

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



- What is information theory?
- 2 Entropy
- 3 Wasserstein metric



Information theory

Information theory



Information theory concerns quantifying the amount of information present in a signal

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

Core ideas

- Unlikely events are more informative because they are unexpected
- Independent events add information because they are unrelated

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]

Entropy



Information

The information of random variable X with pmf P_X is the random variable I(X) with pmf

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

Entropy

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

$$H(X) = \mathbb{E}_X[I(X)] = \sum_i P_X(x_i)I_X(x_i) = -\sum_i P(x_i)\log(P_X(x_i))$$
(2)

Joint entropy



Joint entropy

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

$$H(X,Y) = -\sum_{i} \sum_{j} P_{XY}(x_{i}, y_{i}) \log(P_{XY}(x_{i}, y_{i}))$$
 (3)

Mutual information

The mutual information between two random variables X and Y is

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$
(4)

Special case: I(X,X) = I(X), so information is also called self-information

Cross-entropy



Cross-entropy

The ${\it cross-entropy}$ from random variable X to Y is the expected information of Y with respect to X

$$H(X||Y) = \mathbb{E}_X[I(Y)] = \sum_i P_X(x_i)I_Y(x_i) = -\sum_i P_X(x_i)\log(P_Y(x_i))$$
(5)

Relative entropy / Kullback-Leibler divergence

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

$$\mathcal{D}_{KL}(X||Y) = H(X||Y) - H(X) \tag{6}$$

The KL divergence measures the difference between two distributions

The KL divergence is **not** a **distance metric** because it is **not symmetric** and the triangle inequality fails



Wasserstein metric ("analytic" definition)

The pth Wasserstein metric between two pdfs f_X and f_Y is

$$W_p(f_X, f_Y) = \inf_{\pi \in \Pi(f_X, f_Y)} \left(\int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^p d\Pi(x, y) \right)^{1/p}$$
 (7)

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 .
 - \blacksquare 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

$$W_p(f_X, f_Y) = \inf_{X \sim f_X, Y \sim f_Y} \left(\mathbb{E}_{XY}[\|X - Y\|^p] \right)^{1/p} \tag{8}$$

More facts:

- lacktriangle The two pdfs f_X and f_Y need not both be continuous or discrete
- lacksquare 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 $f_Y(y) = \delta(y-b)$

$$W_p(f_X, f_Y) = ||a - b|| \tag{9}$$

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

$$W_{2}(f_{X}, f_{Y}) = \sqrt{\|\mu_{X} - \mu_{Y}\|^{2} + \mathbf{Tr} \left[\Sigma_{X} + \Sigma_{Y} - 2 \left(\Sigma_{Y}^{\frac{1}{2}} \Sigma_{X} \Sigma_{Y}^{\frac{1}{2}} \right)^{\frac{1}{2}} \right]}$$
(10)

Wasserstein metric



For the interested reader:

- "Statistical aspects of Wasserstein distances" [4]
 - https://arxiv.org/abs/1806.05500
 - Contains a nice introduction on the Wasserstein metric.
- "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



 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.

Bibliography II



[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.