probability

The term "probability" is defined in the Cambridge Dictionary as "the level of possibility of something happening or being true". An empirical probability for an event to happen is a probability that was determined using experiments / historic data. Thus, it can be easily defined as:

$$p(event) = \frac{\text{number of times an event occured}}{\text{number of observations/experiments}}$$

where $p(event)$ stands for the probability of a specific event to occur. An example for an empirical probability distribution p_X for a discrete random variable X could be the probabilities to obtain different grades in the "Free Analytics Environment R" course, where we assume for simplicity that there are only three possible outcomes of the course e.g. grades 3, 4, 5 (= "good", "very good", and "excellent"). The grades obtainable are discrete since they can only take discrete values (in this examples: 3, 4, 5) and not continuous values (e.g. 3.01, 3.02,...). Moreover, the grade is assumed to be a discrete random variable, which means that the grade can only take one out of multiple values (here: 3, 4, 5), each having its own probability to happen.

Let's assume that from the 200 students last year, 100 obtained a grade 3, 80 a grade 4 and 20 a grade 5. Then the corresponding empirical probabilities (as calculated using the formula above) are $p_X = \{0.5, 0.4, 0.1\}$ (since $\frac{100}{200} = 0.5$, $\frac{80}{200} = 0.4$, and $\frac{20}{200} = 0.1$). In this case X is the (discrete) random variable representing the grades, p_X the empirical probability distribution to obtain a grade and a specific grade represents a single event / outcome of the random variable X and is denoted x. So x=5 is the event that a grade 5 is obtained from the random variable X, and $p_X(x = 5\ ) = 0.5$ is the corresponding empirical probability to obtain this grade. It is noteworthy that probabilities take values within $[0, 1]$ - where 0 means an event is impossible to occur and 1 represents the certain occurrence of an event. Moreover, it is important to remember that probabilities sum up to exactly one, meaning that always one of the events happens (and that there are no other events that were not specified but may occur).

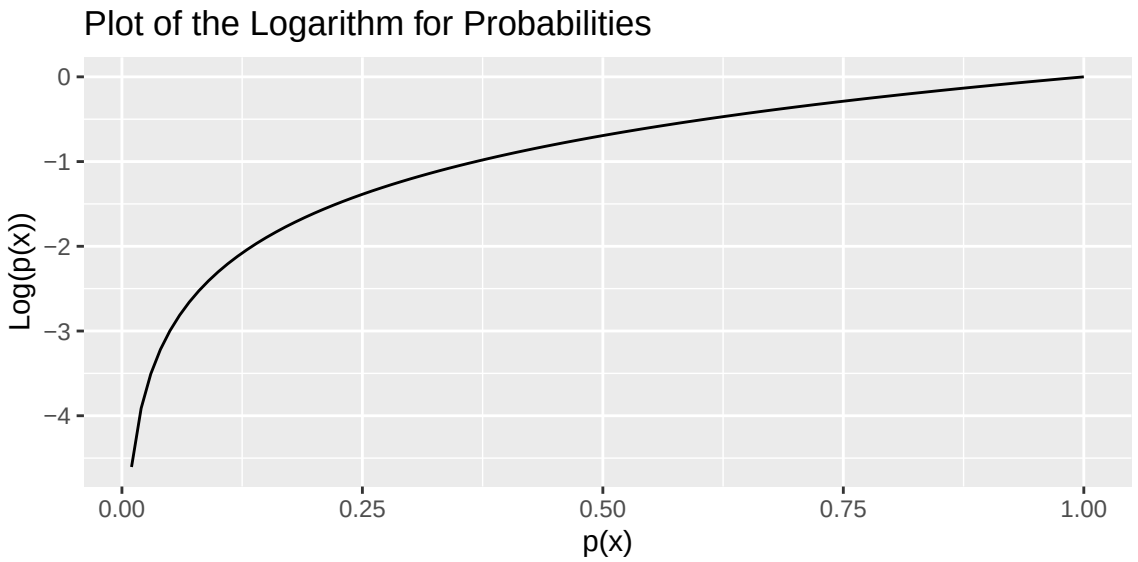
Lastly, we will introduce the concept of joint probability, which refers to the case we have e.g. two random variables X and Y and we want to determine the probability of two events happening at the same time (one event for each of the random variables). A simple example would be to obtain a grade 3 in the course "Free Analytics Environment R" and obtaining a grade 4 in a second course "Marketing Analytics". This would be denoted as $p_{X,Y}(x = 3\ y = 4\ )$ and would be calculated as previously, meaning as the number of times both events occurred simultaneously divided by the number of observations (or experiments).

