Measures of Dependence and Information Theory

CS 4710

Sunday 31$^{\text{st}}$ March, 2019

1 Measures of Dependence

If X and Y are independent, we know a lot about what their probability distributions *aren't*. For instance, we know that $P(X|Y) = P(X)$; Y confers no information about X, so the conditional probability of X doesn't change given Y.

If X and Y are *dependent*, however, we need to be able to measure and describe *how* dependent. Given that X and Y are dependent, there's some functional relationship between X and Y. We can categorize measures of dependence into the different broad families of functions over which our dependency estimation needs to differ drastically.

As we'll see, there are linear, monotonic, and non-linear kinds of dependence. Some of these overlap – linear dependencies will be monotonic as well, for instance. However, some monotonic relationships are not linear. Thus, we need techniques for each of these cases, at the very least. This section will discuss those measures.

1.1 Linear and Monotonic Measures of Dependence

1.1.1 Pearson correlation coefficient ($\rho_{X,Y}$)

Definition. The **mean-adjustment** of a variable X is the random variable A, defined as:

$$A = X - E(X)$$

This effect centers A around the origin, with the same variance as X. (See figure below.)

Definition. The **normalization** of a variable X is the scaling of its variance from σ_X^2 to 1, performed on the mean-adjustment of the variable. (See other figure below.)

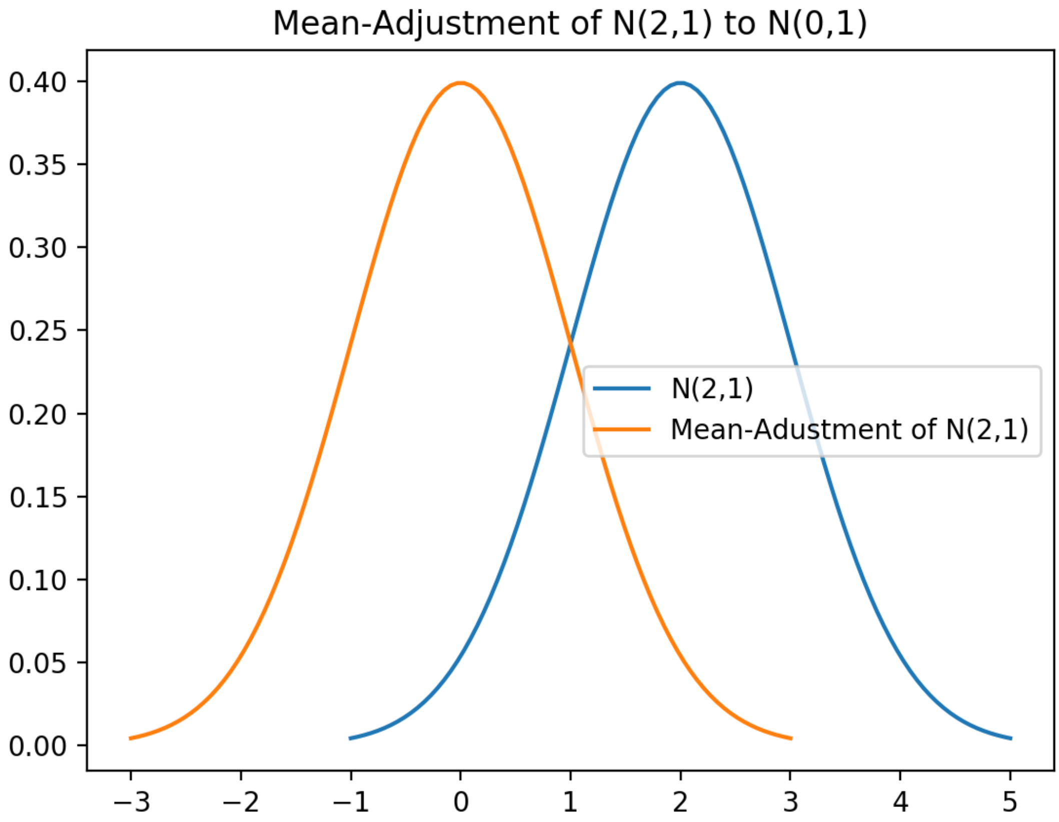
$$Norm(X) = \frac{X-E(X)}{\sigma_X}$$

Pearson correlation ($\rho_{X,Y}$)

$\rho_{X,Y}$ is a value between -1 and 1, capturing maximally extreme anti-correlation at -1 and maximally extreme correlation at 1. 0 corresponds to independence.

