Information Theory

CHEN Si

Contents

1 Introduction

- A branch of applied mathematics
- Fundamental to many areas of electrical engineering and computer science

2 Goals

- In Electrical Engineering
  - Quantify how much information is present in a signal

  - Tells how to design optimal codes and calculate the expected length of messages sampled from specific probability distributions using various encoding schemes

- In Machine Learning
  - Characterize probability distributions
  - Quantify similarity between probability distributions

3 Self-Information

Motivation

- To quantify the information of events

Intuition

- Likely events should have low information content
- Less likely events should have higher information content
- Independent events should have additive information

Definition (Discrete Random Variable) The **self-information** of an event $\mathrm{x} = x$ is defined as

$$I(x) = -\log P(x)$$

Units:

1. **nats**: if log is the natural logarithm with base e
   - One nat is the amount of information gained by observing an event of probability $\frac{1}{e}$
   - In machine learning, we usually use nats
2. **bits** or **shannons**: if log is base-2
   - Information measured in bits is just a rescaling of information measured in nats

Definition (Continuous Random Variable) We use the same definition of information by analogy, but some of the properties from the discrete case are lost

- e.g. An event with unit density still has zero information, despite not being an event that is guaranteed to occur

4 Shannon Entropy

Motivation

- To quantify the amount of uncertainty in an entire probability distribution

Definition

$$H(P) = H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)]$$

- Gives a lower bound on the number of bits needed on average to encode symbols drawn from a distribution P
- Distributions taht are nearly deterministic (where the outcome is nearly certain) have low entropy; distributions that are closer to uniform have high entropy
- When x is continuous, the Shannon entropy is known as the **differential entropy**

5 Kullback-Leibler (KL) divergence

Motivation

- To measure how different two distributions (over the same random variable) are

Definition

$$D_{KL}(P\|Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}[\log P(x) - \log Q(x)]$$

- In the case of descrete variables, it is the extra amount of information needed to send a message containing symbols drawn from probability distribution P, when we use a code that was designed to minimize the length of messages drawn from probability distribution Q

Properties

- Non-negative (proof)
- The KL divergence is 0 if and only if P and Q are the same distribution in the case of discrete variables, or equal "almost everywhere" in the case of continuous variables
- Similar to some sort of distance between these distributions, but it is not symmetric $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$
- The choice of whether to use $D_{KL}(P\|Q)$ or $D_{KL}(Q\|P)$ depends on requirements (see Appendix A for details)

6 Cross-Entropy

Definition

$$H(P, Q) = -\mathbb{E}_{\mathrm{x}\sim P}\left[\log Q(x)\right]$$

Properties

- $H(P, Q) = H(P) + D_{KL}(P\|Q)$
- $\min_Q H(P, Q) \iff \min_Q D_{KL}(P\|Q)$
- $0 \log 0 = \lim_{x\to 0} x \log x = 0$ (by convention in information theory)

Appendix A Asymmetry of KL divergence

1. To minimize $D_{\mathrm{KL}}(p\|q)$
   (a) $q(x)$ needs to be large when $p(x)$ is large
   (b) $D_{\mathrm{KL}}(p\|q)$ is not sensitive to q when $p(x)$ is small
2. To minimize $D_{\mathrm{KL}}(q\|p)$
   (a) $q(x)$ needs to be small when $p(x)$ is small
   (b) $D_{\mathrm{KL}}(q\|p)$ is not sensitive to q when $p(x)$ is large