Introduction to Entropy

Entropy is a measure of uncertainty for a probability distribution p_X for a (discrete) random variable X. The entropy of the (discrete) random variable X is then defined as

$$H(X) = -\sum_{x \in X} p_X(x) * log(p_x(x))$$

The formula shows that we sum over the product of each probability in the probability distribution p_X with the corresponding logarithm of it. The logarithm function used in the equation has the following shape:



The function shows a regressive shape, meaning that it has a positive slope that decreases over time (but remains positive) so that larger values for the probability p_X lead to smaller increases in $log(p_X)$. For a probability close to 1, the corresponding logarithm of it is close to zero, whereas for a probability close to zero, the corresponding logarithm of it is highly negative. Since the $log(p_X)$ for a probability (meaning in the interval $[0, 1]$) is always negative (only exception is when we take $log(1) = 0$), the multiplication of a probability $p_X(x)$ with it will always be negative. Hence, there is a negation sign in front of the sum to obtain a positive entropy value using this equation. Let's have a look at a couple of examples to get a better understanding of entropy and how it works.

No Uncertainty

Since entropy is a measure of uncertainty, the simplest case is that of no uncertainty, where there is only one outcome of X. Let's assume that all students in the past have obtained a grade three so that the event x=3 has an empirical probability of $p_X(x = 3\ ) = 1$. Since no other grade has historically occurred, the empirical probabilities for all other grades is zero. In this case the summation only includes a single element:

$$\begin{aligned} H(X) &= -p_X(x = 3) * log(p_X(x = 3)) \\ &= -1 * log(1) \\ &= 0 \end{aligned}$$

As expected, the entropy value is zero, reflecting that there is no uncertainty. Simply speaking, the empirical probability suggests that it is certain to obtain a grade 3 and it is not possible to obtain any other grade.

Varying Levels of Uncertainty

Assume X has three possible outcomes { 3, 4, 5 } with the following probabilities p_X = { 0.8, 0.1, 0.1 } (which obviously add up to 1). In simple terms, it is with a probability of 80% very likely for x to take a value of 3, whereas values of 4 or 5 are much less likely (probability 10% and 10%). This example clearly contains uncertainty since X has not just a single outcome, but can take multiple values and, hence, is not fully specific. The corresponding entropy value can be calculated as follows:

$$\begin{aligned}
H(X) &= -\sum_{x \in X} p_X(x) * log(p_X(x)) \\
&= -[p_X(x=3) * log(p_X(x=3)) + p_X(x=4) * log(p_X(x=4)) + p_X(x=5) * log(p_X(x=5))] \\
&= -[0.8 * -0.2231436 + 0.1 * -2.3025851 + 0.1 * -2.3025851] \\
&= -[-0.1785148 + -0.2302585 + -0.2302585] \\
&= 0.6390319
\end{aligned}$$

It is apparent that the entropy measure now indicates some uncertainty for the random variable X with the probability distribution p_X. If the uncertainty is now increased in the probability function p_X by making the other outcomes more likely, the entropy value will also increase. For instance, if p_X = { 0.6, 0.2, 0.2 } - so the probability mass is more distributed among the different outcomes of X, then the entropy is:

$$\begin{aligned}
H(X) &= -\sum_{x \in X} p_X(x) * log(p_X(x)) \\
&= -[p_X(x=3) * log(p_X(x=3)) + p_X(x=4) * log(p_X(x=4)) + p_X(x=5) * log(p_X(x=5))] \\
&= -[0.6 * -0.5108256 + 0.2 * -1.6094379 + 0.2 * -1.6094379] \\
&= -[-0.3064954 + -0.3218876 + -0.3218876] \\
&= 0.9502705
\end{aligned}$$

This entropy value is clearly higher than that of the previous example, where the probability mass was mainly in a single outcome (for grade 3).

