Data Mining Tutorial

Mark A. Austin

University of Maryland

austin@umd.edu

ENCE 688P, Fall Semester 2021

October 16, 2021

Overview

Part 02

Entropy

(Quantitative Measure of Uncertainty)

Definition

Definition of Entropy

As it relates to machine learning, entropy is is a measure of the randomness (disorder or uncertainty) of information being processed.

Simple Example: Tossing a Fair Coin (High Entropy):

- A fair coin has no affinity (or preference) for heads or tails.
- The outcome any number of tosses is difficult to predict because there no relationship between coin flipping and the outcome.

Mathematical Models of Entropy

Principle of Maximum Entropy (Jaynes, 1957)

Given some partial information about a random variate, we should choose the probability distribution that is is consistent with the given information (e.g., boundary constraints), but otherwise has maximum entropy associated with it.

Relationship of Entropy to Uncertainty and Probability

- Every probability distribution has some uncertainty associated with it. Entropy provides a quantitative measure of this uncertainty.
- A principle goal of data mining models and algorithms is to reduce uncertainty.

Measuring Uncertainty of a Probability Distribution:

Definition of a Probability Distribution:

Let the probabilities of n possible outcomes A_1, A_2, $\cdots$, A_n, of an experiment be p_1, p_2, $\cdots$, p_n, respectively. The distribution:

$$P = (p_1, p_2, p_3, \cdots, p_n), \tag{2}$$

satisfies the constraints:

$$\sum_{i=1}^{n} p_i = 1, \tag{3}$$

and

$$p_1 \geq 0, p_2 \geq 0, \cdots, p_n \geq 0. \tag{4}$$

Measuring Uncertainty of a Probability Distribution

Requirements for Measuring Uncertainty (Kapur, 1989):

- It should be a function of $p_1, p_2, \cdots, p_n$, i.e.,

$$H = H_n(P) = H(p_1, p_2, \cdots, p_n). \tag{5}$$

- $H_n(P)$ should be a continuous and symmetric function.
- The maximum value of H_n should increase as n increases.
- It should be minimum (and possibly zero) when there is no uncertainty about the outcome. In other words, it should vanish when one of the outcomes is certain.

$$H_n(P) = 0 \text{ when } p_i = 1 \text{ and } p_j = 0, \ (j \neq i). \tag{6}$$

Measuring Uncertainty of a Probability Distribution

- H_n should be maximum when there is maximum uncertainty, which arises when the outcomes are equally likely, i.e.,

$$p_1 = p_2 = \cdots = p_n = \frac{1}{n}. \tag{7}$$

- For two independent probability distributions P and Q,

$$\sum_{i=1}^{n} p_i = 1, \text{ and } \sum_{j=1}^{m} q_j = 1, \tag{8}$$

the uncertainty of the joint scheme $P \cup Q$ should be:

$$H_{m+n}(P \cup Q) = H_n(P) + H_m(Q). \tag{9}$$

If P and Q have outcomes $A_1, A_2, \cdots, A_n$ and $B_1, B_2, \cdots, r_n$, then the joint outcomes are $A_i B_j$ with probabilies $p_i q_j$.

Mathematical Models of Entropy

Shanon's Measure of Entropy

Shanon (1949) proposed the following measure:

$$H_n(P) = \sum_{i=1}^{n} p_i ln(\frac{1}{p_i}) = - \sum_{i=1}^{n} p_i ln(p_i). \qquad (10)$$

Intial Observations:

- This function is continuous, symmetric, and convex.
- When one of the probabilities is 1, the others are zero. The entropy is zero and is a minimum value – no surprise.
- All of the commonly used probability distributions – uniform, normal, poisson, logarithmic – can be framed in terms of maximum entropy subject to constraints.