Key assumptions for $\rho_{X,Y}$

1. Relationship between X and Y is linear



2. X and Y have meaningful averages (e.g. aren't horseshoe-shaped, or something similar)

Definition. Pearson correlation

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Person correlation is the "product moment" of X and Y divided by the product of the standard deviation of X and the standard deviation of Y.

Definition. Product moment of X and Y

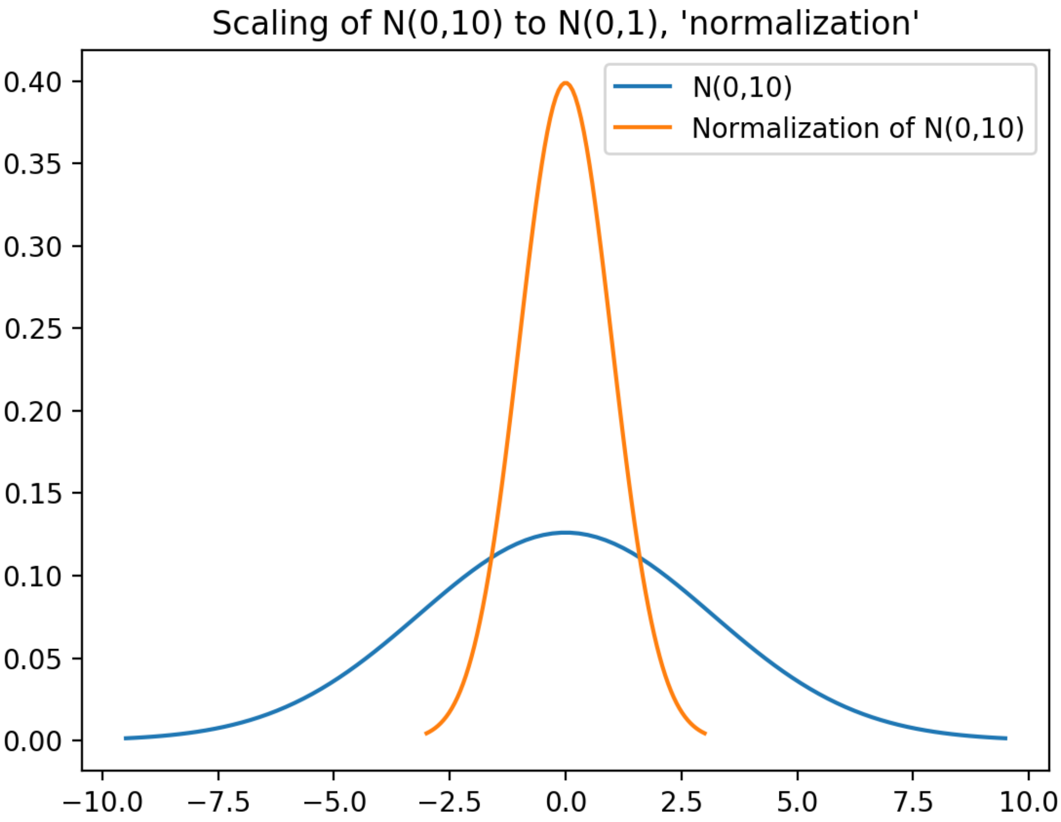
$$E[(X - E(X))(Y - E(Y))]$$

In terms of estimating some of these values (such as σ_X), you can resort to $E(X)$. (To see why, review the definition of variance.)

$$\sigma_X^2 = E(X^2) - (E(X))^2$$

$$\sigma_X = \sqrt{E(X^2) - (E(X))^2}$$

Exploiting facts like this about the connection between expectation and variance allows us to rewrite $\rho X, Y$:



$$E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

$$\rho_{X,Y} = \frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-(E(X))^2}\sqrt{E(Y^2)-(E(Y))^2}}$$

Note if X and Y are independent, then $E(XY) = E(X)E(Y)$, driving $\rho_{X,Y}$ to 0.

