Short definition summary for the first part of chapter 2[1]

31 January 2008

Probability

Consider X and Y, two random variables with X taking values in $\mathcal{X}$ and Y in $\mathcal{Y}$. The *joint probability* is

$$p_{X,Y}(x,y) = \text{probability that X=x and Y=y} \tag{1}$$

From the joint probability we can define the *marginal distributions*

$$\begin{aligned} p_X(x) &= \text{probability that } X = x \text{ irrespective of what } Y \text{ is} \\ p_Y(y) &= \text{probability that } Y = y \text{ irrespective of what } X \text{ is} \end{aligned} \tag{2}$$

and, it follows that

$$\begin{aligned} p_X(x) &= \sum_{y\in\mathcal{Y}} p_{X,Y}(x,y) \\ p_Y(y) &= \sum_{x\in\mathcal{X}} p_{X,Y}(x,y) \end{aligned} \tag{3}$$

We can also define the *conditional probabilities*

$$\begin{aligned} p_{X|Y}(x|y) &= \text{probability that } X = x \text{ if } Y = y \\ p_{Y|X}(y|x) &= \text{probability that } Y = y \text{ if } X = x \end{aligned} \tag{4}$$

These are calculated using Bayes rule, basically this says that the probability of $X = x$ and $Y = y$ is the probability of $X = x$ multiplied by the probability of $Y = y$ given that $X = x$:

$$p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y|x) \tag{5}$$

and, similarily

$$p_{X,Y}(x,y) = p_Y(y)p_{X|Y}(x|y) \tag{6}$$

and, of course, this means

$$\begin{aligned} p_{Y|X}(y|x) &= \frac{p_{X,Y}(x,y)}{p_X(x)} \\ p_{X|Y}(x|y) &= \frac{p_{X,Y}(x,y)}{p_Y(y)} \end{aligned} \tag{7}$$

[1]Conor Houghton, `houghton@maths.tcd.ie` please send me any corrections.

Entropy, information and divergence

The *entropy* of a random variable X with probability distribution $p_x(x)$ is

$$H(X) = -\sum_{x\in\mathcal{X}} p_X(x)\log p_X(x) \tag{8}$$

Now, this can be applied to a joint distribution, after all, a join distribution is just a distribution for $(X,Y)\in\mathcal{X}\times\mathcal{Y}$ so

$$H(X,Y) = -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x,y)\log p_{X,Y}(x,y) \tag{9}$$

The *KL divergence* sometimes called the *relative entropy* of two distributions $p_X(x)$ and $q_X(x)$ for the same random variable X is

$$D(p_X(x)\|q_X(x)) = \sum_{x\in\mathcal{X}} p_X(x)\log\frac{p_X(x)}{q_X x} \tag{10}$$

It can be thought of as measuring the difference between two putative probability distribution. Note that it is not symmetric in $p_X(x)$ and $q_X(x)$.

The *mutual information* of a pair of random variables is

$$I(X,Y) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \tag{11}$$

Hence

$$I(X,Y) = D(p_{X,Y}(x,y)\|p_X(x)p_Y(y)) \tag{12}$$

and the mutual information measures the KL-divergence between the joint probability and the multiple of the marginal probabilities.

Conditional entropy and information

If we have a joint distribution $p_{X,Y}(x,y)$ then for given value of Y, $Y=y$ the conditional distribution can be used to give an entropy. The *conditional entropy* is the average of this entropy over all values of Y:

$$H(X|Y) = \sum_{y\in\mathcal{Y}} p_Y(y)\sum_{x\in\mathcal{X}} p_{X|Y}(x|y)\log p_{X|Y}(x|y) \tag{13}$$

and, by Bayes,

$$H(X|Y) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x,y)\log p_{X|Y}(x|y) \tag{14}$$

There is an important result which can be derived straight from the definitions:

$$H(X,Y) = H(X) + H(Y|X) \tag{15}$$

and

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{16}$$