WIKIPEDIA
The Free Encyclopedia

Earth mover's distance

2 languages

Article Talk Tools

From Wikipedia, the free encyclopedia

In computer science, the **earth mover's distance** (**EMD**)[1] is a measure of dissimilarity between two frequency distributions, densities, or measures, over a metric space *D*. Informally, if the distributions are interpreted as two different ways of piling up earth (dirt) over *D*, the EMD captures the minimum cost of building the smaller pile using dirt taken from the larger, where cost is defined as the amount of dirt moved multiplied by the distance over which it is moved.

Over probability distributions, the earth mover's distance is also known as the Wasserstein metric W_1, Kantorovich–Rubinstein metric, or Mallows's distance.[2] It is the solution of the optimal transport problem, which in turn is also known as the Monge-Kantorovich problem, or sometimes the Hitchcock–Koopmans transportation problem;[3] when the measures are uniform over a set of discrete elements, the same optimization problem is known as minimum weight bipartite matching.

Formal definitions [edit]

The EMD between probability distributions P and Q can be defined as an infimum over joint probabilities:

$$\text{EMD}(P,Q) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}\left[d(x,y)\right]$$

where $\Pi(P,Q)$ is the set of all joint distributions whose marginals are P and Q.

By Kantorovich-Rubinstein duality, this can also be expressed as:

$$\text{EMD}(P,Q) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim P}[f(x)] - \mathbb{E}_{y\sim Q}[f(y)]$$

where the supremum is taken over all 1-Lipschitz continuous functions, i.e. $\|\nabla f(x)\| \leq 1 \quad \forall x$.

EMD between signatures [edit]

In some applications, it is convenient to represent a distribution P as a *signature*, or a collection of *clusters*, where the i-th cluster represents a feature of mass w_i centered at p_i. In this formulation, consider signatures $P = \{(p_1, w_{p1}), (p_2, w_{p2}), \ldots, (p_m, w_{pm})\}$ and $Q = \{(q_1, w_{q1}), (q_2, w_{q2}), \ldots, (q_n, w_{qn})\}$. Let $D = [d_{i,j}]$ be the ground distance between clusters p_i and q_j. Then the EMD between P and Q is given by the optimal flow $F = [f_{i,j}]$, with $f_{i,j}$ the flow between p_i and q_j, that minimizes the overall cost.

$$\min_F \sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} d_{i,j}$$

subject to the constraints:

$$f_{i,j} \geq 0, 1 \leq i \leq m, 1 \leq j \leq n$$

$$\sum_{j=1}^{n} f_{i,j} \leq w_{pi}, 1 \leq i \leq m$$

$$\sum_{i=1}^{m} f_{i,j} \leq w_{qj}, 1 \leq j \leq n$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} = \min \left\{ \sum_{i=1}^{m} w_{pi}, \quad \sum_{j=1}^{n} w_{qj} \right\}$$

The optimal flow F is found by solving this linear optimization problem. The earth mover's distance is defined as the work normalized by the total flow:

$$\mathrm{EMD}(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} d_{i,j}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j}}$$

Variants and extensions [edit]

Unequal probability mass [edit]

Some applications may require the comparison of distributions with different total masses. One approach is to allow for *partial matching*,[1] where dirt from the more massive distribution is rearranged to make the less massive, and any leftover "dirt" is discarded at no cost. Formally, let w_P be the total weight of P, and w_Q be the total weight of Q. We have:

$$\mathrm{EMD}(P, Q) = \frac{1}{\min(w_P, w_Q)} \inf_{\gamma \in \Pi_{\geq}(P, Q)} \int d(x, y) \, \mathrm{d}\gamma(x, y)$$

where $\Pi_{\geq}(P, Q)$ is the set of all measures whose projections are $\geq P$ and $\geq Q$. Note that this generalization of EMD is not a true distance between distributions, as it does not satisfy the triangle inequality.

