Dissimilarity Measures

Prof. Asim Tewari
IIT Bombay

Dissimilarity Measures

- A function *d* is called *dissimilarity* or *distance measure* if for all *x*, *y* ϵ R^p

$$d(x, y) = d(y, x)$$

$$d(x, y) = 0 \quad \Leftrightarrow \quad x = y$$

$$d(x, z) \leq d(x, y) + d(y, z)$$

- From these axioms it follows:

$$d(x, y) \geq 0$$

Dissimilarity Measures

- A class of dissimilarity measures is defined using a norm || . ||of x-y, so

$$d(x, y) = \|x - y\|$$

- A function || . || :R^P -> R^+ is a norm if and only if

$$\|x\| = 0 \Leftrightarrow x = (0, \ldots, 0)$$

$$\|a \cdot x\| \;=\; |a| \cdot \|x\| \quad \forall a \in \mathbb{R}, x \in \mathbb{R}^p$$

$$\|x + y\| \;\leq\; \|x\| + \|y\| \quad \forall x, y \in \mathbb{R}^p$$

Dissimilarity Measures

- The so-called hyperbolic norm

$$\|x\|_h = \prod_{i=1}^{p} x^{(i)}$$

is not a norm according to the previous definition, since it violates

$$\|x\| = 0 \Leftrightarrow x = (0, \ldots, 0)$$

$$\|a \cdot x\| \ = \ |a| \cdot \|x\| \quad \forall a \in \mathbb{R}, x \in \mathbb{R}^p$$

Dissimilarity Measures

A frequently used classes of norms are matrix norms.

The matrix norm is defined as

$$\|x\|_A = \sqrt{xAx^T}$$

with a matrix $A \in \mathbb{R}^{n \times n}$

Dissimilarity Measures

- Euclidean norm

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

- Frobenius or Hilbert-Schmidt norm

$$A = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

Dissimilarity Measures

- Diagonal norm

$$A = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_p \end{pmatrix}$$

- Mahalanobis norm

$$A = \mathrm{cov}^{-1}X = \left(\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^T (x_k - \bar{x}) \right)^{-1}$$

Dissimilarity Measures

- Lebesgue or Minkowski norm

$$\|x\|_\alpha = \sqrt[\alpha]{\sum_{j=1}^{p} \left|x^{(j)}\right|^\alpha}$$

This is the generalized mean (except for a constant factor)

Dissimilarity Measures

Lebesgue norm can lead to

- Infimum norm

  $(\alpha \to -\infty)$

  $$\|x\|_{-\infty} = \min_{j=1,\ldots,p} x^{(j)}$$

- Manhattan or city block distance

  $(\alpha = 1)$

  $$\|x\|_1 = \sum_{j=1}^{p} \left| x^{(j)} \right|$$

- Euclidean norm

  $(\alpha = 2)$

  $$\|x\|_2 = \sqrt{\sum_{j=1}^{p} \left(x^{(j)}\right)^2}$$

- Supremum norm $(\alpha \to \infty)$

  $$\|x\|_{\infty} = \max_{j=1,\ldots,p} x^{(j)}$$

Dissimilarity Measures

- Hamming distance (not a norm)

$$d_H(x, y) = \sum_{i=1}^{p} \rho(x^{(i)}, y^{(i)})$$

with the discrete metric

$$\rho(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{otherwise} \end{cases}$$

For binary features, the Hamming distance is equal to the Manhattan or city block distance.

Similarity Measures

A function s is called similarity or proximity measure if for all $x, y \in \mathbb{R}^p$

$$s(x, y) = s(y, x)$$

$$s(x, y) \leq s(x, x)$$

$$s(x, y) \geq 0$$

The function s is called normalized similarity measure if additionally

$$s(x, x) = 1$$

Similarity Measures

- Any dissimilarity measure d can be used to define a corresponding similarity measure and vice versa, for example using a monotonically decreasing positive function *f* with *f(0)=1* such as

$$s(x, y) = \frac{1}{1 + d(x, y)}$$

Similarity Measures

- Cosine

$$s(x, y) = \frac{\sum_{i=1}^{p} x^{(i)} y^{(i)}}{\sqrt{\sum_{i=1}^{p} \left(x^{(i)}\right)^2 \sum_{i=1}^{p} \left(y^{(i)}\right)^2}}$$

- This is invariant against (positive) scaling of the feature vectors and therefore considers the relative distribution of the features,

$$s(c \cdot x, y) = s(x, y)$$

$$s(x, c \cdot y) = s(x, y)$$

Similarity Measures

- Overlap

$$s(x, y) = \frac{\sum_{i=1}^{p} x^{(i)} y^{(i)}}{\min\left(\sum_{i=1}^{p} \left(x^{(i)}\right)^2, \sum_{i=1}^{p} \left(y^{(i)}\right)^2\right)}$$

Similarity Measures

- Dice

$$s(x, y) = \frac{2 \sum_{i=1}^{p} x^{(i)} y^{(i)}}{\sum_{i=1}^{p} \left(x^{(i)}\right)^2 + \sum_{i=1}^{p} \left(y^{(i)}\right)^2}$$

Similarity Measures

- Jaccard (or sometimes called Tanimoto)

$$s(x, y) = \frac{\sum_{i=1}^{p} x^{(i)} y^{(i)}}{\sum_{i=1}^{p} \left(x^{(i)}\right)^2 + \sum_{i=1}^{p} \left(y^{(i)}\right)^2 - \sum_{i=1}^{p} x^{(i)} y^{(i)}}$$