Notes on statistical distances

Katarzyna Kedzierska

November 09, 2017

Table of Content

Earth Mover's Distance (EMD)

Earth Mover's Distance is the measure of dissimilarity between two distributions. Describes minimal amount of work needed to change one distribution into other. In order to compare two distributions one must present the distribution as a signature - set of clusters represented by the mean or mode and weight equal to the fraction of the distribution given cluster represents.

Given two signatures P and Q:
$P = \{(p_1, w_{p1}), (p_2, w_{p2}), ..., (p_m, w_{pm})\}$
$Q = \{(q_1, w_{q1}), (q_2, w_{q2}), ..., (q_m, w_{qm})\}$
p, q - cluster representatives (mode or mean)
w - weight of the cluster
m, n - number of clusters
$D = [d_{ij}]$
$F = [f_{ij}]$

The goal is to find F such that:

$$WORK(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}$$

is minimized given the following constrains:

$f_{ij} \leq 0$ for $1 \leq i \leq m$; $1 \leq j \leq n$

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

Bhattacharyya distance Hellinger distance Kolmogorov-Smirnov test