The highest uncertainty is reached when all outcomes are equally likely. In this example with three possible outcomes for X, this would mean that p_X = { 0.333, 0.333, 0.333 } so that each outcome of { 3, 4, 5 } will have the same probability of 33.3%. In this case, the entropy value reflects this highest degree of uncertainty concerning which value X will take:

$$\begin{aligned}
H(X) &= -\sum_{x \in X} p_X(x) * log(p_X(x)) \\
&= -[p_X(x=3) * log(p_X(x=3)) + p_X(x=4) * log(p_X(x=4)) + p_X(x=5) * log(p_X(x=5))] \\
&= -[0.333 * -1.0996128 + 0.333 * -1.0996128 + 0.333 * -1.0996128] \\
&= -[-0.3661711 + -0.3661711 + -0.3661711] \\
&= 1.0985132
\end{aligned}$$

Comparing these examples, it is apparent that the elements in the sum take larger negative values for very similar probability values p_x (leading to a higher contribution to the uncertainty), whereas they take smaller negative value if the probability mass is very different, so that one outcome is much more likely than the other outcomes (leading to lower contributions to the uncertainty). A very large probability (close to 1) for instance has only a slightly negative value for the logarithm, resulting in a small negative value when multiplied with the probability. In contrast to that, a very small probability (close to 0) will have a large negative value for the logarithm, which, however, is multiplied with the small probability value, resulting in a small negative value. Probability values that are neither close to 0 nor to 1 do not face situations where either the probability or the logarithm of the probability are close to zero, resulting for the product of these two components in small negative values. Instead, the sum of a sufficiently large probability value (e.g. 0.333) together with the corresponding logarithm of it (e.g. -1.0996128) will result in a larger negative value (e.g. -0.3661711). This illustrates why a uniform distribution (where all outcomes are equally likely) is the most uncertain case, leading to the highest entropy value.

Conditional Entropy

Conditional entropy refers to the entropy of a (discrete) random variable X with respect to another (discrete) random variable Y. The aim of conditional entropy is to measure the remaining uncertainty in variable X when the outcome of Y is known. The formula to determine the conditional uncertainty is as follows:

$$H(X \mid Y) = -\sum_{x \in X} \sum_{y \in Y} p_{X,Y}(x,y) * log(\frac{p_{X,Y}(x,y)}{p_Y(y)})$$

It is apparent that instead of using the probability of a specific outcome x, denoted $p_X(x)$, the joint probability for the outcome x and y at the same time, denoted $p_{X,Y}(x,y)$, is used. Also, this probability is not multiplied with the log of the probability of x denoted $log(p_X(x))$. Instead, the log of the joint probability for a specific outcome for X and Y, denoted $p_{X,Y}(x,y)$, is divided by the probability of a specific outcome of Y, denoted $p_Y(y)$.This quotient in the logarithm is equal to the conditional probability $log(p_{X|y}(x))$. The equation can be re-written to highlight the meaning of the conditional entropy.

$$\begin{aligned} H(X \mid Y) &= -\sum_{x \in X} \sum_{y \in Y} p_{X,Y}(x,y) * log(\frac{p_{X,Y}(x,y)}{p_Y(y)}) \\ &= -\sum_{x \in X} \sum_{y \in Y} p_{X,Y}(x,y) * log(p_{X|y}(x)) \\ &= -\sum_{y \in Y} p_Y(y) \sum_{x \in X} p_{X|y}(x) * log(p_{X|y}(x)) \\ &= \sum_{y \in Y} p_Y(y) * H(X \mid y) \end{aligned}$$

So essentially the entropy of X given Y is the weighted entropy of X for each outcome y of the second (discrete) random variable Y. The weights are simply the probabilities for each outcome of Y.

