Information Theory

Ben Gravell

benjamin.gravell@utdallas.edu
The Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080

Outline

Information theory

Information theory concerns quantifying the amount of information present in signals

- Originally developed for sending and receiving messages over communication channels
- Deals primarily with discrete random variables

Applications

- Machine learning e.g. classify images
- Reinforcement learning e.g. teach robots how to balance

c.f. Ch. 1-3 of Mackay's "Information Theory, Inference, and Learning Algorithms" [1]

c.f. Ch. 3 of Goodfellow's "Deep Learning" [2]

Information

Intuitively, we want a quantity that measures

- The amount of information communicated by an outcome
- How surprising an outcome is

Our definition of “information” or “surprise” should satisfy three axioms:

1. Certain events yield zero information
   - They always happen, so they are not surprising
2. Less probable events yield more information
   - They happen less, so they are more surprising
3. The total information of independent events is the sum of the information of each individual event
   - Their chances of happening are unrelated, so knowing one outcome has no effect on how surprising the other outcome is

Information

The **(Shannon) information** of measuring random variable X with pmf P_X as outcome x is the quantity

$$I_X(x) = -\log_b(P_X(x)) \tag{1}$$

The log base b is an arbitrary choice which has the effect of fixing the units of information. Common choices:

- $b = 2$, “bits”
- $b = e$, “nats”
- $b = 10$, “dits”

Information is a **description of a distribution** like the pmf or cdf.

Sometimes the random variable $I(X) = I_X(X)$ is also called the information.

Entropy

The **entropy** of random variable X is the expected information of X

$$
\begin{aligned}
H(X) &= \mathbb{E}_X[I(X)] && (2) \\
&= \sum_i P_X(x_i) I_X(x_i) && (3) \\
&= -\sum_i P(x_i) \log(P_X(x_i)) && (4)
\end{aligned}
$$

Entropy measures the amount of randomness in X.

Entropy is a **summary statistic** like the mean or variance.

Let X be a Bernoulli random variable with success probability p

Let's compute the entropy of X as a function of the probability p

$$
\begin{aligned}
H(X) &= -\sum_i P(x_i)\log(P_X(x_i)) && (5)\\
&= -p\log(p) - (1-p)\log(1-p) && (6)
\end{aligned}
$$

Exercise: Compute p which maximize and minimize entropy.

Solution:

- Max entropy when $p = 1/2$
  - Most random, heads and tails equally likely
- Min entropy when $p = 0$ or $p = 1$
  - Least random, heads or tails is certain

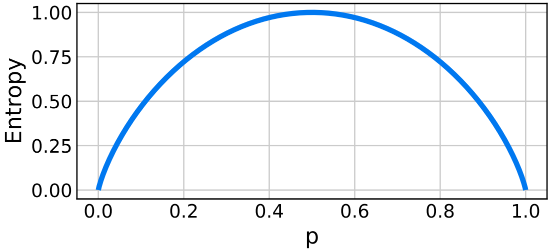

Figure 1: Entropy vs. parameter p for a Bernoulli random variable.
See `entropy_bernoulli.py`

Joint entropy

The **joint entropy** between two random variables X and Y with joint pmf P_{XY} is

$$H(X,Y) = -\sum_i \sum_j P_{XY}(x_i, y_i) \log(P_{XY}(x_i, y_i)) \tag{7}$$

Joint entropy measures the amount of randomness in X and Y.

Special case:
X and Y independent if and only if the joint entropy is additive

$$H(X,Y) = H(X) + H(Y) \tag{8}$$

Mutual information

The **mutual information** between two random variables X and Y is

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{9}$$

$$= \sum_i \sum_j P_{XY}(x_i, y_i) \log \left(\frac{P_{XY}(x_i, y_i)}{P_X(x_i) P_Y(y_i)} \right) \tag{10}$$

Mutual information measures the average reduction in uncertainty about X that results from learning the value of Y.

Special case: $I(X,X) = H(X)$, so entropy can be thought of as "self mutual information"

Cross-entropy

The **cross-entropy** from random variable Y to X is the expected information of Y with respect to X

$$H(X||Y) = \mathbb{E}_X[I(Y)] \tag{11}$$
$$= \sum_i P_X(x_i) I_Y(x_i) \tag{12}$$
$$= -\sum_i P_X(x_i) \log(P_Y(x_i)) \tag{13}$$

Cross-entropy measures the amount of randomness in Y, under the fictitious assumption that Y has the distribution of X for the purpose of computing expectation.

