Data and Similarity

An (alternative) introduction to ML basics

Jeremiah Deng

Information Science
University of Otago

11 July, 2017

Outline

Distance and Similarity

- Many data mining techniques are based on similarity measures between data points
  - Classification: nearest-neighbor, linear discriminant analysis
  - Clustering: k-means, density
  - Visualization: multi-dimensional scaling
- Proximity is a general term to indicate (dis)similarity
- Distance is also used to indicate dissimilarity.
- In mathematics, a distance means a metric.

Metric

- A metric or distance function is a function that defines a distance between each pair of elements of a set (X):

$$d : X \times X \to [0, +\infty)$$

- Distance d satisfies the following:
  1. $d(x, y) \geq 0$
  2. $d(x, y) = 0 \Leftrightarrow x = y$
  3. $d(x, y) = d(y, x)$
  4. $d(x, z) \leq d(x, y) + d(y, z)$

Euclidean distance is a metric

- Two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$
- Euclidean distance

$$L_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- L_2 is a metric.
- Originates from $\mathbb{R}^2$ and $\mathbb{R}^3$; but *scales* to $\mathbb{R}^n$.

Euclidean Distance in Clustering

- How to partition a data set X into k *clusters*?
- The goal is to optimize a score function that links to the compactness of each cluster.
- The most commonly used is the square error criterion:
  $\mathcal{F} = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mathbf{m}_i\|^2$
- Finding the best $\mathbf{m}_i$: $\dfrac{\delta \mathcal{F}}{\delta \mathbf{m}_i} = 0$.
- Given all $\mathbf{x}$ within a cluster, the centroid gives the minimum: $\mathbf{m}_i = E_{\mathbf{x} \in C_i}(\mathbf{x})$.
- This gives the k-means algorithm.

Minkowski Metrics

- Euclidean distance: $L_2 = (\sum_{i=1}^{n} |x_i - y_i|^2)^{1/2}$

$\rightarrow$ Minkowski distance

$$L_p = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p}, p \geq 0$$

- L_1: Metropolitan (city-block): $L_1 = \sum_{i=1}^{n} |x_i - y_i|$
- L_∞: $\max_i |x_i - y_i|$
- $L_{-\infty}$: $\min_i |x_i - y_i|$
- What about L_0?

Is Euclidean Always Good?

- For a high-dimensional space (e.g. $n \geq 10$), data points are more likely lying on hypercubes rather than within hyper spheres.
- So Euclidean distance may easily fail to represent similarity between data points.
  - See P Domingos, "A few useful things to know about machine learning", CACM 55:78-87, 2012
  - C Aggarwal et al., "On the surprising behavior of distance metrics in high dimensional space", LNCS 1973:420-434, 2001.
- This does not *always* happens.
  - See A Zimek et al. (2012), "A survey on unsupervised outlier detection in high-dimensional numerical data", Statistical Analy Data Mining, 5: 363–387

More generalizations

- Mahalanobis distance: ellipsoids.
  - Given data vectors $\{\mathbf{x}\}$, calculate the mean $\boldsymbol{\mu}$ and the covariance matrix Σ. The distance between $\mathbf{x}$ and $\boldsymbol{\mu}$ is $D_M(\mathbf{x}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$
  - Between two data vectors of the same distribution: $d_M(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\boldsymbol{y})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{y})}$
- Distance metric learning (Xing et al., NIPS'02)
  - If $\mathbf{x},\mathbf{y} \in S$, can we learn an optimal metric?