Let's assume the simple example that all information in X is already provided by Y to illustrate the concept of conditional entropy. Let's assume that there are four students and we know their grades for two university courses e.g. "Marketing Analytics"(X) and "Free Analytics Environment R"(Y). We would like to know if the knowledge of their grade in "Free Analytics Environment R" provides us with information on how they score on "Marketing Analytics" or whether there is still a considerable uncertainty concerning their grade in "Marketing Analytics". :

	X	Y
Student A:	3	3
Student B:	3	3
Student C:	4	4
Student D:	5	5

In this example, outcomes and the probabilities for X and Y are { 3, 4, 5 } and $p_X = p_Y = \{0.5, 0.25, 0.25\}$. It is apparent that knowing Y leaves no uncertainty in the values X takes given that the values of Y are the same as those for X. In other words, knowing the grade in "Free Analytics Environment R"(Y) essentially means also knowing the grade in "Marketing Analytics"(X). The joint probabilities, meaning the probability for outcomes x and outcomes y being observed at the same time, are:

$$p_{X,Y} = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.25 & 0 \\ 0 & 0 & 0.25 \end{bmatrix}$$

As before, the probabilities need to sum up to 1 overall, which they do. The elements can be interpreted as follows: the value in row 1 and column 1 represents $p_X(x = 3, y = 3)$, which is the joint probability of the outcome of X being 3 when the outcome of Y is 3. In other words, the probability that a student obtains grade 3 in both courses. Such an outcome was observed for 2 out of the 4 students, meaning an empirical probability of 0.5. This probability is the same as $p_y(y = 3)$ since when y is 3, x is also always 3 (and vice versa). This is true for all outcomes that X and Y can take, so it is clear that knowing the outcome of Y means knowing the outcome of X with certainty. Applying the formula for the conditional entropy presented above will lead to the same conclusion (with a bit simplified notation):

$$\begin{aligned}
H(X \mid Y) &= -\sum_{x \in X} \sum_{y \in Y} p_{X,Y}(x, y) * log(\frac{p_{X,Y}(x, y)}{p_Y(y)}) \\
&= -[p(x = 3, y = 3) * log(\frac{p(x = 3, y = 3)}{p(y = 3)}) + p(x = 3, y = 4) * log(\frac{p(x = 3, y = 4)}{p(y = 4)}) + \cdots \\
&+ p(x = 5, y = 4) * log(\frac{p(x = 5, y = 4)}{p(y = 4)}) + p(x = 5, y = 5) * log(\frac{p(x = 5, y = 5)}{p(y = 5)})] \\
&= -[0.5 * log(\frac{0.5}{0.5}) + 0 * log(\frac{0}{0.25}) + 0 * log(\frac{0}{0.25}) + 0 * log(\frac{0}{0.5}) + 0.25 * log(\frac{0.25}{0.25}) \\
&+ 0 * log(\frac{0}{0.25}) + 0 * log(\frac{0}{0.5}) + 0 * log(\frac{0}{0.25}) + 0.25 * log(\frac{0.25}{25})] \\
&= 0
\end{aligned}$$

It is apparent that the conditional entropy $H(X \mid Y)$ is equal to zero, showing that the knowledge of Y leaves no uncertainty in X. This is reflected in the calculation where (1) each joint probability of $p_{X,Y}(x, y)$ that is zero will lead to zero contribution to the sum (due to the multiplication with it) and (2) each joint probability $p_{X,Y}(x, y)$ that is larger zero is always equal to $p_Y(y)$, so that a *log* of 1 is taken, which leads to a zero contribution to the sum as well (due to multiplication with zero). Hence, the conditional entropy reflects that the knowledge of the grade in "Free Analytics Environment R"(Y) provides full knowledge about the grade in "Marketing Analytics"(X).

