Mutual Information

# Entropy

 Jesse: Your style is reasonably clear, but you need to increase the information density, i.e. each sentence must say more. Also, we have to give a solid intuition before we get technical. The intuition should be as close as possible to the real concept (that’s why I didn’t like the colleagues talking to one another). Also, at some point we should mention the relationship between MI and correlation and also how well MI captures combinatorial settings (e.g. need A and B together).

Mutual Information measures how much each of two variables tells us about the other. For example, you may have no idea whether I’m sick or whether I have a temperature above the normal, but if you know one then you’ll have a good guess about the other. Sickness and high temperature have a high mutual information. On the other hand, sickness and the color of my kitchen cupboards have very low mutual information.

Technically, the uncertainty about a variable is measured by *entropy*. To calculate the entropy of a random variable $X$, we find the *probability mass function* of $X$, the probability of obtaining any given value $x$ in $X$. As an example of a probability mass function, imagine rolling a fair die. The probability of getting each value is $frac{1}{6}$, and so the probability mass function is $frac{1}{6}$ for each value $x \in (1 ... 6)$, and $0$ everywhere else. (For continuous random variables, the probability density function (discussed in Section XX {this should be discussed in the Bayes section}) is used instead.) We then calculate the entropy of $X$, written as $H\left(X\right)$ in the following way,

$H\left(X\right) = - \sum{n}{i=1}{p \left( x\_i \right) log\_b p\left(x\_i\right)}$

where $p\left(x\_i\right)$ is the probability mass function’s value when $X=x\_i$ and $b$ is a chosen base of the log, usually $2$, $e$, or $10$. So, if we were to calculate the entropy of a fair die roll use log base $2$, we would end up with

$H\left(X\right) = - \sum{6}{i=1}{\frac{1}{6} log\_2 \left(\frac{1}{6}\right)}$

$H\left(X\right) = - 6 \* (frac{1}{6} log\_2 (\frac{1}{6}))$

$H\left(X\right) = 2.585$

The entropy of a fair die roll is about $2.585$ *bits*. Bits are the unit of entropy obtained when using log base $2$. Other units can be *nits* (log base $e$), and *dits* (log base $10$). The number obtained above can be interpreted as being the amount of information each toss of the die gives us. Because this is a fair die, this is the highest possible entropy for this problem, because it is the most random. If we used a die that was weighted so $2$ appeared more often than the other numbers, we would have a lower entropy value. For example, if we modify the probability of $2$ with a weight, so there is a $0.5$ probability of $x = 2$, and keeping all other values at equal probability ($0.1$), we obtain

$H\left(X\right) = - 5 \* (frac{1}{10} log\_2 (\frac{1}{10})) + 1 \* (frac{1}{2} log\_2 (\frac{1}{2}) $H\left(X\right) = 1.161$

Now, each roll of the die is giving us *less* information. If the die always fell on 2, then we would learn nothing from a roll of the die and the entropy would be zero.

# Mutual Entropy

 An important part of calculating the Mutual Information in genomic applications is calculating the conditional entropy between two genes, or $H \left(Y | X \right)$ where $X$ and $Y$ are two different genes, each with expression values at $n$ time points. Intuitively, the conditional entropy of $Y$ given $X$ is the amount of entropy left over in $Y$ once $X$’s entropy is taken into account.

 As an example of conditional entropy, suppose we have information about club membership and voting patterns among undergraduates. We can set up possible values for $Club$ and $Votes?$ as such,

 $Club \in \left( Debate Team \\

 Computer Club \\

 History Club \right)$

 $ Votes? \in \left( Yes \\

 No \right) $

 Below is our data from the survey.

| Club | Votes? |
| --- | --- |
| Debate Team | Yes |
| Debate Team | Yes |
| Computer Club | No |
| Computer Club | No |
| History Club | Yes |
| History Club | No |

 We can see from the table that both of the people from the Debate Team said that they planned to vote, wheras no one from the Computer Club would vote, and the History Club was perfectly split. Thus, the club determines voting for any member of the Computer or Debate Clubs, but not for the History Club. We calculate these specific conditional entropies as follows:

 $ H \left( Y | X = “Debate Team” \right) = 0 $

 $ H \left( Y | X = “Computer Club” \right) = 0$

 $ H \left( Y | X = “History Club” \right) = 1$

 We can the full calculate the conditional entropy by summing over all values of $H \left( Y | X = v\_i \right)$, weighted by the probability of that $v\_i$ occurs. Formally,

 $ H \left( Y | X \right) = \sum{i = 1,...,n}{P \left( X = v\_i \right) \* H \left( Y | X = v\_i \right) }$

 If we work this out, we get:

 $ H \left( Y | X \right) = 0.33 \* 0 + 0.33 \* 0 + 0.33 \* 1 $

 $ H \left( Y | X \right) = 0.33 $

 So the conditional entropy $H \left( Y | X \right)$ is $0.33$. Calculating entropy and conditional entropy for continuous variables (such as gene expression values) follows the same idea, but is instead done using the integral of the values $v\_i$ and the probability density functions of $X$ and $Y$.

# Calculating the Mutual Information

The Mutual Information value can be calculated by combining the above two values. So, the MI between two random variables $X$ and $Y$ is defined as

$I(X;Y)=H(X) – H(X|Y)$

MI can also be formulated in terms of the probability density functions of $X$ and $Y$.

$I(X;Y) = \integral{\_Y\integral{\_X p(x,y) log\_b \left( frac{p(x,y)}{p\_1(x)p\_2(y)}\right)}dx}dy

Where the base $b$ of the log is $2$, $e$, or $10$, $p(x,y)$ is the joint probability distribution function of $X$ and $Y$, $p\_1(x)$ is the marginal probability density function of $X$, and $p\_2(y)$ is the marginal probability density function of $Y$. For a more in-depth discussion of probability density functions, please see the section on Bayesian approaches. If $p(x)$ and $p(y)$ are independent of each other, then $p(x,y) = p(x)p(y)$, which will make the quotient in equation XX equal to $1$, which turns to $0$ when we take the log of it, dropping the MI value to $0$. Intuitively, this makes sense: if $x$ and $y$ are independent, then knowing one doesn’t give any information about knowing the other and so they have zero mutual information.

 In genomics applications, we are interested in what the values of gene $X$ tell us about the values of gene $Y$. Does gene $X$ give us information about what gene $Y$ is doing and if so how much? If a lot, we may want to connect them with an edge. -- Add an example using the small gene network example