Special case: $H(X||X) = H(X)$, so entropy can be thought of as "self cross-entropy"

Relative entropy

Relative entropy / Kullback-Leibler divergence

The **relative entropy** or **Kullback–Leibler (KL) divergence** from random variable Y to X is

$$\mathcal{D}_{KL}(X||Y) = H(X||Y) - H(X) \tag{14}$$

$$= \sum_i P_X(x_i) \log \left(\frac{P_X(x_i)}{P_Y(x_i)} \right) \tag{15}$$

KL divergence measures the **difference between two distributions**.

KL divergence is **not a distance metric** because

1. It is not symmetric
2. The triangle inequality fails

See `kl_divergence.py`

Wasserstein metric

Wasserstein metric ("analytic" definition)

The **pth Wasserstein metric** between two pdfs f_X and f_Y is

$$W_p(f_X, f_Y) = \inf_{\pi \in \Pi(f_X, f_Y)} \left(\int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^p d\Pi(x, y) \right)^{1/p} \tag{16}$$

where $\Pi(f_X, f_Y)$ is the space of joint pdfs with marginals f_X and f_Y.

- There are ∞ different joint pdfs with marginals f_X and f_Y!
- The joint pdf π defines a **transport map** between f_X and f_Y.
  - π is a plan for moving the mass from f_X to f_Y (and vice versa)
  - Finding the infimal π is a special case of the general **optimal transport problem** c.f. [3]
  - In many cases, this ∞-dim infimization problem can be solved analytically or by reformulating as a finite-dim optimization program

Wasserstein metric

Wasserstein metric ("probabilistic" definition) [4]

The **pth Wasserstein metric** can be expressed as

$$W_p(f_X, f_Y) = \inf_{X \sim f_X,\ Y \sim f_Y} \left(\mathbb{E}_{XY}[\|X - Y\|^p]\right)^{1/p} \tag{17}$$

More facts:

- The two pdfs f_X and f_Y need not both be continuous or discrete
- $p = 1$ and $p = 2$ are the most common choices

Comparison with KL divergence:

- Like the KL divergence, the Wasserstein metric measures the **difference between two distributions**
- Unlike the KL divergence, the Wasserstein metric **is a valid distance metric**
  - Formal analysis using generic results for distance metrics is easier

Special case: pth Wasserstein metric of two Dirac deltas
$f_X(x) = \delta(x-a)$ and $f_Y(y) = \delta(y-b)$

$$W_p(f_X, f_Y) = \|a-b\| \tag{18}$$

Special case: 2nd Wasserstein metric of two Gaussians
$f_X = \mathcal{N}(\mu_X, \Sigma_X)$ and $f_Y = \mathcal{N}(\mu_Y, \Sigma_Y)$

$$W_2(f_X, f_Y) = \sqrt{\left\|\mu_X - \mu_Y\right\|^2 + \mathbf{Tr}\left[\Sigma_X + \Sigma_Y - 2\left(\Sigma_Y^{\frac{1}{2}}\Sigma_X\Sigma_Y^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]} \tag{19}$$

For the interested reader:

1. *"Statistical aspects of Wasserstein distances"* [4]
   - https://arxiv.org/abs/1806.05500
   - Contains a nice introduction on the Wasserstein metric.
2. *"Data-Driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations"* [5]
   - https://arxiv.org/abs/1505.05116
   - Quickly becoming a classic.
   - Details how to use the Wasserstein metric to solve optimization problems involving random problem data with unknown distribution while being robust to the worst-case distribution.

Bibliography I

[1] David JC MacKay and David JC Mac Kay.
Information theory, inference and learning algorithms.
Cambridge university press, 2003.
https://www.inference.org.uk/itila/.

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville.
Deep Learning.
MIT Press, 2016.
http://www.deeplearningbook.org.

[3] Cédric Villani.
Optimal transport: old and new, volume 338.
Springer, 2009.
https://cedricvillani.org/sites/dev/files/old_images/2012/08/preprint-1.pdf.

[4] Victor M Panaretos and Yoav Zemel.
Statistical aspects of wasserstein distances.
Annual review of statistics and its application, 6:405–431, 2019.
https://arxiv.org/pdf/1806.05500.pdf.

[5] Peyman Mohajerin Esfahani and Daniel Kuhn.
Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations.
Mathematical Programming, 171(1):115–166, 2018.
https://arxiv.org/pdf/1505.05116.pdf.