Mutual Information

Mutual information brings the two concepts of entropy and conditional entropy together. It is formulated as follows:

$$I(X;Y) = H(X) - H(X \mid Y)$$

This means that mutual information is the difference between the uncertainty in X minus the uncertainty in X given the knowledge of Y. In case X and Y are independent - so the knowledge of Y has no impact on the probability of X, the mutual information of both of these variables is zero. So they do not have information in common (hence the name "mutual information"). Thus, X and Y are independent and mutual information as a measure of dependence reflects this by giving a score of zero. On the other hand, if Y decreases the uncertainty in X then the (original) entropy of X, $H(X)$, will be larger than the conditional entropy $H(X \mid Y)$, leading to a positive mutual information value. This would indicate some degree of dependence between these two variables since knowledge of Y leads to a lower uncertainty in X.

It is noteworthy that the perspective from which the mutual information $I(X;Y)$ is measured does not matter, meaning $I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$. Moreover, it is important to understand that mutual information is non-negative meaning it cannot take a negative value. The simple reason for this is that any variable cannot have more uncertainty given another variable than by itself, meaning that e.g. for any variable X it holds that $H(X) \geq H(X \mid Y)$. In other words, the variable X cannot be less certain having knowledge of another variable Y since, in the worst case, Y simply does not provide any information on X and, thus, $H(X) = H(X \mid Y)$. In case Y provides at least some information on X, the knowledge of it will reduce the uncertainty in X so that $H(X) > H(X \mid Y)$, leading to a positive mutual information value.

Let's consider several examples to illustrate this measure. Let's consider first the previous example with four students and their grades in "Marketing Analytics"(X) and "Free Analytics Environment R" (Y).

The entropy of X in this case is

$$\begin{aligned}
H(X) &= -\sum_{x \in X} p_X(x) * log(p_x(x)) \\
&= -[p_X(x=3) * log(p_x(x=3)) + p_X(x=4) * log(p_x(x=4)) + p_X(x=5) * log(p_x(x=5))] \\
&= -[0.5 * -0.6931472 + 0.25 * -1.3862944 + 0.25 * -1.3862944] \\
&= -[-0.3465736 + -0.3465736 + -0.3465736] \\
&= 1.0397208
\end{aligned}$$

The entropy value is clearly larger than zero, indicating that there is uncertainty in X. In other words, there is uncertainty in the grades that students get in the course "Marketing Analytics" since each student can either obtain a grade 3, 4 or 5.

The conditional entropy of X given Y (as previously calculated) is:

$$\begin{aligned}
H(X \mid Y) &= -\sum_{x \in X} \sum_{y \in Y} p_{X,Y}(x,y) * log(\frac{p_{X,Y}(x,y)}{p_Y(y)}) \\
&= 0
\end{aligned}$$

In other words, knowledge of Y leaves no uncertainty in the value X takes. Thus, the mutual information between X and Y is:

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= 1.0397208 - 0 \\ &= 1.0397208 \end{aligned}$$

The high positive value for mutual information (MI) shows that there is a strong dependency between X and Y. In particular, we know that this is the maximum dependence for these two variables since $I(X;Y) = H(X)$, meaning the information that Y and X have in common is equivalent to the entire uncertainty in X. It is apparent that for each student the results of the two courses are the same, so this result is not surprising - also, intuitively, because both courses require similar interest and skill in analytics.

