Some maths background for Information Theory: Expectations and sums of variables

Iain Murray

September 9, 2014

This note reviews some of the course pre-requisites. I will expect you to know the material in this note thoroughly so that you can understand the course material. If anything here isn't clear after working through it, come to office hour or agree to meet me before a lecture. If this material is hard-going you might consider requesting a late transfer to a less mathematical course.

1 Probability Distributions / Ensembles

We will often assume that we are interested in a so-called *ensemble*, $X = \{x, \mathcal{A}_X, \mathcal{P}_X\}$, that will give us an outcome, x, from a discrete set or 'alphabet' $\mathcal{A}_X = \{a_1, a_2, \ldots, a_I\}$, with corresponding probabilities $\mathcal{P}_X = \{p_1, p_2, \ldots, p_I\}$. This notation is chosen to match the MacKay course text.

At first we will assume that we know the complete definition of X, and simply wish to compute properties of it, or define codes for encoding and decoding a future outcome, x, that we haven't yet observed.

Examples

A standard six-sided die has $\mathcal{A}_X = \{1, 2, 3, 4, 5, 6\}$ with corresponding probabilities $\mathcal{P}_X = \{1/6, 1/6, 1/6, 1/6, 1/6, 1/6\}$.

A Bernoulli distribution, which has probability distribution

$$P(x) = \begin{cases} 1 & p, \\ 0 & \text{otherwise}, \end{cases}$$

defines an ensemble with $\mathcal{A}_X = \{1, 0\}$ with $\mathcal{P}_X = \{p, 1-p\}$.

2 Expectations

Expectations are properties of probability distributions that we can compute. The expectation of some function, f, of an outcome, x, is:

$$\mathbb{E}_{P(x)}[f(x)] = \sum_{i=1}^{I} p_i f(a_i).$$

Often the subscript $P(x)$ is dropped from the notation because the reader knows under which distribution the expectation is being taken.

The expectation is sometimes a useful representative value of a random function value. The expectation of the identity function, $f(x) = x$, is the 'mean', which is one measure of the centre of a distribution.

The expectation is a *linear operator*:

$$\mathbb{E}[f(x) + g(x)] = \mathbb{E}[f(x)] + \mathbb{E}[g(x)] \qquad \text{and} \qquad \mathbb{E}[cf(x)] = c\mathbb{E}[f(x)].$$

These properties are apparent if you explicitly write out the summations.

The expectation of a constant with respect to x is the constant:

$$\mathbb{E}[c] = c\sum_{i=1}^{I} p_i = c,$$

because probability distributions sum to one ('probabilities are normalized').

The expectation of independent outcomes separate:

$$\mathbb{E}[f(x)g(y)] = \mathbb{E}[f(x)]\,\mathbb{E}[g(y)].$$

True if x and y are independent.

Exercise 1: prove this.

3 The mean

The mean of a distribution over a number, is simply the 'expected' value of the numerical outcome.

$$\text{'Expected Value'} = \text{'mean'} = \mu = \mathbb{E}[x] = \sum_{i=1}^{I} p_i a_i.$$

For a six-sided die:

$$\mathbb{E}[x] = \frac{1}{6}\times 1 + \frac{1}{6}\times 2 + \frac{1}{6}\times 3 + \frac{1}{6}\times 4 + \frac{1}{6}\times 5 + \frac{1}{6}\times 6 = 3.5.$$

In every day language I wouldn't say that I 'expect' to see 3.5 as the outcome of throwing a die…I expect to see an integer! However, 3.5 is the expected value as defined. Similarly a single Bernoulli outcome will be a zero or a one, but its 'expected' value is a fraction,

$$\mathbb{E}[x] = p\times 1 + (1-p)\times 0 = p,$$

the probability of getting a one.

Change of units: I might have a distribution over heights measured in metres, for which I have computed the mean. If I multiply the heights by 100 to obtain heights in centimetres, the mean in centimetres can be obtained by multiplying the mean in metres by 100. Formally: $\mathbb{E}[100\,x] = 100\,\mathbb{E}[x]$.

4 The variance

The variance is also an expectation, measuring the average squared distance from the mean:

$$\text{var}[x] = \sigma^2 = \mathbb{E}[(x-\mu)^2] = \mathbb{E}[x^2] - \mathbb{E}[x]^2,$$

where $\mu = \mathbb{E}[x]$ is the mean.

Exercise 2: prove that $\mathbb{E}[(x-\mu)^2] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$.

Exercise 3: show that $\text{var}[cx] = c^2\,\text{var}[x]$.

