Information Theory **DCC/ICEx/UFMG**
Prof. Mário S. Alvim **2017/02**

SOLUTION OF HOMEWORK

Dependent Random Variables
(MacKay - Chapter 8)

Necessary reading for this assignment:

- *Information Theory, Inference, and Learning Algorithms* (MacKay): *Information Theory, Inference, and Learning Algorithms* (MacKay):
  - Chapter 8.1: *More about entropy*

Note: The exercises are labeled according to their level of difficulty: `[Easy]`, `[Medium]` or `[Hard]`. This labeling, however, is subjective: different people may disagree on the perceived level of difficulty of any given exercise. Don't be discouraged when facing a hard exercise, you may find a solution that is simpler than the one the instructor had in mind!

Review questions.

1. The entropy $H(X) = -\sum_x p(x) \log p(x)$ can be interpreted as the uncertainty one has about the random variable X. With that in mind, for each of the items below, give its name, its mathematical formula and explain its meaning in terms of uncertainty.

   (a) $H(X, Y)$.

   Instructor's solution: The joint entropy
   $$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y)$$
   can be interpreted as the uncertainty of random variables X and Y taken together.

   (b) $H(X \mid Y)$.

   Instructor's solution: The conditional entropy
   $$H(X, Y) = \sum_y p(y) H(X \mid Y = y) = -\sum_{x,y} p(x, y) \log p(x \mid y)$$
   of X given Y can be interpreted as the uncertainty of random variable X given that the value of random variable Y is known.

   (c) $I(X; Y)$.

   Instructor's solution: The mutual information
   $$I(X; Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) = I(Y; X)$$
   among random variables X and Y can be interpreted as the information that random variable X caries about Y, that is, the amount by which the knowledge of Y reduces the uncertainty of X. (Note that the mutual information is symmetric.)

   (d) $I(X; Y \mid Z)$.

Instructor's solution: The mutual information

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

among random variables X and Y, given random variable Z, can be interpreted as the information that random variable X caries about Y, that is, the amount by which the knowledge of Y reduces the uncertainty of X, all that under the assumption that the value of random variable Z is known. (Note that the conditional mutual information is symmetric.)

2. State the following "laws" of information theory.

   (a) The chain rule for entropy $H(X_1, X_2, \ldots, X_n)$.

   Instructor's solution: The chain rule for entropy is

   $$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_1 \ldots, X_{i-1}).$$

   (b) The chain rule for mutual information $I(X_1, X_2, \ldots, X_n; Y)$.

   Instructor's solution: The chain rule for mutual information is

   $$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_1, \ldots, X_{i-1}),$$

   where

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

   is the conditional mutual information between X and Y given Z.

   (c) The data-processing inequality (DPI), and explain what it intuitively means.

   Instructor's solution: If $X \to Y \to Z$ form a Markov Chain, that is, if $p(x,y,z) = p(x)p(y \mid x)p(z \mid y)$ for all x, y, z in X, Y, Z, then $I(X;Z) \leq I(Y;Z)$.
   Intuitively, DPI means that if X is processed to produce Y, and then Y is processed to produce Z, then Z can carry no more information about X than Y does. In other words, if you post-process output Y to obtain Z, you can never gain information about input X. Yet in other words: "post-processing cannot create information".

Exercises.

3. (MacKay 8.1) `[Medium]`

Instructor's solution:

$$\begin{aligned}
H(X,Y) &= H((U,V),(V,W)) \\
&= H(U,V,V,W) \\
&= H(U) + H(V \mid U) + H(V \mid U,V) + H(W \mid U,V,V) && \text{(by the chain rule)} \\
&= H(U) + H(V) + 0 + H(W) && (*) \\
&= H_u + H_v + H_w,
\end{aligned}$$

where in $(*)$ we used the fact that U, V and W are independent, and the fact that $H(V \mid U, V) = 0$ (since when both U and V are known, there isn't any uncertainty about V).

$$\begin{aligned} H(X \mid Y) &= H((U,V) \mid (V,W)) \\ &= H(U, V \mid V, W) \\ &= H(U \mid V, W) + H(V \mid V, W, U) && \text{(by the chain rule)} \\ &= H(U) + 0 && (**) \\ &= H_u, \end{aligned}$$

where in $(**)$ we used the fact that U, V and W are independent, and the fact that $H(V \mid V, W, U) = 0$.

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= H((U,V)) - H_u && \text{(by the calculation above)} \\ &= H_u + H_v - H_u && \text{(since } U \text{ and } V \text{ are independent)} \\ &= H_v. \end{aligned}$$

4. (MacKay 8.2) `[Medium]`

   Instructor's solution: Given in the textbook.

