Solutions 8: Statistical dependence

BIO-369

Prof. Anne-Florence Bitbol
EPFL

1 Some examples

a) See Jupyter notebook. Since $X = Y$, for each draw, we have $x = y$, and thus the arrays of results are identical for the two random variables. All markers in the scatter plot are on the $x = y$ diagonal.

b) Since $X = Y$, we expect the correlation coefficient between X and Y to be equal to 1 (positive correlation, maximum value). See Jupyter notebook.

c) See Jupyter notebook. We can check on examples that the known identities (that are not already used to estimate the various quantities) are satisfied, for instance that $I(X;Y)+H(X|Y)-H(X) = 0$.

d) See Jupyter notebook. Since $X = Y$, $H(X) = H(Y)$. Here, for all x and y, the joint probability $P(x, y)$ satisfies $P(x, y) = P(x)\delta_{x,y}$ where $\delta_{x,y}$ is 0 if $x \neq y$ and 1 if $x = y$. Therefore

$$H(X,Y) = -\sum_{x\in\Omega, y\in\Omega} P(x,y)\log_2[P(x,y)] = -\sum_{x\in\Omega} P(x)\log_2[P(x)] = H(X)\,. \tag{1}$$

Thus, we have $I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X)$. Besides, $H(X|Y) = H(X,Y) - H(Y) = 0$, and similarly $H(Y|X) = 0$.

Let us turn to the interpretation of these results. $H(Y|X)$ is the average amount of uncertainty (missing information) remaining about Y (the value y that Y takes) after X (the value x that X takes) is known. Here, because $X = Y$, this remaining uncertainty is zero. Then, $I(X;Y)$ is the amount of uncertainty in Y that is removed by knowing X, as can be seen by $I(X;Y) + H(Y) - H(Y|X)$. Indeed, it's the amount of uncertainty (missing information) in Y, namely $H(Y)$, minus the amount of uncertainty in Y that remains after X is known, namely $H(Y|X)$. Here, all the uncertainty about Y, namely $H(Y)$, is removed by knowing X, and thus $I(X;Y) = H(Y)$, which is also equal to $H(X)$.

e) See Jupyter notebook. Here, we no longer have $X = Y$, but a substantial part of Y is still determined by X, we can view this as Y being equal to X plus some noise. Accordingly, the correlation coefficient of X and Y is still positive and quite large but it is not 1. The entropies of X and Y are now different but have similar values, in agreement with the fact that the two histograms are similar. However, now, the joint entropy of X and Y is substantially larger than $H(X)$ (or $H(Y)$): since y no longer exactly tracks x, there is more uncertainty about the pair (x, y) than about x or y alone. The mutual information between X and Y is still substantial (compared e.g. to $H(X)$) but significantly smaller than $H(X)$ and $H(Y)$, which makes sense as knowing the value of X removes some uncertainty on Y, but not all of it. And accordingly, here $H(Y|X)$ and $H(X|Y)$ are substantial (compared respectively to $H(Y)$ and $H(X)$).

f) See Jupyter notebook. In this case, the correlation between X and Y is close to zero, while y is a deterministic function of x. This is due to the fact that their relationship is nonlinear and non-monotonic. Specifically, the contributions to the correlation cancel out between the decreasing and the increasing part of the scatter plot.

Besides, here, the histograms show that the distribution of Y is more peaked than that of X. Therefore, we expect to have $H(Y) < H(X)$, which is the case. Here, the mutual information is

substantial (compared to $H(Y)$ or $H(X)$), which captures the fact that there is a strong statistical dependence between these two random variables, while this was missed by correlation. In addition, here we have $H(X|Y) > H(Y|X)$, which makes sense since $H(X) > H(Y)$ and $H(X) = I(X;Y) + H(X|Y)$ while $H(Y) = I(X;Y) + H(Y|X)$. Let us interpret the fact that $H(X|Y) > H(Y|X)$. Since $H(Y|X)$ is the average amount of uncertainty (missing information) remaining about Y (the value y that Y takes) after X (the value x that X takes) is known, $H(X|Y) > H(Y|X)$ means that there is more missing information about X if Y is known than vice-versa, which is consistent with intuition when looking at the scatter plot (for many values of y, there are two possible ranges for x).

g) Here the correlation is almost zero, as in the previous case. The two histograms are less different than in the previous case, and accordingly, entropies are less different. We also observe that in the decompositions $H(X) = I(X;Y) + H(X|Y)$ and $H(Y) = I(X;Y) + H(Y|X)$, the relative importance of mutual information has decreased, while that of conditional entropies have increased. This makes sense, as now there is more uncertainty left about one variable after the other one is known.

2 Coevolving sites in interacting proteins

a) See Jupyter notebook. There are 21 possible states in this data, the 20 natural amino acids and the alignment gap.

b) See Jupyter notebook.

c) See Jupyter notebook. We find that some pairs of sites have particularly large mutual information. In the histogram, we can see a few pairs with scores larger than 0.5. In the colorscale matrix, we can see that some of them are localized rather close to one another.

d) See Jupyter notebook.

e) See Jupyter notebook. Comparing visually the contact map obtained here to the colorscale representation of the matrix of mutual informations obtained before, we see that some of the pairs of sites with high mutual information seem to coincide with contacts.

f) See Jupyter notebook. We find that most of the pairs of sites with high mutual information correspond to a pair of sites that are in contact in the three-dimensional structure of the HK-RR complex. Specifically, with the threshold chosen on mutual information, 13 out of 14 of these pairs are in contact, and moreover, they are the 13 top ones. This confirms the impression from visualizing the two matrices. The pairs of sites with high mutual information are often pairs of sites that are in contact between the HK and the RR, they are at the interface between the two molecules and they determine the specificity of the interaction between an HK and its cognate RR.