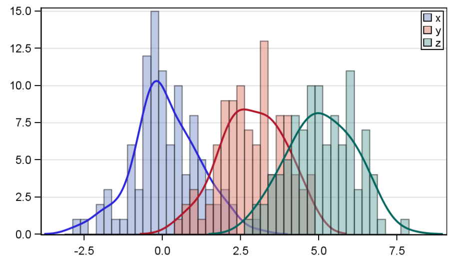
$$d_A(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T\mathbf{A}(\mathbf{x}-\mathbf{y})}$$

  Matrix $\mathbf{A}$ is positive semi-definite.
  - The best A for clustering can be solved for

$$\begin{array}{ll} \min\limits_A & \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in S} \|\mathbf{x}_i-\mathbf{x}_j\|_A^2 \\ s.t. & \sum_{(\mathbf{x}_i,\mathbf{x}_j)\notin S} \|\mathbf{x}_i-\mathbf{x}_j\|_A \geq 1 \end{array}$$

How to compare histograms?

- Sometimes, data vectors are actually histograms – discrete probability models.
- The difference on bin values matters; the distance between bins also matters.
- Euclidean distance is not a good representation any more!



Divergence Between Two Probability Distributions

- Kullback-Leibler divergence between $p(x)$ and $q(x)$:

$$\mathrm{KLD}(p,q) = \sum_i p_i \log\frac{p_i}{q_i}$$

- Assume that the N dimensions of the data are independent and Gaussian distributed, a simplified Kullback-Leiberler divergence can be worked out in close form (Mathiassen 2002) for two models p and q:

$$\mathrm{KL}(q;p) = \frac{1}{2}\sum_{j=1}^{N}\left(\log\left(\frac{\sigma_j^{(p)}}{\sigma_j^{(q)}}\right)^2 + \left(\frac{\mu_j^{(q)}-\mu_j^{(p)}}{\sigma_j^{(p)}}\right)^2 + \left(\frac{\sigma_j^{(q)}}{\sigma_j^{(p)}}\right)^2 - 1\right)$$

Variation: Earth moving distance (EMD)

EMD (Rrubner, 1998) is defined over weighted point sets.
Suppose each point set is configured by a normalized weight set.
Denote a point set as $A = \{a_1, a_2, ..., a_m\}$, with $a_i = \{(x_i, w_i)\}$, $x_i \in R^k$, and $w_i \in R^+ \cup \{0\}$.
EMD: the minimum amount of work needed to transform one configuration to another by moving weight under constraints.
Denote the set of all feasible flows as $F = \{f_{ij}\}$, where i is a point label for set A, and j for B. These flows are subject to certain constraints.
EMD between the two point sets can then be define as

$$\mathrm{EMD}(A,B) = \min_{f\in F}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}d_{ij}. \quad (1)$$

Content-based Image Retrieval

- Traditionally, information retrieval is concept-based – problematic for images.
- CBIR aims at retrieve images using low-level feature matching (e.g. search by example, search by sketch etc.)
- Skips costly image annotation; circumvents word matching issues (keyword subjective; thesauri needed)
- Hot topic during 1990s/2000s
- ☹ Haunted by the "semantic gap"
- ☺ Leveraged research on image feature extraction, database, and similarity-based pattern recognition
- Key question: how to compare the histograms (colour / shape / texture)?

More Variations

- Rubner et al., "Empirical evaluation of dissimilarity measures for color and texture", CVIU 2001
- *Minkowski distance* L_p
- Kolmogorov–Smirnov distance distance between two cumulative probability functions $F(X)$ and $F(Y)$:

$$\mathrm{KS}(X, Y) = \max_i |F(i; X) - F(i; Y)|$$

- *Kullback-Leibler divergence (KLD)*
- Jensen-Shannon divergence
- Findings: best "distance" is data dependent, but L_2 and L_∞ consistently inferior (!)

Set-to-set distances?

- Hausdorff distance between two non-empty sets X and Y

$$d_H(X, Y) = \max\left(\sup_{x\in X} \inf_{y\in Y} d(x, y), \sup_{y\in Y} \inf_{x\in X} d(x, y)\right)$$

  - HD is a metric.
  - Used in computer vision in comparing shapes.
- Jaccard index
  - Between two sets A and B: $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, $d_J(A, B) = 1 - J(A, B)$
  - Generalized $J(X, Y) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)}$
- Other ideas?

