Information Theory

Jianhua MEI

RSE Machine Learning and Econometrics Reading Group

November 6, 2023

Introduction to Information Theory

In this chapter, we introduce a few basic concepts from the field of information theory.

- Entropy
- Cross Entropy
- Joint Entropy
- Conditonal Entropy
- Relative Entropy(KL Divergence)
- Mutual information
- Conditional Mutual Information
- Normalized Mutual Information
- Maximal Information Coefficient

Supervised Learning - Decision Trees

(a)

(b)

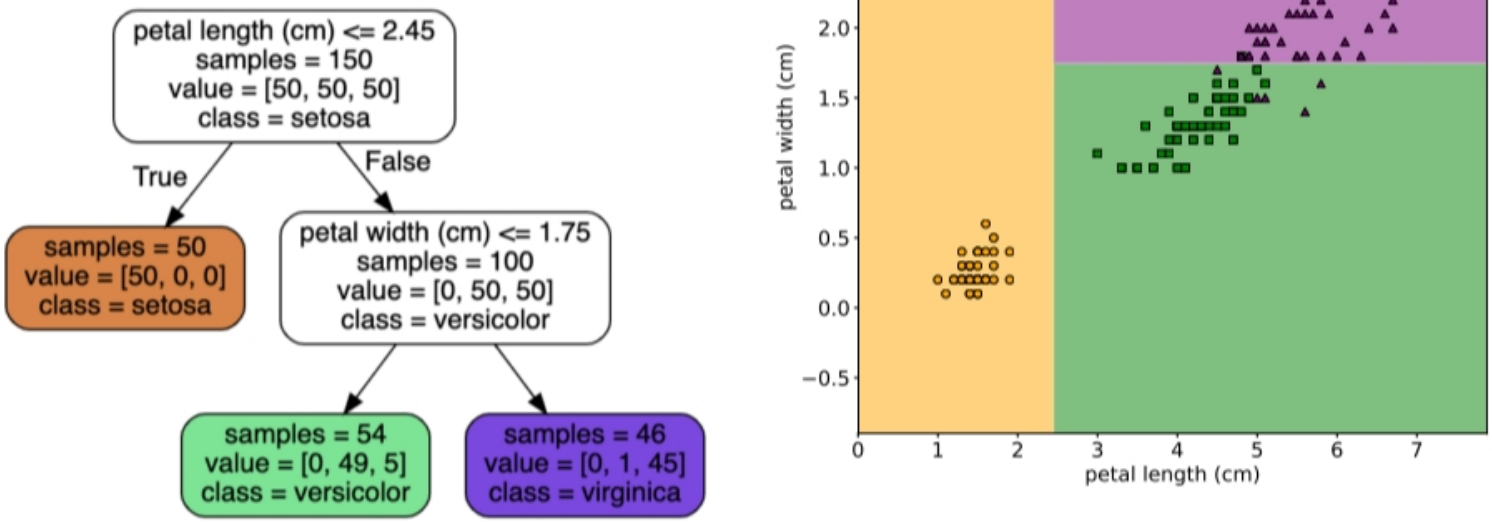
(c)

Figure 1.1: Three types of Iris flowers: Setosa, Versicolor and Virginica. Used with kind permission of Dennis Kramb and SIGNA.

index	sl	sw	pl	pw	label
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
	...				
50	7.0	3.2	4.7	1.4	Versicolor
	...				
149	5.9	3.0	5.1	1.8	Virginica

Table 1.1: A subset of the Iris design matrix. The features are: sepal length, sepal width, petal length, petal width. There are 50 examples of each class.

Supervised Learning - Decision Trees



(a) (b)

Figure 1.4: Example of a decision tree of depth 2 applied to the Iris data, using just the petal length and petal width features. Leaf nodes are color coded according to the predicted class. The number of training samples that pass from the root to a node is shown inside each box; we show how many values of each class fall into this node. This vector of counts can be normalized to get a distribution over class labels for each node. We can then pick the majority class. Adapted from Figures 6.1 and 6.2 of [Gér19]. Generated by iris_dtree.ipynb.