Let's make a second example where we look at the relationship between two different courses that likely do not rely on a similar set of skills. Let's imagine again the "Marketing Analytics"(X) course but now also a course on "Creative Writing"(Y). The skills required to perform well in these courses are quite different so students that are very creative and have a high language skill will likely be good in "Creative Writing" while those good in marketing and analytics will probably perform well on "Marketing Analytics". Notwithstanding, there may be students that excel at both of these disciplines or may simply not be good at either of the two. The student results on both courses may be as follows:

	X	Y
Student A:	3	3
Student B:	3	5
Student C:	4	3
Student D:	5	5

The corresponding probabilities for X are $p_X = \{0.5, 0.25, 0.25\}$ whereas the probabilities for Y are $p_Y = \{0.5, 0, 0.5\}$ since no student obtained a grade 4 in "Creative Writing".
Student A, for instance, is only good at both courses, whereas student D excels at both. Student B and C are only doing very well in one of the courses but perform comparably worse at the other. Thus, in this example we would expect only a small degree of dependence between "Marketing Analytics" (X) and "Creative Writing" (Y) since the performance of neither of the two courses provides much information on how the performance on the second course will be. The joint probability distribution is now different than previously:

$$p_{X,Y} = \begin{bmatrix} 0.25 & 0 & 0.25 \\ 0.25 & 0 & 0 \\ 0 & 0 & 0.25 \end{bmatrix}$$

The probability of $x = 3$ is now distributed among $p(x = 3\ , y = 3)$ and $p(x = 3\ , y = 5)$, which both represent one out of four cases ($= 0.25$).

The entropy for X is the same as previously since the results for all students in "Marketing Analytics" (X) did not change:

$$
\begin{aligned}
H(X) &= -\sum_{x \in X} p_X(x) * log(p_x(x)) \\
&= -[p_X(x=3) * log(p_x(x=3)) + p_X(x=4) * log(p_x(x=4)) + p_X(x=5) * log(p_x(x=5))] \\
&= -[0.5 * -0.6931472 + 0.25 * -1.3862944 + 0.25 * -1.3862944] \\
&= -[-0.3465736 + -0.3465736 + -0.3465736] \\
&= 1.0397208
\end{aligned}
$$

But the conditional entropy of X given Y, meaning the remaining uncertainty in X given the knowledge in Y will not be zero as in the previous example since knowing the grade in "Creative Writing" (Y) still leaves uncertainty about the grade in "Marketing Analytics" (Y):

$$
\begin{aligned}
H(X \mid Y) &= -\sum_{x \in X} \sum_{y \in Y} p_{X,Y}(x,y) * log(\frac{p_{X,Y}(x,y)}{p_Y(y)}) \\
&= -[p(x=3,y=3) * log(\frac{p(x=3,y=3)}{p(y=3)}) + p(x=3,y=4) * log(\frac{p(x=3,y=4)}{p(y=4)}) + \cdots \\
&+ p(x=5,y=4) * log(\frac{p(x=5,y=4)}{p(y=4)}) + p(x=5,y=5) * log(\frac{p(x=5,y=5)}{p(y=5)})] \\
&= -[0.25 * log(\frac{0.25}{0.5}) + 0 * log(\frac{0}{0}) + 0.25 * log(\frac{0.25}{0.5}) + 0.25 * log(\frac{0.25}{0.5}) + 0 * log(\frac{0}{0}) \\
&+ 0 * log(\frac{0}{0.5}) + 0 * log(\frac{0}{0.5}) + 0 * log(\frac{0}{0}) + 0.25 * log(\frac{0.25}{0.5})] \\
&= -[-0.1732868 + 0 + -0.1732868 + -0.1732868 + 0 + 0 + 0 + 0 + -0.1732868] \\
&= 0.6931472
\end{aligned}
$$

It is apparent that the conditional entropy is not zero, showing that the knowledge of the grade in "Creative Writing" (Y) leaves considerable uncertainty for the grade in "Marketing Analytics" (X). In the formula there are four elements that are non-zero, so that contribute to the remaining uncertainty in X having the knowledge of Y. Those are for the outcomes $p(x = 3\ , y = 3)$, $p(x = 3\ , y = 5)$, $p(x = 4\ , y = 3)$, and $p(x = 5\ , y = 5)$. It is easy to understand why this is the case. These four probabilities cover the cases when y is 3 or NA (which is in this example all grades y can take). When y=3, then x can take either a value of 3 or of 4. Hence, in both are part of the remaining uncertainty. A similar case is the one when y=5, then x can take either a value of 3 or of 5 also showing that there is uncertainty in x when we have knowledge of this value of y.

