Lecture 3, Tuesday w2, 2014-09-23

Main lecture points:

- Block coding idea: identify a list of possible files with almost all the probability mass under a model. Just use a fixed width code for them. (And deal with the rest any way you like, they happen rarely, so it doesn't matter.)
- The law of large numbers says an average of samples will be close to their mean.
- Central limit theorem (CLT): *Close to the mean*, the sum or mean of N independent variables with bounded mean and variance tend towards being Gaussian distributed as N increases.
- Chebyshev's inequality: bounds tail probabilities of *any* distribution far from the mean.
- Information content: intuition, and definition, $\log(1/p) = -\log(p)$.

Check your progress

Get familiar with information contents:

- How do you convert between information contents in bits, nats, or bans?
- The submarine game (MacKay, p71): can you show that if you keep asking questions until you identify the submarine, the total information content experienced is always 6 bits? MacKay shows this on p72.

Applying the law of large numbers, CLT, and/or Chebyshev's:

Imagine you have a machine learning system that makes a real-valued prediction (e.g., temperature). You measure the absolute error made on each case in a large test set of size N, and compute the mean absolute error $\hat{m}$. This estimator $\hat{m}$ is a random variable, it depends on the particular test set that you gathered. If you gathered a new test set, you'd get a different estimate. What can you say about how $\hat{m}$ is distributed (and under what assumptions)? It may be useful to talk about the true mean absolute error m, and its variance σ^2, which you might also have to estimate.

That is: do you know how to put a standard error bar on an estimate and know what that means? If you do any experimental work (including numerical experiments) in your project, you'll probably want to put error bars on some estimates.

Recommended reading

We are part way through the 'week 2' slides. MacKay pp66–73 gives more detail for the intuitions behind information content.

Ask on NB if anything is unclear, or too compressed.

For keen people

Again, it's a good idea to try reproducing plots from the slides. Plotting graphs is a useful research skill. And if you have to implement something, it really tests whether you know where the plot came from.