Exercise 4: show that $\text{var}[x+y] = \text{var}[x] + \text{var}[y]$, for independent outcomes x and y.

Exercise 5: Given outcomes distributed with mean μ and variance σ^2, how could you shift and scale them to have mean zero and variance one?

Change of units: If the outcome x is a height measured in metres, then x^2 has units m^2; x^2 is an area. The variance also has units m^2, it cannot be represented on the same scale as the outcome, because it has different units. If you multiply all heights by 100 to convert to centimetres, the variance is multiplied by 100^2. Therefore, the relative size of the mean and the variance depends on the units you use, and so often isn't meaningful.

Standard deviation: The standard deviation σ, the square root of the variance, *does* have the same units as the mean. Therefore it is a meaningful number to use as a typical distance from the mean. Often variances are used in intermediate calculations because they are easier to deal with: it is variances that add, not standard deviations.

5 Sums of independent variables: "random walks"

A drunkard starts at the centre of an alleyway, with exits at each end. He takes a sequence of random staggers either to the left or right along the alleyway. His position after N steps is $k_N = \sum_{n=1}^{N} x_n$, where the outcomes, $\{x_n\}$, the staggering motions, are drawn from some ensemble with zero mean and finite variance σ^2. For example $\mathcal{A}_X = \{-1, +1\}$ with $\mathcal{P}_X = \{1/2, 1/2\}$, which has $\mathbb{E}[x_n]=0$ and $\mathrm{var}[x_n]=1$.

If the drunkard started in the centre of the alleyway, will he ever escape? If so, roughly how long will it take? (If you don't already know, have a think…)

The expected, or mean position after N steps is $\mathbb{E}[k_N] = N\mathbb{E}[x_n] = 0$. This doesn't mean we don't think he'll escape. There are ways of escaping both left and right, and 'on average' he'll stay in the middle.

The variance of the position is $\mathrm{var}[k_N] = N\mathrm{var}[x_n] = N\sigma^2$. The standard deviation of the position is then $\mathrm{std}[k_N] = \sqrt{N}\sigma$, and is a measure of the width of the distribution over the distance from the centre of the alleyway. If we double the length of the alley, then it will typically take four times the number of random steps to escape.

Worthwhile remembering: *the sum of N independent variables scales with $\sqrt{N}$.* Sometimes you might have to work out the σ for your problem, or do other detailed calculations. But sometimes the scaling of the width of the distribution is all that really matters.

Solutions

As always, you are *strongly* recommended to work hard on a problem yourself before looking at the solutions. As you transition into doing research, there won't be any answers, and you have to build confidence in getting and checking your own answers.

Exercise 1: For independent outcomes x and y, $p(x, y)=p(x)p(y)$ and so
$\mathbb{E}[f(x)g(y)] = \sum_x \sum_y p(x)p(y)f(x)g(y) = \sum_x p(x)f(x) \sum_y p(y)g(y) = \mathbb{E}[f(x)]\mathbb{E}[g(y)]$.

Exercise 2: $\mathbb{E}[(x-\mu)^2] = \mathbb{E}[x^2 + \mu^2 - 2x\mu] = \mathbb{E}[x^2] + \mu^2 - 2\mu\mathbb{E}[x] = \mathbb{E}[x^2] - \mu^2$.

Exercise 3: $\mathrm{var}[cx] = \mathbb{E}[(cx)^2] - \mathbb{E}[cx]^2 = \mathbb{E}[c^2x^2] - (c\mathbb{E}[x])^2 = c^2(\mathbb{E}[x^2] - \mathbb{E}[x]^2) = c^2\mathrm{var}[x]$.

Exercise 4: $\mathrm{var}[x+y] = \mathbb{E}[(x+y)^2] - \mathbb{E}[x+y]^2 = \mathbb{E}[x^2] + \mathbb{E}[y^2] + 2\mathbb{E}[xy] - (\mathbb{E}[x]^2 + \mathbb{E}[y]^2 + 2\mathbb{E}[x]\mathbb{E}[y]) = \mathrm{var}[x] + \mathrm{var}[y]$, if $\mathbb{E}[xy]=\mathbb{E}[x]\mathbb{E}[y]$, true if x and y are independent variables.

Exercise 5: $z = (x-\mu)/\sigma$ has mean 0 and variance 1. Note division by the standard deviation, not the variance. Prove this result for yourself by applying the other results in this note.

Notice that using the expectation notation where possible, rather than writing out the summations explicitly, makes the answers concise.