Entropy

- The entropy of a probability distribution can be interpreted as a measure of uncertainty, or lack of predictability, associated with a random variable drawn from a given distribution.
- We can also use entropy to define the *information content* of a data source.
- For example, suppose we observe a sequence of symbols $X_n \sim p$ generated from distribution p. If p has high entropy, it will be hard to predict the value of each observation X_n. Hence we say that the dataset $D = (X_1, \ldots, X_n)$ has high information content.

Entropy for discrete random variables

The entropy of a discrete random variable X with distribution p over K states is defined by

$$H(X) \triangleq -\sum_{k=1}^{K} p(X=k) \log_2 p(X=k) = -E_X[\log_2 p(X)]$$

- Usually we use log base 2, in which case the units are called bits.
- If we use log base e, the units are called nats.
- For example, if $X \in 1, ..., 5$ with histogram distribution p = [0.25, 0.25, 0.2, 0.15, 0.15], we find H = 2.29 bits.
- The discrete distribution with maximum entropy is the uniform distribution. Hence for a K-ary random variable, the entropy is maximized if $p(x=k) = 1/K$; in this case, $H(X) = \log_2 K$.

Entropy for discrete random variables

The discrete distribution with maximum entropy is the uniform distribution. Hence for a K-ary random variable, the entropy is maximized if $p(x = k) = 1/K$; in this case, $H(X) = \log_2 K$. To see this, note that

$$H(X) = -\sum_{k=1}^{K} \frac{1}{K} \log(1/K) = -\log(1/K) = \log(K)$$

Entropy for discrete random variables

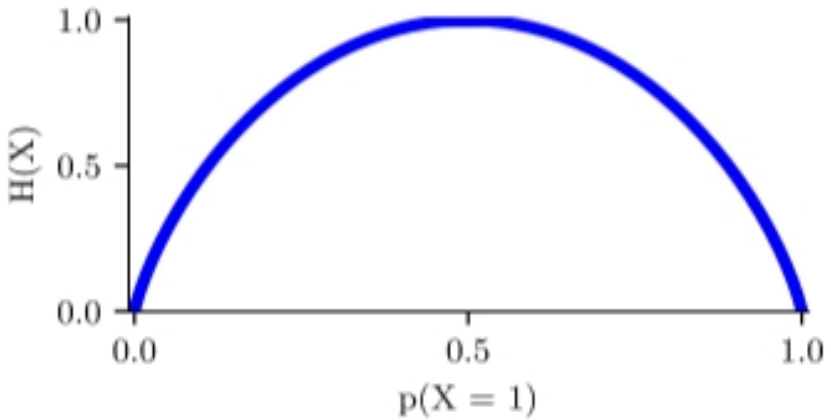


Figure 6.1: Entropy of a Bernoulli random variable as a function of θ. The maximum entropy is $\log_2 2 = 1$. Generated by bernoulli_entropy_fig.ipynb.

Cross Entropy

The cross entropy between distribution p and q is defined by

$$H_{ce}(p, q) \triangleq -\sum_{k=1}^{K} p_k \log q_k \quad (6.7)$$

One can show that the cross entropy is the expected number of bits needed to compress some data samples drawn from distribution p using a code based on distribution q. This can be minimized by setting $q = p$, in which case the expected number of bits of the optimal code is $H_{ce}(p, p) = H(p)$ — this is known as *Shannon's source coding theorem*.