The standard way to calculate $\rho_{X,Y}$ is called "Pearson's r" (this is often calculated by software implementations of linear regression):

$$r_{x,y} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

where $\bar{X}$ is the sample average.

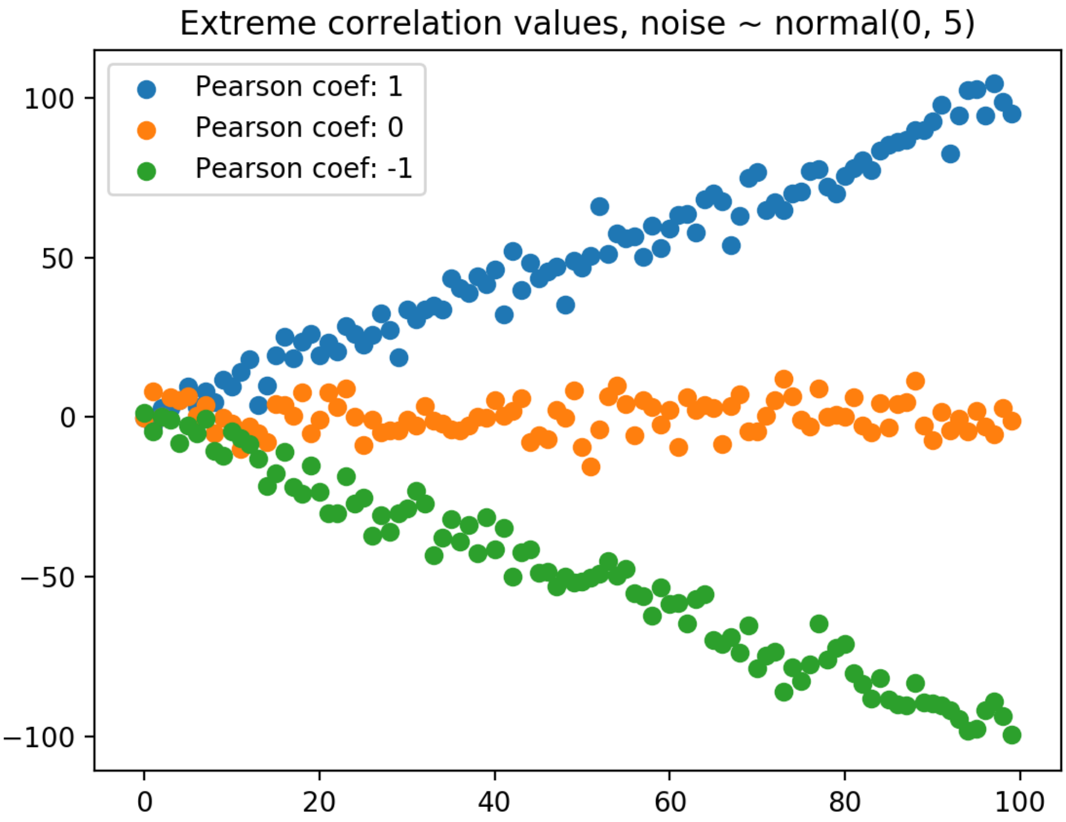
$\rho_{X,Y}$ has *relational symmetry*:

$$\rho_{X,Y} = \rho_{Y,X}$$

$\rho_{X,Y}$ is *invariant* under some shifting and scaling transformations of X, Y:

$$\rho_{X,Y} = \rho_{f(X),g(Y)}$$

where



$$f(X) = a + bX$$
$$g(Y) = c + dY$$
$$\text{for } a, b, c, d > 0$$

If $\rho_{X,Y} > 0$ then joint observations of both X, Y both tend to the same side of $E(X), E(Y)$ at the same time. In other words, when X is observed above its mean, the corresponding observation of Y tends to be above its mean as well. The same holds for being below the mean.