Mathematical Models of Entropy

Maximum Value of Entropy

We can use Lagrange's equations to find a maximum value, i.e.

$$-\sum_{i=1}^{n} p_i ln(p_i) - \lambda \left[\sum_{i=1}^{n} p_i - 1 \right]. \tag{11}$$

This gives (uniform distribution):

$$p_1 = p_2 = \cdots = p_n = \frac{1}{n}. \tag{12}$$

The maximum value of H_n is:

$$H_n = -\sum_{i=1}^{n} \frac{1}{n} ln(\frac{1}{n}) = ln(n) \rightarrow \text{ increases linearly with n.} \tag{13}$$

Mathematical Models of Entropy

Illustrative Example

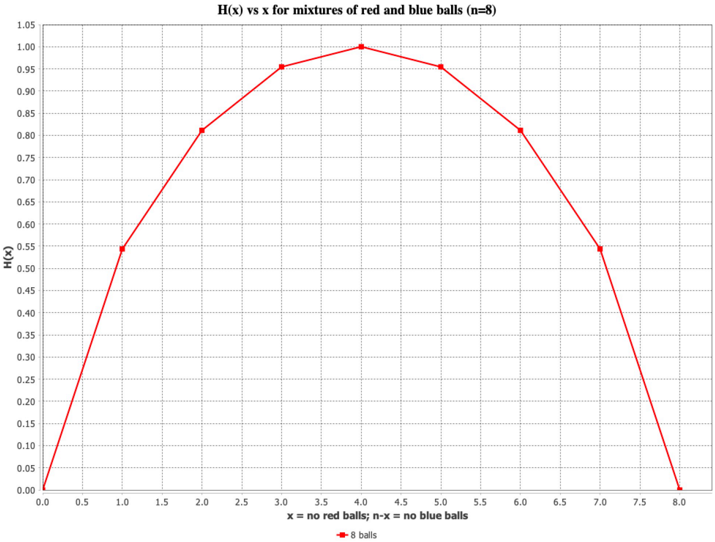
Suppose that an urn contains a mixture of red (n_r) red and blue (n_b) balls (i.e., $n = n_r + n_b$). The entropy is:

$$H_2(P) = - \left[\frac{n_r}{n} \right] \log_2 \left[\frac{n_r}{n} \right] - \left[\frac{n_b}{n} \right] \log_2 \left[\frac{n_b}{n} \right]. \tag{14}$$

Sample Calculation. Let $n_r = 2$, $n_b = 6$.

$$\begin{aligned} H_2(P) &= - \left[\frac{2}{8} \right] \log_2 \left[\frac{2}{8} \right] - \left[\frac{6}{8} \right] \log_2 \left[\frac{6}{8} \right] \\ &= \frac{1}{4} \cdot 2.0 + \frac{3}{4} \cdot 0.415 = 0.811 \end{aligned} \tag{15}$$

Mathematical Models of Entropy



Mathematical Models of Entropy

Key Points:

- Minimum values of entropy occur when the urn contains only red balls (i.e., x = 0) or only blue balls (i.e., x = 8). There is no disorder.
- The maximum value of entropy occurs when the urn system has maximum disorder – that is, four blue balls and four red balls.

$$H_2(P) = -\left[\frac{4}{8}\right]\log_2\left[\frac{4}{8}\right] - \left[\frac{4}{8}\right]\log_2\left[\frac{4}{8}\right] = 1.0 \tag{16}$$

- Even higher levels of entropy (disorder) can be obtained by adding more colors to the urn, e.g., 2 blue balls, 2 green balls, 3 red balls, 1 purple ball. Now, P = $(\frac{1}{4}, \frac{1}{4}, \frac{3}{8}, \frac{1}{8})$.

References

- Jaynes E.T., Information Theory and Statistical Mechanics. II, Phys. Rev. 108, 171, October 1957.
- Kapur J.N., Maximum-Entropy Models in Science and Engineering, John Wiley and Sons, 1989.
- Mitchell T.M., Machine Learning and Data Mining, Communications of the ACM, Vol. 42., No. 11, November 1999.
- Russell S., and Norvig P., Artificial Intelligence: A Modern Approach (Third Edition), Prentice-Hall, 2010.
- Shanon C.E., and Weaver W., The Mathematical Theory of Communication, University of Illinois, Urbana, Chicago, 1949.
- Witten I.H., Frank E., Hall M.A., and Pal C.J., Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2017.