Machine learning - Information theory

Yonghoon Dong

March 8, 2024

Entropy

1. The **entropy** of a probability distribution can be interpreted as a measure of uncertainty.
2. We can also use entropy to define the **information content** of a data source.

For example, let we observe a sequence of symbols X_n 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 $\mathcal{D}$ **high information content**.

Entropy : Discrete case

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

$$\mathbb{H}(X) := -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k) = -\mathbb{E}_X[\log p(X)] \geq 0$$

If there is any confusion, we could alternatively write $\mathbb{H}(p(X))$ instead of $\mathbb{H}(X)$.

Entropy : Continuous case (Differential entropy)

Differential entropy (also referred to as continuous entropy) is a concept in information theory that began as an attempt by Claude Shannon to extend the idea of (Shannon) entropy, a measure of average (surprisal) of a random variable, to continuous probability distributions.

$$h(X) = -\int_{\mathcal{X}} p(x) \log_2 p(x) dx = -\mathbb{E}_X[\log p(X)]$$

Cross entropy

The **cross entropy** between distribution p and q is defined by

$$\mathbb{H}_{ce}(p, q) := -\sum_{k=1}^{K} p_k \log q_k$$

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 because $-\log q_k$ can be interpreted as a number of bits needed to compress.

Joint entropy

The **joint entropy** of two random variables X and Y is defined as

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

It satisfies the following property

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

Intuitively, this says combining variables together does not make the entropy go down: we can't reduce uncertainty merely by adding more unknowns to the problem

Conditional entropy

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

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

Conditional entropy

1. If Y is a deterministic function of X, then knowing X completely determines Y, so $\mathbb{H}(Y|X) = 0$
2. If X and Y are independent, $\mathbb{H}(Y|X) = \mathbb{H}(Y)$ because knowing X tells us nothing about Y
3. Since $\mathbb{H}(X, Y) \leq \mathbb{H}(Y) + \mathbb{H}(X)$, $\mathbb{H}(Y|X) \leq \mathbb{H}(Y)$ with equality iff X and Y are independent.

This shows that, on average, conditioning on data never increases one's uncertainty. The caveat **"on average"** is necessary because one may get $\mathbb{H}(Y|x) > \mathbb{H}(Y)$ for any particular observation value of X

Conditional entropy

We already know that

$$\mathbb{H}(X_1, X_2) = \mathbb{H}(X_1) + \mathbb{H}(X_2|X_1)$$

This can be generalized to get the **chain rule for entropy**

$$\mathbb{H}(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} \mathbb{H}(X_i|X_1, \ldots, X_{i-1})$$

Relative entropy (KL divergence)

Given two distribution p and q, it is often useful to define a **distance metric** to measure how close they are. We will be more general and consider a **divergence measure** $D(p, q)$ which quantifies how far q is from p. This metric satisfies

1. Positive definiteness : $D(p, q) \geq 0$ with equality iff $p = q$
2. Triangle inequality : $D(q, r) \leq D(p, q) + D(q, r)$

Note that the symmetry condition for satisfying the metric is **not** required.

Relative entropy (KL divergence)

There are many possible divergence measure we can use but normally people use **KL divergence**, also known as the **information gain** or **relative entropy**

Definition

For discrete distribution, the KL divergence is defined as

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

For a continuous distribution,

$$D_{\mathrm{KL}}(p \| q) := \int p(x) \log \frac{p(x)}{q(x)} dx$$

Interpretation of KL divergence

We can rewrite the KL divergence as follows

$$\begin{aligned} D_{\mathrm{KL}}(p \| q) &= \sum_{k=1}^{K} p_k \log p_k - \sum_{k=1}^{K} p_k \log q_k \\ &= -\mathbb{H}(p) + \mathbb{H}_{\mathrm{ce}}(p, q) \end{aligned}$$

Since $\mathbb{H}(p) \geq 0$, the cross entropy $\mathbb{H}_{\mathrm{ce}}(p, q)$ is a lower bound. Thus we can interpret the KL divergence as the "extra number of bits" we need to pay when compressing the data sample if we use the incorrect distribution q as the basis of our coding scheme compared to the true distribution p.