On the other hand, if $\rho_{X,Y} < 0$ then X and Y tend to sit on *opposite* sides of their respective means. When $X_i > E(X)$ then typically $Y_i < E(Y)$ (same goes when reversing both inequalities).

We can use Pearson's correlation to construct a basic test of independence.

1. Calculate $r_{X,Y}$.
2. For a "small enough" value of $r_{X,Y}$, conclude it is approximately zero.
3. If $r_{X,Y}$ is approximately zero, conclude X and Y are independent. Otherwise, conclude they are dependent.

Above we make reference to a "small enough" value in our test without defining what "small enough" is. We'll see how to set this threshold value later in hypothesis testing.

Pearson's r is **asymptotically unbiased**:

$$\begin{gathered}\text{Suppose } X \sim P(X|\theta)\text{, with estimator } \hat{\theta}.\\ Bias(\hat{\theta},\theta) = E_{X|\theta}(\hat{\theta}) - \theta\\ Bias(\hat{\theta},\theta) = E_{X|\theta}(\hat{\theta} - \theta)\\ \text{Asymptotically unbiased:}\\ \lim_{n\to\infty} E_{X|\theta}(\hat{\theta_n} - \theta) = 0\end{gathered}$$

So as our data set grows to arbitrary size n, the difference of $\rho_{X,Y}$ and $r_{X,Y}$ approaches 0. In other words, r zeroes in on ρ given enough data. This result doesn't tell us *how much* data (for some particular problem), but it does tell us that over time r becomes a tighter estimate.

1.1.2 Spearman's rank correlation ($S_{X,Y}$)

Spearman's rank correlation is defined based on Pearson correlation. Furthermore, it is a measure of *monotonic* relationships (not linear relationships in general). In other words, it is a measure of dependence that tracks whether or not two variables rise and fall together (or the opposite), but not the degree to which it happens linearly.

Monotonic relationships are more general than linear ones. Any ranking can have Spearman's rank correlation calculated over it; furthermore, any metric dataset can be converted to ranked data (based on the dataset's weak order). For instance, you could ask people to rank a set of movies by personal enjoyment and then calculate the Spearman's correlation between any two respondents' rankings. This won't capture scale at all (the size of the difference between the rankings), so Pearson's correlation is not directly applicable – still, the monotonic relationships that may be present between rankings is captured by $S_{X,Y}$.

Definition. Spearman correlation between X and Y

$$S_{X,Y} = \rho_{rank(X),rank(Y)}$$

Therefore, we can estimate S with estimate $\hat{S}$, constructed using substituting in Pearson's r for ρ:

$$\hat{S_{X,Y}} = r_{rank(X),rank(Y)}$$

where $rank(X), rank(Y)$ are any ordinal labelings of X and Y consistent with their weak orders.

1.2 Non-Linear Measures of Dependence

This section will cover non-linear measures of dependence. In order to do so, we need to introduce some basic information theory. Information theory gives us the tools to capture non-linear forms of dependence through the concept of *entropy.*

1.2.1 Information

To introduce some non-linear measures of dependence, we need to introduce a different (yet equivalent) perspective to probability – **information**. Unlike probability, information is a non-normalized score (it can get quite large and has no maximum in general like probability does). Information is measured in units, and we'll see this boils down to the choice of a logarithmic base (with 2 corresponding to the "bit").

We'll start by giving the four fundamental properties of information (similar to the axioms of probability). These will imply a functional form for information, as we'll see soon.

Four fundamental properties of information

For probability $p = P(X)$ and event X:

1. $(I(p) = 0) \Leftrightarrow (p = 1)$ – probability 1 events yield 0 information
2. $(p > p') \Leftrightarrow (I(p) < I(p'))$ – as p gets larger, $I(p)$ gets smaller (and vice versa)
3. $I(p) \geq 0$ for all X – information is non-negative
4. $I(p, p') = I(p) + I(p')$ for p and p' probabilities of *independent events*

Theorem. If $I(p)$ satisfies fundamental properties 1-4, then

$$I(P(X)) = log_k(\tfrac{1}{P(X)})$$

Any choice of k will do, but we will choose $k = 2$ as this makes the unit of information the "bit". We're going to skip the proof on this, but you can find it in Claude Shannon's original formulation of information theory, published while working at Bell Labs.

