1 facts about classical probability theory I often forget or get confused

See also Preskill's notes at `http://www.theory.caltech.edu/people/preskill/ph229/notes/chap5.pdf` which are a very useful reference.

- As should be obvious $p(x) = \int dy\ p(x,y) = \int dy\ p(x|y)p(y)$
- The following should be considered an axiom of probability theory

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

  I usually have to think carefully to convince myself that this should be an axiom. This is pretty much the entire reason I sometimes get rusty on probability theory. From the above it follows that

$$p(x|y,z) = \frac{p(x,y,z)}{p(y,z)}$$

- Bayes' theorem follows trivially from the above

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

- One can reparemeterize a distribution using

$$|p(x)dx| = |p(\xi)d\xi|$$

  in other words

$$p(\xi) = p(x)\left|\det\left(\frac{\partial x}{\partial \xi}\right)\right|$$

- The Shannon entropy is defined as

$$H(X) = -\int dx\ p(x)\log(p(x))$$

  And is the expectation value of the log of the distribution. One can see intuitively that this is an entropy by noting that the Shannon entropy of a δ function vanishes while the entropy of a constant is maximal.

- One can define the conditional Shannon entropy

$$H(X|Y) = -\int dxdy\ p(x,y)\log(p(x|y)) = H(X,Y) - H(Y)$$

  which is the expectation value of the conditional probability. This can simply be interpretted as the amount of "randomness" in X for a particular fixed y.

- One can define the mutual information

$$I(X,Y) = H(X) - H(X|Y) = \int dxdy\ p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

  This can be thought of as a measure of how well X is determined by Y. Note that if Y fully determines X then $H(X|Y)$ vanishes and $I(X,Y) = H(X)$, its maximum value. If X and Y are completely independent, then $H(X|Y) = H(X)$ and $I(X,Y) = 0$. A particularly nice way of thinking about the mutual information is using the following, which is obtained trivially from the above

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

That is, the mutual information is the entropy of $p(x,y)$ if X and Y were independent minus the actualy entropy of $p(x,y)$. It follows from this that the mutual information is symmetric in its arguments.

- The mutual information $I(X,Y)$ is independent of invertable reparemeterizations of the probability measures. This follows immediately from the explicit definition of $I(X,Y)$ and from $|p(x)dx| = |p(\xi)d\xi|$. (Note that such reparemeterizations cannot mix X and Y.)

- Entropy is subadditive
$$H(X,Y) \leq H(X) + H(Y)$$
Of course the quantum analog of this also holds true. This also guarantees that the mutual information is positive.

- The classical shannon entropy of a whole system is always greater than the entropies of each of its constituent parts
$$H(X,Y) \geq H(X) \qquad H(X,Y) \geq H(Y)$$
This is obviously quite different from the quantum case where the strongest analogous statement is the triangle inequality
$$S(\rho_{AB}) \geq |S(\rho_A) - S(\rho_B)|$$
This is essentially the statement that there is no entanglement in classical information theory.

- From the above it is trivial to see the following lower and upper bounds on the entropy
$$H(X) \leq H(X,Y) \leq H(X) + H(Y)$$

- Very confusingly, the cross-entropy between two distributions is often written as $H(X,Y)$. To avoid this confusion, I'll refer to the cross-entropy as
$$\sigma(p,q) = -\int dx\ p(x)\log(q(x))$$
This is another measuer of "difference" between two distributions. I find its meaning far less intuitive than the other measures talked about here, but it can serve as a useful objective function when trying to match distributions.