An alternative approach is to allow for mass to be created or destroyed, on a global or local level, as an alternative to transportation, but with a cost penalty. In that case one must specify a real parameter α, the ratio between the cost of creating or destroying one unit of "dirt", and the cost of transporting it by a unit distance. This is equivalent to minimizing the sum of the earth moving cost plus α times the L^1 distance between the rearranged pile and the second distribution. The resulting measure $\widehat{EMD}_{\alpha}$ is a true distance function.[4]

More than two distributions [edit]

The EMD can be extended naturally to the case where more than two distributions are compared. In this case, the "distance" between the many distributions is defined as the optimal value of a linear program. This generalized EMD may be computed exactly using a greedy algorithm, and the resulting functional has been shown to be Minkowski additive and convex monotone.[5]

Computing the EMD [edit]

The EMD can be computed by solving an instance of transportation problem, using any algorithm for minimum-cost flow problem, e.g. the network simplex algorithm.

The Hungarian algorithm can be used to get the solution if the domain D is the set $\{0,1\}$. If the domain is integral, it can be translated for the same algorithm by representing integral bins as multiple binary bins.

As a special case, if D is a one-dimensional array of "bins" of length n, the EMD can be efficiently computed by scanning the array and keeping track of how much dirt needs to be transported between consecutive bins. Here the bins are zero-indexed:

$$\begin{aligned}\mathrm{EMD}_0 &= 0\\ \mathrm{EMD}_{i+1} &= P_i + \mathrm{EMD}_i - Q_i\\ \text{Total Distance} &= \sum_{i=0}^{n} |\mathrm{EMD}_i|\end{aligned}$$

EMD-based similarity analysis [edit]

EMD-based similarity analysis (EMDSA) is an important and effective tool in many multimedia information retrieval[6] and pattern recognition[7] applications. However, the computational cost of EMD is super-cubic to the number of the "bins" given an arbitrary "D". Efficient and scalable EMD computation techniques for large scale data have been investigated using MapReduce,[8][9] as well as bulk synchronous parallel and resilient distributed dataset.[10]

Applications [edit]

An early application of the EMD in computer science was to compare two grayscale images that may differ due to dithering, blurring, or local deformations.[11] In this case, the region is the image's domain, and the total amount of light (or ink) is the "dirt" to be rearranged.

The EMD is widely used in content-based image retrieval to compute distances between the color histograms of two digital images.[*citation needed*] In this case, the region is the RGB color cube, and each image pixel is a parcel of "dirt". The same technique can be used for any other quantitative pixel attribute, such as luminance, gradient, apparent motion in a video frame, etc..

More generally, the EMD is used in pattern recognition to compare generic summaries or surrogates of data records called signatures.[1] A typical signature consists of list of pairs ((x_1,m_1), ... (x_n,m_n)), where each x_i is a certain "feature" (e.g., color in an image, letter in a text, etc.), and m_i is "mass" (how many times that feature occurs in the record). Alternatively, x_i may be the centroid of a data cluster, and m_i the number of entities in that cluster. To compare two such signatures with the EMD, one must define a distance between features, which is interpreted as the cost of turning a unit mass of one feature into a unit mass of the other. The EMD between two signatures is then the minimum cost of turning one of them into the other.

EMD analysis has been used for quantitating multivariate changes in biomarkers measured by flow cytometry, with potential applications to other technologies that report distributions of measurements.[12]

History [edit]

The concept was first introduced by Gaspard Monge in 1781,[13] in the context of transportation theory. The use of the EMD as a distance measure for monochromatic images was described in 1989 by S. Peleg, M. Werman and H. Rom.[11] The name "earth mover's distance" was proposed by J. Stolfi in 1994,[14] and was used in print in 1998 by Y. Rubner, C. Tomasi and L. G. Guibas.[1]

See also [edit]

- Monge–Ampère equation

References [edit]