1.2.2 Average information (entropy)

For random variable X, we will write $I(X)$ as shorthand for $I(P(X))$. As X is a random variable, $I(X)$ is one as well. Thus we can calculate its expectation:

$$E(I(X)) = E(\tfrac{1}{P(X)})$$

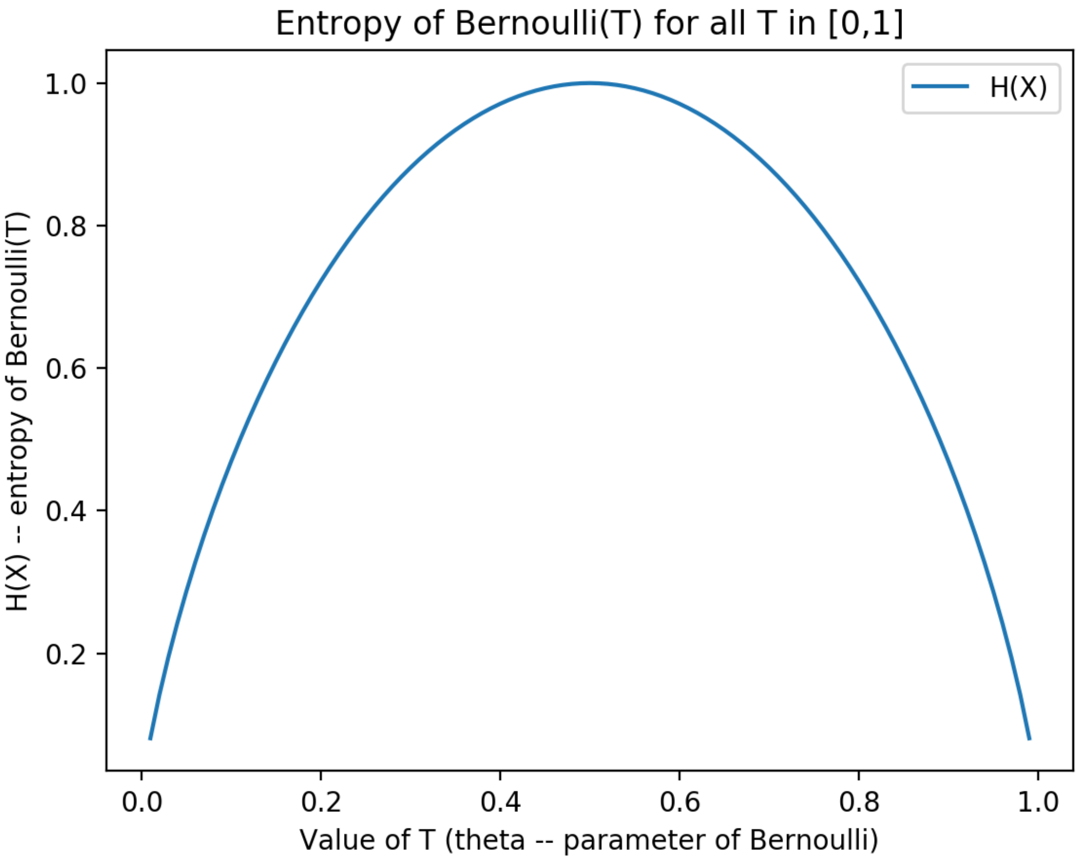
$$= \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \log_2(\tfrac{1}{P(X_i)})$$

$$= - \textstyle\sum_{i=1}^{n} (P(X_i) \log(P(X_i)))$$

For physics students, it may be useful to note this expression is equivalent to Boltzmann's H (function for entropy):

$$H(X) = - \textstyle\sum_{i=1}^{n} (P(X_i) \log(P(X_i)))$$

It may help to view Bernoulli random variables in terms of their entropy. We can take our parameter θ and treat it as a variable, estimating $H(X)$ across the different values of θ's support: Note that the maximum entropy situation is where $P(X = 1) = \frac{1}{2} = \frac{1}{n}$ for $n = |\Omega|$. In general for discrete distributions this will be true – entropy is maximized when we are maximally unsure of the result. Thus, the distribution must be *uniform* – every alternative is equally likely.