The mutual information between X and Y is then:

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X \mid Y) \\
&= 1.0397208 - 0.6931472 \\
&= 0.3465736
\end{aligned}
$$

Hence, there is some dependence between X and Y (e.g. knowing that y=3 means that x cannot take a value 5) but the knowledge in Y still leaves a considerable level of uncertainty in X.

Implementation of mutual information in R

One way to implement mutual information in R is using the *mutinformation* function in the **infotheo** package. In particular, after the variables X and Y are defined, the function call can be as follows:

```
# Example "Marketing Analytics"(X) and "Free Analytics Environment R" (Y)
X <- c(3,3,4,5)
Y <- c(3,3,4,5)
mutinformation(X, Y, method="emp")
```

```
## [1] 1.039721
```

```
# Example "Marketing Analytics"(X) and "Creative Writing" (Y)
X <- c(3,3,4,5)
Y <- c(3,5,3,5)
mutinformation(X, Y, method="emp")
```

```
## [1] 0.3465736
```

We can see that the results obtained using this function call are the same as those we calculated manually. Hence, we are assured that the function correctly calculates mutual information (and it was also verified that our calculation of mutual information was correct).

Symmetric Uncertainty

Symmetric uncertainty is a normalized version of mutual information. It simply uses the mutual information $I(X;Y)$ and standardizes it using the entropy of X and the entropy of Y:

$$SU(X;Y) = \frac{2 * I(X;Y)}{H(X) + H(Y)}$$

The value for symmetric uncertainty will be in the interval $[0, 1]$. This is advantageous if we want to interpret the dependence between two variables since we clearly know the range that the values can take. This was not directly true for the mutual information. For instance, in the first example where X and Y contained entirely the same information, the mutual information value was over one. We only realized that this value was high given that we knew that the conditional entropy was zero, so that the mutual information was equal to the entropy of X. Without this knowledge, it would not have been that apparent to us that the mutual information reflects the fact that variable X and Y overlap entirely in the information they provide. However, this knowledge is exactly helpful for us to understand how the standardization works. In case X and Y contain the same information, the entropy (=uncertainty) in X will be the same as in Y so $H(X) = H(Y)$. We already know that in case both variables share exactly the same information, that $H(X \mid Y) = 0$, so $I(X;Y) = H(X)$. This is the maximum value mutual information can take. Given that in this case $H(X) = H(Y)$, we can easily see that the symmetric uncertainty will be one (by re-writing the formula for this specific case):

$$\begin{aligned} SU(X;Y) &= \frac{2 * I(X;Y)}{H(X) + H(Y)} \\ &= \frac{2 * H(X)}{H(X) + H(X)} \\ &= 1 \end{aligned}$$

On the other extreme, X and Y could be entirely independent. In that case the conditional entropy would be the same as the (unconditional) entropy. So $H(X \mid Y) = H(X)$, since the knowledge of Y does not provide any information on the values that variable X will take. Thus, the mutual information as a measure of dependence would be zero, indicating that there is no dependence between X and Y. In the calculation this is apparent since $I(X;Y) = H(X) - H(X \mid Y)$ which in this case would be $I(X;Y) = H(X) - H(X) = 0$. Given this extreme case, the minimum value of symmetric uncertainty will be one (by filling in the knowledge we just obtained):

$$\begin{aligned} SU(X;Y) &= \frac{2 * I(X;Y)}{H(X) + H(Y)} \\ &= \frac{2 * 0}{H(X) + H(Y)} \\ &= 0 \end{aligned}$$

For all other levels of dependence between X and Y that are not complete dependence and independence (two extreme cases) values in between 0 and 1 are obtained - reflecting the level of information these two variables have in common.