Convex set

Definition (Convex set)

Let S be a vector space or an affine space over the real numbers. A subset C of S is convex if

$$\forall x, y \in C, \theta x + (1-\theta)y \in C \text{ for all } 0 \leq \theta \leq 1$$

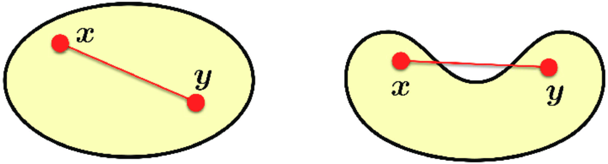

Figure: Geometric illustration of convex set and non-convex set

Epigraph

Definition (epigraph)

Let $f : R^n \to R$

$$\text{epi } f = \{(x, t) \in R^{n+1} |\ x \in \text{dom } f, f(x) \leq t\}$$

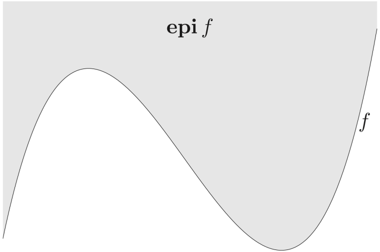

Figure: Geometric illustration of a epigraph

Convex function

Definition (Convex function)

$f : \mathbb{R}^n \to \mathbb{R}$ is convex if dom f is a convex set and

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$$

for all $x, y \in$ dom $f, 0 \leq \theta \leq 1$, or equivalently f is convex if and only if epi f is a convex set

Definition (Strictly convex function)

$f : \mathbb{R}^n \to \mathbb{R}$ is strictly convex if dom f is a convex set and

$$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta)f(y)$$

for all $0 < \theta < 1$ and $x, y \in X$ such that $x \neq y$

Convex function

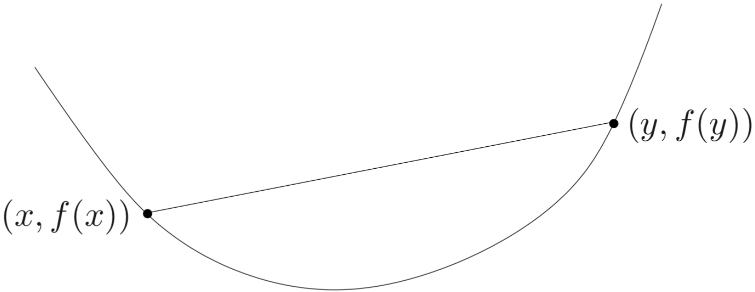

Figure: Geometric illustration of a epigraph

Jensen's inequality

Theorem (Jensen's inequality)

For any convex function f,

$$f(\sum_{i=1}^{n} \lambda x_i) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$

where $\lambda_i \geq 0$ and $\sum_{i=1}^{n} \lambda_i = 1$. In probability perspective, it can be written as

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(x)]$$

Proof.

Using the definition of the convex function recursively and take $\lambda_i = p(x)$ □

Non negativity of KL divergence

Theorem

Let $A = \{x : p(x) > 0\}$ *be the support of* $p(x)$*.*

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

Non negativity of KL divergence

Proof.

$$
\begin{aligned}
D_{\mathrm{KL}}(p\|q) &= -\sum_{x\in A} p(x)\log\frac{q(x)}{p(x)} && (1)\\
&\geq -\log\sum_{x\in A} p(x)\frac{q(x)}{p(x)} = -\log\sum_{x\in A} q(x) \text{ (By Jensen's inequality)} && (2)\\
&\geq -\log\sum_{x\in\mathcal{X}} q(x) = -\log 1 = 0 \text{ (Since } \log q(x)\leq 0, \text{ for all } x) && (3)
\end{aligned}
$$

Since $\log(x)$ is strictly concave function, equality holds only when $p(x)/q(x)$ is constant for all $x\in A$ where $\mathcal{X}\subset A$ □

Non negativity of KL divergence

Proof.

Let $p(x) = cq(x)$ for all $x \in A$. To satisfy (3), $\sum_{x \in \mathcal{X}} q(x) = \sum_{x \in A} q(x)$. Therefore,

$$1 = \sum_{x \in A} p(x) = c \sum_{x \in A} q(x) = c$$

Then, $c = 1$. Hence $D_{\mathrm{KL}}(p \| q) = 0$ if and only if $p(x) = q(x)$ for all x. $\square$