1.2.3 Mutual information and conditional entropy

We can now consider mutual information and conditional entropy – two-variable comparisons that allow us to characterize the entropy in one of the variables if the second is known.

Definition. Mutual information

$$M(X,Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2(\frac{P(x,y)}{P(x)P(y)})$$

Note when X and Y are independent, $P(X,Y) = P(X,X) = P(X)P(Y)$ and so $M(X,Y) = 0$.

Definition. Conditional entropy of X given Y

$$H(X|Y) = E(I(X)|Y)$$

$$= -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log(\frac{P(x,y)}{P(x)})$$

Conditional entropy gives us the "unexplained residue" for X given Y. In the case where $P(X) = P(Y)$, $P(X,Y) = P(X)$, driving conditional entropy to 0. In other words, if you know X, then X has no entropy – you can reconstruct X from itself with no errors.

1.2.4 Variation of information

The measures we've looked at so far are not distance metrics. This means that while they help us assess the similarity of X and Y in terms of functions of their mutual information, they don't necessarily tell us "how far apart" X and Y are informationally. Variation of information is a function of two events that satisfies the four axioms of distance metrics, and tells us precisely how much X and Y differ (in a way that helps us compare them both to event Z, and so on).

Definition. Variation of information

$$V_I X, Y = H(X|Y) + H(Y|X)$$

This can be thought of as the "sum of unexplained residues" between X and Y in both directions. $H(X|Y)$ tells us how much information we need on average to represent the unavoidable errors (on average) when reconstructing X given Y. For instance, for some particular value of Y, there may be two values of X seen each with equal proportion. Then the condition entropy of X given that value of Y is 1 bit, the same as a coin flip.

$V_I(X, Y)$ satisfies the four properties of distance metrics, making it a distance metric.

1. $V_I(X, Y) \geq 0$ – non-negativity
2. $(V_I(X, Y) = 0) \Leftrightarrow (X = Y)$ – identity of indiscernibles
3. $V_I(X, Y) = V_I(Y, X)$ – relational symmetry
4. $V_I(X, Z) \leq V_I(X, Y) + V(Y, Z)$ for all X, Y, Z – triangle inequality

2 Examples for Information Theory

Example: fair coin

Suppose $X_1 \sim \mathit{Bernoulli}(\theta_{X_1})$, $X_2 \sim \mathit{Bernoulli}(\theta_{X_2})$. Then:

$$(X_1 \text{ independent of } X_2) \Rightarrow (P(X_1, X_2) = P(X_1)P(X_2))$$

For X_n with $X_i \sim \mathit{Bernoulli}(\theta_{X_i})$ with $1 \leq i \leq n$:

$$(X_1 \text{ independent of } \ldots \text{ independent of } X_n) \Rightarrow (P(X_1, \ldots, X_n) = \textstyle\prod_{i=1}^n P(X_i))$$

$$\Rightarrow I(X_1, \ldots, X_n) = \textstyle\sum_{i=1}^n I(P(X_i))$$

Note each X_i contributes somewhere between 0 and 1 bit on average to the joint information of X and Y.

Example. Suppose $k = 2$, $X \sim \mathit{Bernoulli}(\theta = 0.5)$.

$$P(X = 1) = 0.5 = P(X = 0)$$

$$I(P(X = 1)) = \log_2(\tfrac{1}{P(X=1)}) = \log_2(1/0.5) = 1$$

$$E(I(P(X))) = \frac{1}{n}\sum_{i=1}^{n}\log_2(\frac{1}{P(X_i)})$$

$$= \frac{1}{2}(1+1) = 1$$

Alternatively this could have been calculated with the following expression:

$$-\sum_{x \in X} P(x)\log_2(P(x))$$

As we can see, a communication channel communicating the result of the toss of a fair coin would take 1 bit of information to encode on average.

