Probability Unit 5

Topic: Entropy and measures of association

- Date: @February 22, 2024
- Lane: *Probability, Statistics*

What it covers

- Joint Distributions, Entropy measures, covariance (tests of dependence)

Requires

Random variables, discrete distributions, expectation, variance

Required by

Linear regression

Density estimation

Entropy as a measure of informativeness

By taking the expected value of the log of a probability distribution one gets a quantity that increases as the probability is more diffuse and uncertain.

$$H(P(X)) = -E[\log_2 P(x)] = -\sum_i P(x_i) \log_2 P(x_i)$$

where the negative sign makes the entropy positive since the log terms of quantities less than one are negative.

Joint Entropy, similar to the chain rule for probabilities

$$H(P, Q) = H(P) + H(Q \mid P) = H(Q) + H(P \mid Q)$$

Similar to probabilities one can define the entropy of a conditional probability:

$$H(P(X \mid Y)) = -E_{xy}[\log_2 P(X \mid Y)] = -\sum_{ij} P(x_i y_j) \log_2 P(x_i \mid y_j)$$

However the conditional entropy is a scalar, not a function of *Y* the conditioning r.v.

Typically entropy is attributed to the “amount of information” or more precisely the lack thereof represented by the distribution. The idea is, as if one was expecting an uncertain message or signal, the higher the entropy, the less certain one is about the content of the message and hence more informative. A message where one can entirely anticipate its content has an entropy equal to zero. The term, originally developed in communication theory - sometimes called “Shannon Entropy” - is borrowed from Thermodynamics, where entropy is a measure of the disorder attributed to the number of possible states that a system - say that a collection of atoms- can be in. The convention is to use log to the base 2, so the units become the number of “bits” needed to encode the message.

For discrete probability distributions of a (finite) number of items, entropy is a linear function of the log of the number of combinations of ways in which the items can be distributed.

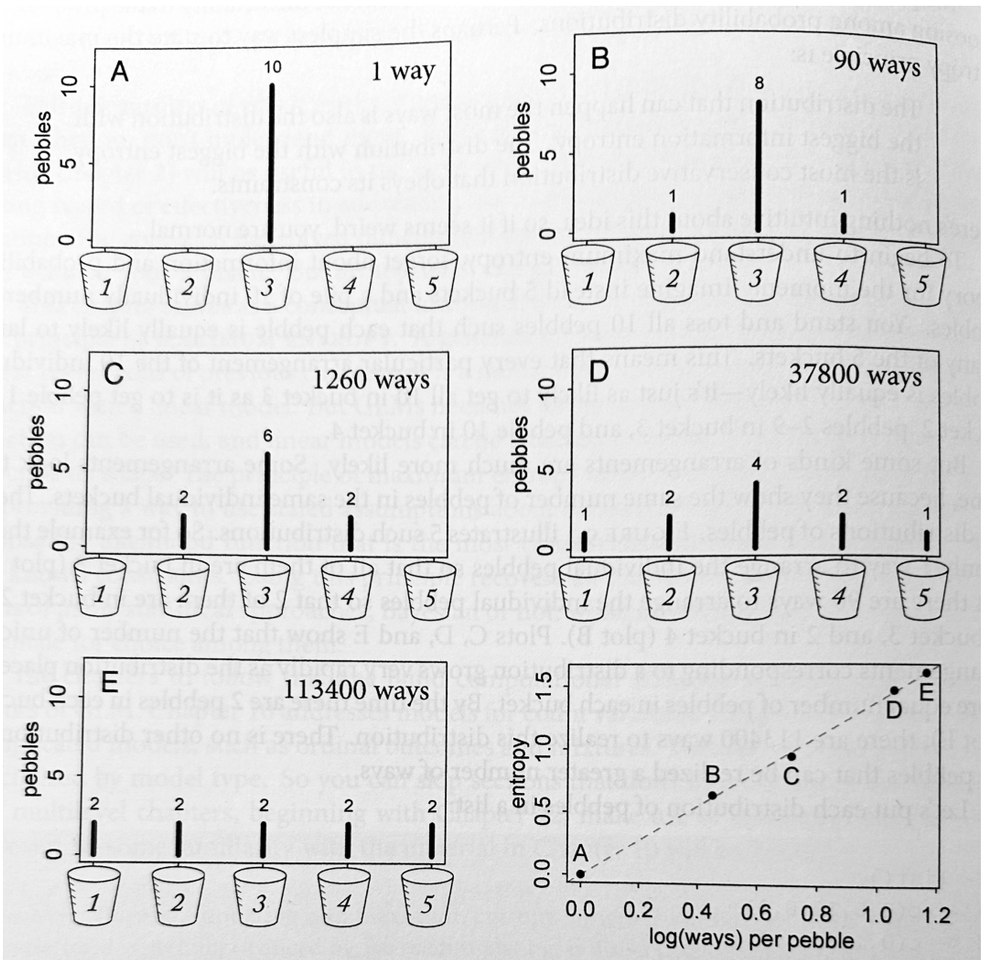

FIGURE 9.1. Entropy as a measure of the number of unique arrangements of a system that produce the same distribution. Plots A through E show the numbers of unique ways to arrange 10 pebbles into each of 5 different distributions. Bottom-right: The entropy of each distribution plotted against the log number of ways per pebble to produce it.

(From R. McElreath “Statistical Rethinking” (2016) CRC Press p.270)

Entropy

- increases with the number of possibilities (the size of the state space),
- is largest when probabilities are equally distributed over the states, and
- is an additive property of probability distributions.

So in cases where we multiple probabilities, we add their entropies.

