Quantitative Risk Management
Case for Week 2

John A. Dodson

September 13, 2017

Primer: Entropy

The entropy of a (univariate) random variable is a measure of how much information you know about the eventual value it will take. This may seem like a very abstract notion, but it turns out to be very important in the setting of fitting probability models to data. In fact, one might describe the goal of fitting a model to data as finding the version of the model that minimizes the average entropy of each observation.

$$H_X \triangleq -\operatorname{E}\log f_X(X)$$

where $f_X(\cdot)$ is the probability density function[1] and the operator represent expectation.

- The entropy of a non-random (or degenerate random) variable is zero.
- The entropy of a fair coin toss is $\log 2 \approx 0.7$ "nats" (or 1 "bit" using the base-two logarithm).
- The entropy of a uniformly random "byte" (a discrete random variable which can take can take on any of 256 equally-probable values) is 8 bits or about 5.5 nats.

You might intuit that, for a continuous random variable with a location and a dispersion, the value of the location parameter probably does not matter to the entropy, but the value of the dispersion parameter probably does. After all, the notion of variance evokes uncertainly.

Let's explore this intuition. For $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, we have

$$-\log f_X(X) = \log\sqrt{2\pi\sigma^2} + \frac{(X-\mu)^2}{2\sigma^2}$$

so

$$H_X = \log\sqrt{2\pi e\sigma^2}$$

which is an increasing function in the variance parameter σ^2. The entropy of a standard normal is $\log\sqrt{2\pi e} \approx 1.4$ nats or just over two bits.

What about a more general case? Say you have a univariate random variable X with a known entropy H_X, and you are interested in the entropy of a new random variable $Y = aX + b$ for some fixed numbers a and b. We call this an "affine transformation", and it encompasses any arbitrary shift or change of units of the mean or standard deviation (if these quantities exist).

[1] The measure $\mu(\sigma)$ of a set σ from the sigma algebra is $\int_{x\in\sigma} f_X(x)\,dx$

Since the probability density of Y is

$$f_{aX+b}(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

in terms of the probability density of X,

$$H_Y = -\int \log\left(\frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)\right) \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)\, dy$$

Changing variables and expanding the logarithm,

$$\begin{aligned} &= -\int \left(\log f_X(x) - \log|a|\right) f_X(x)\, dx \\ &= H_X + \log|a| \end{aligned}$$

so we see that indeed entropy is unchanged under a location transformation and increasing under a scale transformation.

It might be useful to think of entropy as a generalization of variance.

Exercise: Bivariate Normal

The bivariate normal has five parameters: two means, two standard deviations, and a correlation. The standard version has only one parameter, $-1 \leq \rho \leq 1$. Its density is

$$f_{(X,Y)}(x,y) = \frac{1}{2\pi} e^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}} \frac{1}{\sqrt{1-\rho^2}}$$

1. Evaluate the marginal density of Y
2. Calculate its entropy
3. Evaluate the (conditional) density of $Y|X$ for the event $\mathrm{P}\{X=1\}=1$
4. Show that conditioning had reduced its entropy

Solution: Joint Density

The density for a multivariate normal random variable is

$$f_{\vec{X}}(x) = (2\pi)^{-n/2}\, e^{-\frac{1}{2}(\vec{x}-\vec{\mu})'\Sigma^{-1}(\vec{x}-\vec{\mu})} \sqrt{|\Sigma^{-1}|}$$

for positive-definite covariance matrix Σ of dimension n and mean vector $\vec{\mu}$.

For the exercise, we have $n=2$, $\mu=0$, and

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

with correlation parameter $-1 < \rho < 1$. Hence

$$\Sigma^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$$

and the density simplifies to

$$f_{(X,Y)}(x,y) = \frac{1}{2\pi} e^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}} \frac{1}{\sqrt{1-\rho^2}}$$

Solution: Conditional Density

According to the exercise, we are interested in the event $\mathrm{P}\{X = 1\} = 1$ and what it tells us about Y. Prior to the observation, we only have the marginal density for Y,

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)\, dx \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \end{aligned}$$

which has entropy

$$H_Y = \mathrm{E}\left[-\log f_Y(Y)\right] = \log\sqrt{2\pi e}$$

Once we have the observation on X, we can work with the conditional density,

$$\begin{aligned} f_{Y|X}\left(y\right) &= \frac{f_{(X,Y)}\left(1,y\right)}{\int_{-\infty}^{\infty} f_{(X,Y)}(1,y)\, dy} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\rho)^2}{2(1-\rho^2)}} \frac{1}{\sqrt{1-\rho^2}} \end{aligned}$$

which has entropy

$$H_{Y|X} = \mathrm{E}\left[-\log f_{Y|X}\left(Y\right)\middle|\, X\right] = \log\sqrt{2\pi e\left(1-\rho^2\right)}$$

Clearly $H_{Y|X} \le H_Y$. To the extent that there is correlation, the observation acts to lower the entropy of the r.v. we are interested in describing.

Extension: Regression

We have demonstrated that (in the bivariate normal case at least) conditioning conveys information. We can see this in a more familiar light by considering the general bivariate result with four additional parameters and an arbitrary conditioning event, $\mathrm{P}\{X = x\} = 1$.

You can confirm that the general conditional density works out to be

$$f_{Y|X}\left(y\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}\left(x-\mu_X\right)\right)^2}{2\sigma_Y^2\left(1-\rho^2\right)}} \frac{1}{\sigma_Y\sqrt{1-\rho^2}}$$

In other words,

$$Y|X \sim \mathcal{N}\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}\left(x - \mu_X\right), \sigma_Y^2\left(1 - \rho^2\right)\right)$$

suggesting the transformation

$$Y|X \triangleq \alpha + \beta x + \epsilon \tag{1}$$

where

$$\beta = \frac{\operatorname{cov}\left(X, Y\right)}{\operatorname{var} X} \tag{2a}$$

$$\alpha = \operatorname{E} Y - \beta \operatorname{E} X \tag{2b}$$

$$\epsilon \sim \mathcal{N}\left(0, \operatorname{var} Y - \beta^2 \operatorname{var} X\right) \tag{2c}$$

You may recognize this as the “population” result from classical statistics for ordinary least squares regression.

This observation is useful because it gives us guidance not only on grounding classical regression in modern estimation theory, but also on how we might adapt to situations where relationships are not well described by linear correlations.