5. (MacKay 8.6) `[Easy]`

   Instructor's solution: Given in the textbook.

6. (MacKay 8.7) `[Medium]`

   Instructor's solution: Given in the textbook.

7. (MacKay 8.9) `[Hard]`

   Instructor's solution: Given in the textbook.

8. (MacKay 8.10) `[Medium]`

   Instructor's solution:

   (a) The experiment is to drawn a card. There are 6 possible outcomes for a card, since there are 3 ways of picking a card, and 2 ways of choosing what side is up.

   Let us call l the value of the lower face, and u the value of the upper face of the card drawn.

   We want to decide what color is the lower face if we know that the upper face is black, so we should calculate what is the posterior probability distribution induced on the lower face given that the upper face is black.

First let's calculate $p(l = \textit{white} \mid u = \textit{black})$:

$$\begin{aligned} p(l = \textit{white} \mid u = \textit{black}) &= \frac{p(l = \textit{white}, u = \textit{black})}{p(u = \textit{black})} && \text{(by the Bayes theorem)} \\ &= \frac{1/6}{3/6} && (*) \\ &= \frac{1}{3}, \end{aligned}$$

where in $(*)$ we used the fact that $p(l = \textit{white}, u = \textit{black}) = 1/6$ since only one of the 6 equally probable ways of picking a card leads to a black upper side and a white lower side, and the fact that $p(u = \textit{black}) = 1/6$ because three of the 6 equally probable ways of picking a card leads to a black upper side.

Now let's calculate $p(l = \textit{black} \mid u = \textit{black})$:

$$\begin{aligned} p(l = \textit{black} \mid u = \textit{black}) &= \frac{p(l = \textit{black}, u = \textit{black})}{p(u = \textit{black})} && \text{(by the Bayes theorem)} \\ &= \frac{2/6}{3/6} && (*) \\ &= \frac{2}{3}, \end{aligned}$$

where in $(*)$ we used the fact that $p(l = \textit{black}, u = \textit{black}) = 2/6$ since two of the 6 equally probable ways of picking a card leads to a black upper side and a black lower side, and the fact that $p(u = \textit{black}) = 1/6$ because three of the 6 equally probable ways of picking a card leads to a black upper side.

Hence, given that the upper side of the card is black, there's a $2/3$ probability that the lower side is black, against only a $1/3$ probability that the lower side is white, so it is sensible to guess that the lower side is black.

(b) Before the any card is drawn, there is a $1/2$ probability that a given card has a white lower side, and a $1/2$ probability that the card has a black lower site (because there are 6 equally likely outcomes for picking a card, and 3 lead to a white lower side, while 3 lead to a black lower side).

Calling L the random variable associated with lower side of a card, the entropy of L is $H(L) = H(0.5, 0.5) = 1$ bit.

Similarly, calling U the random variable associated with upper side of a card, the entropy of U is $H(U) = H(0.5, 0.5) = 1$ bit.

Now, if know that the value of the upper face, we have the following distributions (similar to the calculation in item (a)):

- if the upper side is black, we get the distribution $p(l = \textit{white} \mid u = \textit{black}) = 1/3$ and $p(l = \textit{black} \mid u = \textit{black}) = 2/3$, which means that

$$\begin{aligned} H(L \mid U = \textit{black}) &= H(1/3, 2/3) \\ &= -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} \\ &= \log 3 - \frac{2}{3} \end{aligned}$$

- if the upper side is white, we get the distribution $p(l = \textit{white} \mid u = \textit{white}) = 2/3$ and

$p(l = \textit{black} \mid u = \textit{white}) = 1/3$, which means that

$$\begin{aligned} H(L \mid U = \textit{white}) &= H(2/3, 1/3) \\ &= -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} \\ &= \log 3 - \frac{2}{3} \end{aligned}$$

Hence, the entropy of the lower side given the upper side is

$$\begin{aligned} H(L \mid U) &= p(U = \textit{black}) \cdot H(L \mid U = \textit{black}) + p(U = \textit{white}) \cdot H(L \mid U = \textit{white}) \\ &= \frac{1}{2} \cdot \left(\log 3 - \frac{2}{3}\right) + \frac{1}{2} \cdot \left(\log 3 - \frac{2}{3}\right) \\ &= \log 3 - \frac{2}{3}. \end{aligned}$$

Hence, the mutual information among the upper and lower face is

$$\begin{aligned} I(U; L) &= H(L) - H(L \mid U) \\ &= 1 - \left(\log 3 - \frac{2}{3}\right) \\ &= \frac{5}{3} - \log 3, \end{aligned}$$

which means that by learning the upper side of the card we learn $5/3 - \log 3 \approx 0.082$ bits of information about the lower side card.