1. ^ *a* *b* *c* *d* Rubner, Y.; Tomasi, C.; Guibas, L.J. (1998). "A metric for distributions with applications to image databases". *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*. Narosa Publishing House. pp. 59–66. doi:10.1109/iccv.1998.710701. ISBN 81-7319-221-9. S2CID 18648233.
2. ^ C. L. Mallows (1972). "A note on asymptotic joint normality". *Annals of Mathematical Statistics*. **43** (2): 508–515. doi:10.1214/aoms/1177692631.
3. ^ Singiresu S. Rao (2009). *Engineering Optimization: Theory and Practice* (4th ed.). John Wiley & Sons. p. 221. ISBN 978-0-470-18352-6.
4. ^ Pele, Ofir; Werman, Michael (2008). "A Linear Time Histogram Metric for Improved SIFT Matching". *Computer Vision – ECCV 2008*. Lecture Notes in Computer Science. Vol. 5304. Springer Berlin Heidelberg. pp. 495–508. doi:10.1007/978-3-540-88690-7_37. eISSN 1611-3349. ISBN 978-3-540-88689-1. ISSN 0302-9743.
5. ^ Kline, Jeffery (2019). "Properties of the d-dimensional earth mover's problem". *Discrete Applied Mathematics*. **265**: 128–141. doi:10.1016/j.dam.2019.02.042. S2CID 127962240.
6. ^ Mark A. Ruzon; Carlo Tomasi (2001). "Edge, Junction, and Corner Detection Using Color Distributions". *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
7. ^ Kristen Grauman; Trevor Darrel (2004). "Fast contour matching using approximate earth mover's distance". *Proceedings of CVPR 2004*.
8. ^ Jin Huang; Rui Zhang; Rajkumar Buyya; Jian Chen (2014). "MELODY-Join: Efficient Earth Mover's Distance Similarity Joins Using MapReduce". *Proceedings of IEEE International Conference on Data Engineering*.
9. ^ Jia Xu; Bin Lei; Yu Gu; Winslett, M.; Ge Yu; Zhenjie Zhang (2015). "Efficient Similarity Join Based on Earth Mover's Distance Using MapReduce". *IEEE Transactions on Knowledge and Data Engineering*.

10. ^ Jin Huang; Rui Zhang; Rajkumar Buyya; Jian Chen, M.; Yongwei Wu (2015). "Heads-Join: Efficient Earth Mover's Distance Join on Hadoop". *IEEE Transactions on Parallel and Distributed Systems*.
11. ^ *a* *b* S. Peleg; M. Werman; H. Rom (1989). "A unified approach to the change of resolution: Space and gray-level". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **11** (7): 739–742. doi:10.1109/34.192468. S2CID 18415340.
12. ^ Orlova, DY; Zimmerman, N; Meehan, C; Meehan, S; Waters, J; Ghosn, EEB (23 March 2016). "Earth Mover's Distance (EMD): A True Metric for Comparing Biomarker Expression Levels in Cell Populations". *PLOS One*. **11** (3): e0151859. Bibcode:2016PLoSO..1151859O. doi:10.1371/journal.pone.0151859. PMC 4805242. PMID 27008164.
13. ^ "Mémoire sur la théorie des déblais et des remblais". *Histoire de l'Académie Royale des Science, Année 1781, avec les Mémoires de Mathématique et de Physique*. 1781.
14. ^ J. Stolfi, personal communication to L. J. Guibas, 1994, as cited by Rubner, Yossi; Tomasi, Carlo; Guibas, Leonidas J. (2000). "The earth mover's distance as a metric for image retrieval" (PDF). *International Journal of Computer Vision*. **40** (2): 99–121. doi:10.1023/A:1026543900054. S2CID 14106275.

External links [edit]

- C code for the Earth Mover's Distance (archived here)
- Python implementation with references
- Python2 wrapper for the C implementation of the Earth Mover's Distance
- C++ and Matlab and Java wrappers code for the Earth Mover's Distance, especially efficient for thresholded ground distances
- Java implementation of a generic generator for evaluating large-scale Earth Mover's Distance based similarity analysis
- Demonstration of Minkowski additivity, convex monotonicity, and other properties of the Earth Movers distance

Category: Statistical distance