Background - joint probability distributions.

Joint probability distributions $P(E_1, E_2, \ldots)$ are distributions of two or more random variables. For discrete events they assign a probability over all combinations of the events. This is equivalent to the distribution over all intersections of the events. For two binary events this is shown by a 2 by 2 table:

	E2	not E2
E1	P(E1 and E2)	P(E1 and not E2)
not E1	P(not E1 and E2)	P(not E1 and not E2)

As shown in the table, one can look for interdependence among the rows and columns, by coming up with various measures. Independence, such as $P(AB) = P(A)P(B)$ would be the case if the table was a rank one matrix, formed by the product of the probabilities in as vectors AB^T.

For continuous r.v.s variance has a natural extension to *covariance,* which recovers the linear dependence between a pair of variables. Similar are the *partial regression coefficients,* the "betas" in multi-variate regression. Regression coefficients are normalized covariances (in the orthogonal case):

$$\beta_{y.x} = \frac{\operatorname{cov}(X,Y)}{\operatorname{cov}(X,X)}$$

Normalizing the covariance for both variables obtains the *correlation coefficient:*

$$\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\operatorname{cov}(X,X)\operatorname{cov}(Y,Y)}$$

All three of these measures are linear measures for continuous r.v.s, and apart from their scaling, agree on the degree of dependence.

Mutual Information, cross entropy

Entropy-based measures of dependence apply to discrete variables, (in fact just event probabilities are required, although they have been expanded to continuous variables) and are not limited to linear dependence between r.vs.

Mutual Information

$$MI(P;Q) = MI(Q;P) = H(P) - H(P \mid Q) = H(Q) - H(Q \mid P)$$

When *P & Q* are independent, then both terms are equal, and mutual information is zero. It is maximized when conditional probabilities approach 0 or 1, so that the second term goes to zero and mutual information approaches the entropy of *P.*

Mutual Information can be used to rate which features have more predictive value in a regression when the feature and the outcome variables are not continuous variables.

Cross entropy

Cross entropy measures similarity between distributions.

$$CE(P,Q) = -E_P[\log Q] = -\sum_i P(x_i) \log Q(x_i)$$

The definition of entropy has an interesting property that can be used to compare similarity — called the "deviance" — between probability distributions. When the log probability is computed for a different distribution than used for taking the expectation, the result is never larger than the entropy of the distribution. It is maximized when *P* and Q are the same.

Cross entropy is used extensively as the optimization criterion in deep learning, when models are used to predict probabilities.

Incidentally both cross entropy and mutual information derive from another entropy-based measure: *KL divergence.*

Maximum (log) Likelihood, "Information criteria"

Maximum likelihood is a general purpose method for estimation — for fitting probability models to data, such as fitting regressions. It reduces statistics problems to optimization.

Note the similarity between log likelihood and cross entropy by comparing one distribution to another. Log likelihood, when applied to data compares the empirical distribution of the data to that of an estimated model (for example a prediction). Maximization of the log-likelihood gives the best model fit. The optimal parameters it finds are an estimate of the most likely values of the model parameters. (Hence the name.)

Maximum likelihood can be used to compare models. However it is a relative measure, for different models applied to the same data, unlike accuracy-based measures that attempt to compare models

Maximum likelihood obtains a point estimate of the parameter. To get interval estimates we will resort to simulation, using “the bootstrap”.

Likelihood is the conditional probability considered as a function of the parameters beta. Log likelihood converts the product of the likelihoods (assuming “identical independently distributed x-s) to a sum:

$$\Lambda(\beta) = \Sigma_i \log P(x_i \mid \beta) = \log \prod_i N(x_i \mid \beta)$$

So for the normal distribution the terms are:

$$\log N(x_i \mid \beta) = \log \left((2\pi\sigma^2)^{-1/2} \exp(-\frac{1}{2\sigma^2}(y - \beta x_i)^2) \right) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - \beta x_i)^2$$

Linear regression via Maximum Likelihood

Here’s how the expression for the beta parameters in linear regression are derived by maximum likelihood. If assume the data is normally distributed then to only term that remains after taking the derivative is $\frac{1}{2\sigma^2}(y - \beta_0 - \beta_1 x_i)^2$ and the maximum likelihood equations are identical to the sum of squares minimization:

So, when maximizing the log likelihood, the terms are the same as when minimizing RSS:

$$0 = \frac{d\Lambda(\beta)}{d\beta} = \frac{RSS(\beta)}{d\beta}$$

Read:

Evans, Chapter 6.1, 6.2 Maximum Likelihood Estimation

Evans Chapter 10.1 For using regression coefficients to define dependence among variables.

Assignment

1. Use the definition of Entropy to derive the definition of mutual information:

$$MI(Q;P) = H(P) - H(P \mid Q)$$

2. Derive:

$$MI(Q;P) = -\sum_i \sum_j P(x_i, y_j) \log[\frac{P(x_i)P(y_j)}{P(x_i, y_i)}]$$

Note the minus sign. How would the formula change without the minus sign?

References

Our textbooks don’t cover entropy except as specifics about different machine learning models.

A detailed sourcebook is
David J. C. MaKay “Information Theory, Inference and Learning Algorithms” (2003) Cambridge.

The content from this unit come from R. McElreath “Statistical Rethinking” (2016) CRC Press

There are a set of lectures based on the book on youtube that are a good introduction to Bayesian (anti-) statistics. They are a great follow on to this course:

https://www.youtube.com/watch?v=FdnMWdICdRs&list=PLDcUM9US4XdPz-KxHM4XHt7uUVGWWVSus&index=1

Comprehensive libraries for ...