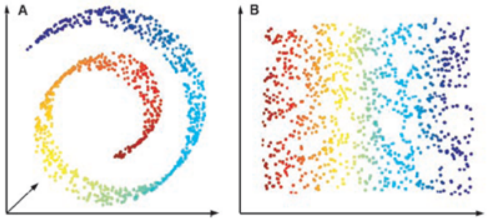
Another case against Euclidean distance

According to Winston Churchill

It is a mistake to look too far ahead. Only one link of the chain of destiny can be handled at a time.

- Sometimes distance between two data points should not be measured directly but by how many hops they are separate by neighbours.
- A number of algorithms exploit the neighbourhoods and turn a dataset into a graph.
- E.g. Isomap by Tenenbaum et al., *Science*, v290(5500), 2000, pp.2319-2323.

Isomap



"Swiss roll" expanded by Isomap (Tenenbaum et al., 2000)

Isomap – a nonlinear MDS

- Connect each point to its k nearest neighbors to form a graph.
- Approximate pairwise geodesic distances using Dijkstra's algorithm on this graph.
- Apply Metric MDS to recover a low dimensional isometric embedding.
- *t.b.c.*: manifold learning

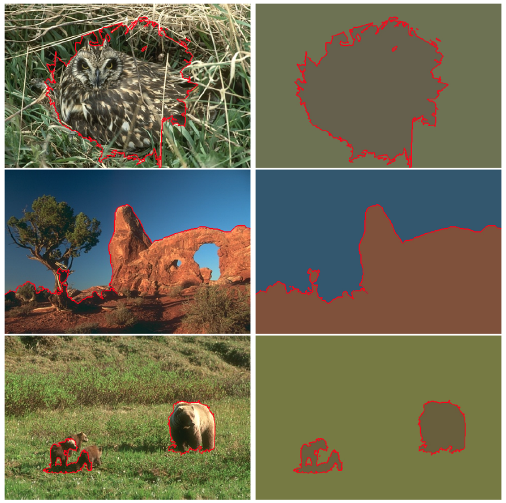
A Distance metric for covariance matrices

- Image segmentation: pixels ($\rightarrow$ superpixels) $\rightarrow$ regions
- In addition to pixel color, Gu et al. (2014) proposed to incorporate covariance matrices for image segments.
- Förstner & Moonen metric on two covariance matrices Σ_A, Σ_B: $d(\Sigma_A, \Sigma_B) = \sqrt{\sum_{r=1}^{n} \ln^2 \lambda_r}$, where Σ_A, Σ_B are of dimension $n \times n$, $\lambda_r (r = 1, 2, \cdots, n)$ are the eigenvalues from the generalized eigenvalue problem $|\lambda\Sigma_A - \Sigma_B| = 0$.
- Two different similarity matrices, W_c and W_Σ, representing color and color covariance respectively
- Spectral clustering based on similarity measures combining W_c and W_Σ

Image segmentation

Table 2. Performance comparison over the BSD database with K adjusted manually

Algorithms	PRI	VoI	GCE	BDE
SAS [7]	0.8319	1.6849	**0.1779**	11.2900
ℓ_0-sparse-coding [9]	0.8355	1.9935	0.2297	**11.1955**
Ours (using W_{HP})	**0.8495**	**1.6260**	0.1785	12.3034
Ours (using W_{DP})	0.8345	2.1169	0.2341	12.0008
Ours (using W_{AD})	0.8397	2.0359	0.2308	11.8868



(Gu, Deng, Purvis 2014)

"Top 10% Paper" ICIP'14 Paris

Recap

- Choices on distance metrics / similarity measures can make a difference.
- "Distance" can be measured between data points (vectors), histograms, covariance matrices, and sets / set profiles.
- New metrics / similarity measures in high-dimensional data spaces, and their combinations, remain interesting research topics.
- Notable directions:
  - Tensor, manifold learning
  - Kullback-Leibler divergence
  - Distance metric learning