Joint Entropy

The Joint entropu of two random variables X and Y is defined as

$$H(X, Y) = -\sum_{x,y} p(x, y) \log_2 p(x, y)$$

- If X and Y are independent, then $H(X, Y) = H(X) + H(Y)$
- If Y is a deterministic function of X, then $H(X, Y) = H(X)$.
- If X is a deterministic function of Y, then $H(X, Y) = H(Y)$.
- So

$$H(X, Y) \geq \max\{H(X), H(Y)\} \geq 0$$

Conditional Entropy

The conditional entropy of Y given X is the uncertainty we have in Y after seeing X, averaged over possible values for X:

$$\begin{aligned} H(Y|X) &= \mathbb{E}_{p(x)}\left[H(p(Y|X=x))\right] \\ &= \sum_x p(x) H(p(Y|X=x)) \\ &= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) \\ &= H(X,Y) - H(X) \end{aligned}$$

- $H(Y|X) \leq H(Y)$ with equality iff X and Y are independent.
- $H(Y|X) \geq 0$ with equality iff Y is a deterministic function of X.

Relative entropy (KL divergence)

Given two distributions p and q, it is often useful to define a distance metric to measure how "close" or "similar" they are.
For discrete distributions, the KL divergence is defined as follows:

$$D_{KL}(p \parallel q) \triangleq \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}$$

This naturally extends to continuous distributions as well:

$$D_{KL}(p \parallel q) \triangleq \int p(x) \log \frac{p(x)}{q(x)} dx$$

We can rewrite the KL as follows:

$$D_{KL}(p \parallel q) = \sum_{k=1}^{K} p_k \log p_k - \sum_{k=1}^{K} p_k \log q_k = -H(p) + H_{ce}(p, q)$$

Lower bound:

$$D_{KL}(p \parallel q) \geq 0 \quad \text{with equality iff} \quad p(x) = q(x)$$

Mutual information

The KL divergence gave us a way to measure how similar two distributions were. How should we measure how dependant two random variables are? One thing we could do is turn the question of measuring the dependence of two random variables into a question about the similarity of their distributions.
The mutual information between rv's X and Y is defined as follows:

$$I(X;Y) \triangleq D_{\mathrm{KL}}(p(x,y) \parallel p(x)p(y)) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Show that:

$$I(X;Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

Thus we can interpret the MI between X and Y as the reduction in uncertainty about X after observing Y, or, by symmetry, the reduction in uncertainty about Y after observing X.

Normalized Mutual information

For some applications, it is useful to have a normalized measure of dependence, between 0 and 1. We now discuss one way to construct such a measure.

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \leq H(X) \\ &= H(Y) - H(Y|X) \leq H(Y) \end{aligned}$$

so

$$0 \leq I(X;Y) \leq \min(H(X), H(Y))$$

Therefore we can define the normalized mutual information as follows:

$$NMI(X,Y) = \frac{I(X;Y)}{\min(H(X), H(Y))} \leq 1$$

This normalized mutual information ranges from 0 to 1.

Maximal information coefficient

It is useful to have a normalized estimate of the mutual information, but this can be tricky to compute for real-valued data.

$$MIC(X, Y) = \max_G \left[\frac{I(X, Y)|_G}{\log ||G||} \right]$$

- G is the set of 2d grids
- $I(X, Y)|_G$ represents a discretization of the variables onto this grid
- $||G||$ is $\min(G_x, G_y)$, where G_x is the number of grid cells in the x direction, and G_y is the number of grid cells in the y direction.
- They suggest restricting grids so that $G_x G_y \leq B(n)$, where $B(n) = n^{0.6}$, where n is the sample size.

Maximal information coefficient

- Construct a grid over the range of the data for variables X and Y.
- Calculate the probability of data points for X and Y falling into each bin within this grid.
- Compute the mutual information on the grid based on these probabilities.
- Calculate the mutual information for all possible grid sizes and normalize each by a logarithmic scaling factor.
- The MIC value is the maximum of these normalized ratios.
- The MIC ranges from 0 to 1, where 0 indicates that the two variables are independent, and 1
- indicates the presence of some kind of perfect functional relationship between them.

Reference

[1] Kevin P.Murphy: Probabilistic Machine Learning Ch1