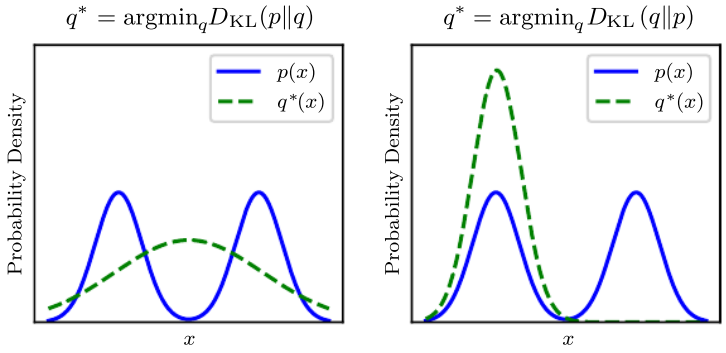

Figure 3.6: The KL divergence is asymmetric. Suppose we have a distribution $p(x)$ and wish to approximate it with another distribution $q(x)$. We have the choice of minimizing either $D_{\mathrm{KL}}(p\|q)$ or $D_{\mathrm{KL}}(q\|p)$. We illustrate the effect of this choice using a mixture of two Gaussians for p, and a single Gaussian for q. The choice of which direction of the KL divergence to use is problem dependent. Some applications require an approximation that usually places high probability anywhere that the true distribution places high probability, while other applications require an approximation that rarely places high probability anywhere that the true distribution places low probability. The choice of the direction of the KL divergence reflects which of these considerations takes priority for each application. *(Left)*The effect of minimizing $D_{\mathrm{KL}}(p\|q)$. In this case, we select a q that has high probability where p has high probability. When p has multiple modes, q chooses to blur the modes together, in order to put high probability mass on all of them. *(Right)*The effect of minimizing $D_{\mathrm{KL}}(q\|p)$. In this case, we select a q that has low probability where p has low probability. When p has multiple modes that are sufficiently widely separated, as in this figure, the KL divergence is minimized by choosing a single mode, to avoid putting probability mass in the low-probability areas between modes of p. Here, we illustrate the outcome when q is chosen to emphasize the left mode. We could also have achieved an equal value of the KL divergence by choosing the right mode. If the modes are not separated by a sufficiently strong low-probability region, then this direction of the KL divergence can still choose to blur the modes.

Appendix B Theorems & Proofs

Theorem 1 (Jensen's Inequality) If X is a random variable and f a convex function, then

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

Moreover, if f is strictly convex, the equality in implies that $X = \mathbb{E}[X]$ with probability 1 (i.e., X is a constant).

Proof We prove this for discrete distributions by induction on the number of mass points.

- For a two-mass-point distribution, the inequality becomes

  $$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$$

  which follows directly from the definition of convex functions.

- Suppose that the theorem is true for distributions with $k1$ mass points. Then writing $p'_i = p_i/(1 - p_k)$ for $i = 1, 2, \ldots, k-1$, we have

$$\begin{aligned}
\sum_{i=1}^{k} p_i f(x_i) &= p_k f(x_k) + (1 - p_k) \sum_{i=1}^{k-1} p'_i f(x_i) \\
&\geq p_k f(x_k) + (1 - p_k) f\left(\sum_{i=1}^{k-1} p'_i x_i \right) \\
&\geq f\left(p_k x_k + (1 - p_k) \sum_{i=1}^{k-1} p'_i x_i \right) \\
&= f\left(\sum_{i=1}^{k} p_i x_i \right)
\end{aligned}$$

□

(see more for continous case)

Theorem 2 (Information Inequality) Let $p(x)$, $q(x)$, $x \in \mathcal{X}$, be two probability mass functions. Then

$$D_{\mathrm{KL}}(p \| q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x.

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

$$-D(p||q) = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \tag{2.83}$$

$$= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \tag{2.84}$$

$$\leq \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \tag{2.85}$$

$$= \log \sum_{x \in A} q(x) \tag{2.86}$$

$$\leq \log \sum_{x \in \mathcal{X}} q(x) \tag{2.87}$$

$$= \log 1 \tag{2.88}$$

$$= 0, \tag{2.89}$$

where (2.85) follows from Jensen's inequality. Since $\log t$ is a strictly concave function of t, we have equality in (2.85) if and only if $q(x)/p(x)$ is constant everywhere [i.e., $q(x) = cp(x)$ for all x]. Thus, $\sum_{x \in A} q(x) = c \sum_{x \in A} p(x) = c$. We have equality in (2.87) only if $\sum_{x \in A} q(x) = \sum_{x \in \mathcal{X}} q(x) = 1$, which implies that $c = 1$. Hence, we have $D(p||q) = 0$ if and only if $p(x) = q(x)$ for all x. □