Example: unfair coin

Suppose $Y \sim Bernoulli(\theta = 0.75)$.

$$-\sum_{y \in Y}\log_2(P(y)) = -(0.75 * \log_2(0.75) + 0.25 * \log_2(0.25))$$

$$\approx -(-0.3113 - 0.5) \approx 0.8113 \text{ bits on average}$$

Notice how when the "coin" in question isn't fair, it requires less than the full 1 bit to encode *on average*. This is to say there would be a potential to compress the signal (in the abstract at least).

Example: conditional entropy

Suppose you're an exobiologist studying alien veggies. You're extra brave among exobiologists and have decided to test the sourness of various alien veggies by eating them yourself. Your hypothesis is that the color of the veggie might give you information about its sourness.

Suppose our dataset looks as follows:

Color (X)	Sour? (Y)
yellow	yes
yellow	yes
green	no
green	no
red	yes
red	yes
red	no
red	no

We can calculate our empirical joint distribution by adding up the cases for particular combinations of X and Y, and dividing out by the total size of our dataset n.

	Sour	Not Sour
Yellow	0.25	0
Green	0	0.25
Red	0.25	0.25

Note that the sum of the sum of the columns is 1, as is the sum of the sum of the rows.

$$H(X) = -0.5\log_2(0.5) - 0.25\log_2(0.25) - 0.25\log_2(0.25) = 1.5$$

$$H(Y) = -0.5\log_2(0.5) - 0.5\log_2(0.5) = 1$$

Note that $H(X) \neq H(Y)$, and so we know straight away X and Y are not perfect guides to each other.

Now we'll calculate $H(Y|X)$ for a few values of X, to see how good of a guide X is for Y (given our dataset, which is admittedly quite small).

First let X = yellow:

$$H(Y|X = \text{yellow}) = -P(Y = \text{yes}|X = \text{yellow})\log_2(P(Y = \text{yes}|X = \text{yellow})) - P(Y = \text{no}|X = \text{yellow})\log_2(P(Y = \text{no}|X = \text{yellow}))$$

$$= -0 * log_2(0) - 1 * log_2(1) = 0$$

Given our dataset, knowing the value of "yellow" reduces the entropy of sourness to 0. In other words, the empirical probability of sourness given yellow is 1 (given our dataset).

Now let X = red:

$$H(Y|X = \text{red}) = -(P(Y = \text{yes}|X = \text{red})\log_2(P(Y = \text{yes}|X = \text{red})) + P(Y = \text{no}|X = \text{red})\log_2(P(Y = \text{no}|X = \text{red}))$$

$$= -(0.5\log_2(0.5) + 0.5\log_2(0.5)) = 1$$

If the color is red, then (given our dataset) it's equivalent to a fair coin flip whether the veggie is sour or not in terms of entropy.

Now let's extend the example to discuss variation of information. We'll have to calculate both $H(X|Y)$ an $H(Y|X)$ to do this:

$$H(X|Y) = -\sum_x \sum_y P(x,y)\log_2(\tfrac{P(x,y)}{P(x)})$$

$$H(X|Y) = -(0.25\log_2(0.25/0.25) + (0.25\log_2(0.25/0.25)) + (0.25\log_2(0.25/0.5)) + (0.25\log_2(0.25/0.5))$$

$$= -(-0 - 0 - 0.25 - 0.25) = 0.5$$

And for $H(Y|X)$:

$$H(Y|X) = -(0.25\log_2(0.25/0.5) + 0.25\log_2(0.25/0.5) + 0.5\log_2(0.25/0.5) + 0.5\log_2(0.25/0.5)$$

$$= -(-0.25 - 0.25 - 0.5 - 0.5) = 1.5$$

And so for variation of information:

$$V_I(X,Y) = H(X|Y) + H(Y|X) = 2$$

which should accord with the dataset intuitively. There are two unavoidable discrepancies between X and Y – where color is red but, on two occasions, but the veggie is sour (as opposed to not sour, or vice versa). As these are two binary values, the variation of information between X and Y is measured as two bits (it would take 2 bits on average to encode the difference between X and Y given what we've seen).