Information Theory

Jinhwan Suk

Department of Mathematical Science, KAIST

April, 2, 2020

Introduction

- Information theory is a branch of applied mathematics.
- Originally proposed by Claude Shannon in 1948.
- A key measure in information theory is entropy.

The basic intuition

Learning that an unlikely event has occurred is more informative than learning that a likely event has occurred.

Example

- Message 1 : "the sun rose this morning"
- Message 2 : "there was a solar eclipse this morning"

Message 1 is useless!!

Message 2 is much more informative than Message 1.

Formalization

We would like to quantify information in a way that formalizes this intuition.

- Likely events should have low information. And, events that are guaranteed to happen should have no information content whatsoever.
- Less likely events should have higher information content.
- Independent events should have additive information. e.g. a tossed coin has com up as head twice.

From the above intuitions, we can expect following properties of Information function $I(x) = I_X(x)$.

- $I(x)$ is a function of $P(x)$.
- $I(x)$ is inversely proportional to $P(x)$.
- $I(x) = 0$ if $P(x) = 1$.
- $I(X_1 = x_1, X_2 = x_2) = I(X_1 = x_1) + I(X_2 = x_2)$

Formalization

Write $I(x) = I(P(X = x))$,

$P(X_1 = x_1) = p_1$, and $P(X_2 = x_2) = p_2$.

Then, we can express the last property as

$$I(p_1 p_2) = I(p_1) + I(p_2).$$

Thus, $I(p) = k \log p$ for some $k < 0$.

Self-information

To satisfy all three of these properties, we define the **self-information** of an event $X = x$ to be

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

When X is continuous, we use the same definition but some of the properties from the discrete case are lost. (Property 2)

self-information is a measure of information(or, uncertainity) of a certain single event.

Shannon entropy

Shannon entropy quantifies the amount of uncertainity in an entire probability distribution.

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

The Shannon entropy of a distribution(or, random variable) is the expected amount of information in an event drawn from that distribution.

Entropy is a measure of the unpredictability of the state, or equivalently, of its average information content.

Classification problem

- In classification problem, we usually want to describe $\mathbb{P}(Y|X)$ for each input X.
- So many models(c_θ) aim to estimate conditional probability distribution by choosing optimal $\hat{\theta}$ such that

$$c_{\hat{\theta}}(x)[i] = \mathbb{P}(Y = y_i|X = x),$$

  like softmax classifier or Logistic regresor.
- So we can regard the classification problem as the regression problem such that minimizes

$$R(c_\theta) = \mathbb{E}_X[\mathcal{L}(c_\theta(X), \mathbb{P}(Y|X))]$$

  ($\mathcal{L}$ measures closeness between two probability distribution)

Total variation distance

Goal : Find an estimator $\hat{\theta}$ such that $\mathbb{P}_{\hat{\theta}}$ is close to $\mathbb{P}_{\theta^*}$.

This means : $|\mathbb{P}_{\hat{\theta}}(A) - \mathbb{P}_{\theta^*}(A)|$ is **small** for all event A.

Definition

The *total variation distance* between two probability measures $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta^*}$ is defined by

$$TV(\mathbb{P}_\theta, \mathbb{P}_{\theta^*}) = \max_{A:events} |\mathbb{P}_\theta(A) - \mathbb{P}_{\theta^*}(A)|.$$

Total variation distance measures the difference between two probability distribution only within the point of view toward Probability measure.

Kullback-Leibler divergence

Goal : Find an estimator $\hat{\theta}$ such that $\mathbb{P}_{\hat{\theta}}$ is close to $\mathbb{P}_{\theta^*}$.
Probabilitic view : $|\mathbb{P}_{\hat{\theta}}(x) - \mathbb{P}_{\theta^*}(x)|$ is **small** $\forall x \in \mathcal{X}$.
Informational view : $|\log \mathbb{P}_{\hat{\theta}}(x) - \log \mathbb{P}_{\theta^*}(x)|$ is **small** $\forall x \in \mathcal{X}$.

Definition

The *KL divergence* between two probability measures P and Q is defined by

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

$D_{KL}(P||Q)$ is the expected value of difference in information between two probability distribution P and Q with respect to P.

Cross-entropy

$$\begin{aligned} D_{KL}(P||Q) &= \mathbb{E}_{X \sim P}[\log P(x) - \log Q(x)] \\ &= \mathbb{E}_{X \sim P}[\log P(x)] - \mathbb{E}_{X \sim P}[\log Q(x)] \\ &= \textit{constant} - \mathbb{E}_{X \sim P}[\log Q(x)] \end{aligned}$$

Hence, minimizing the KL divergence is equivalent to minimizing $-\mathbb{E}_{X \sim P}[\log Q(x)]$, whose name is **cross-entropy**. And the estimation by using estimator that minimizes *KL divergence* or *Cross-entropy* is called **maximum likelihood principle**.

Loss function application

Return to Main Goal : Find an estimator $\hat{\theta}$ that minimizes

$$R(c_\theta) = \mathbb{E}_X[\mathcal{L}(c_\theta(X), \mathbb{P}(Y|X))].$$

Suppose that $X_1, X_2, ..., X_n$ are i.i.d and *cross-entropy* is used for $\mathcal{L}$.

$$\begin{aligned}
\mathbb{E}_X[\mathcal{L}(c_\theta(X), \mathbb{P}(Y|X))] &\sim \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(c_\theta(X_i), \mathbb{P}(Y|X_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n} -\mathbb{E}_{Y|X_i \sim \mathbb{P}_{Y_{true}|X_i}}[\log c_\theta(X_i)] \\
&= \frac{1}{n}\sum_{i=1}^{n} -\log\{c_\theta(X_i)[Y_{i,true}]\}.
\end{aligned}$$