Spazi vettoriali e misure di similaritá

R. Basili

Corso di *Web Mining* e *Retrieval*
a.a. 2009-10

March 25, 2010

Outline

Outline

- Spazi vettoriali a valori reali
- Operazioni tra vettori
- Indipendenza Lineare
- Basi
- Prodotto Interno
- Norma di un vettore e Proprietá
- Vettori unitari
- Ortogonalitá
- Similaritá
- Norme e similaritá

Real-valued Vector Space

Vector Space definition:

A *vector space* is a set V of objects called *vectors* $\underline{x} = \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} = |\underline{x}\rangle$

where we can simply refer to a vector by $\underline{x}$, or using the specific realization called *column vector*, (*Dirac* notation $|\underline{x}\rangle$)

Real-valued Vector Space

Vector Space definition:

A vector space need to satisfy the following axioms:

Real-valued Vector Space

Vector Space definition:

A vector space need to satisfy the following axioms:

Sum

To every pair, $\underline{x}$ and $\underline{y}$, of vectors in V there corresponds a vector $\underline{x}+\underline{y}$, called the sum of $\underline{x}$ and $\underline{y}$, in such a way that:

1. sum is commutative, $\underline{x}+\underline{y}=\underline{y}+\underline{x}$
2. sum is associative, $\underline{x}+(\underline{y}+\underline{z})=(\underline{x}+\underline{y})+\underline{z}$
3. there exist in V a unique vector Φ (called the origin) such that $\underline{x}+\Phi=\underline{x}\ \forall \underline{x}\in V$
4. $\forall \underline{x}\in V$ there corresponds a unique vector $-\underline{x}$ such that $\underline{x}+(-\underline{x})=\Phi$

Real-valued Vector Space

Vector Space definition:

A vector space need to satisfy the following axioms:

Sum

To every pair, $\underline{x}$ and $\underline{y}$, of vectors in V there corresponds a vector $\underline{x}+\underline{y}$, called the sum of $\underline{x}$ and $\underline{y}$, in such a way that:

1. sum is commutative, $\underline{x}+\underline{y}=\underline{y}+\underline{x}$
2. sum is associative, $\underline{x}+(\underline{y}+\underline{z})=(\underline{x}+\underline{y})+\underline{z}$
3. there exist in V a unique vector Φ (called the origin) such that $\underline{x}+\Phi=\underline{x}\ \forall\underline{x}\in V$
4. $\forall\underline{x}\in V$ there corresponds a unique vector $-\underline{x}$ such that $\underline{x}+(-\underline{x})=\Phi$

Scalar Multiplication

To every pair α and $\underline{x}$, where α is a scalar and $\underline{x}\in V$, there corresponds a vector $\alpha\underline{x}$, called the product of α and $\underline{x}$, in such a way that:

1. associativity $\alpha(\beta\underline{x})=(\alpha\beta)\underline{x}$
2. $1\underline{x}=\underline{x} \qquad \forall\underline{x}\in V$
3. mult. by *scalar* is distributive wrt. vector addition $\alpha(\underline{x}+\underline{y})=\alpha\underline{x}+\alpha\underline{y}$
4. mult. by *vector* is distributive wrt. scalar addition $(\alpha+\beta)\underline{x}=\alpha\underline{x}+\beta\underline{x}$

Vector Operations

Sum of two vector $\underline{x}$ and $\underline{y}$

$$\underline{x}+\underline{y}=|\underline{x}\rangle+|\underline{y}\rangle=\begin{pmatrix} x_1+y_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n+y_n \end{pmatrix}$$

Vector Operations

Sum of two vector $\underline{x}$ and $\underline{y}$

$$\underline{x} + \underline{y} = |\underline{x}\rangle + |\underline{y}\rangle = \begin{pmatrix} x_1 + y_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n + y_n \end{pmatrix}$$

Linear combination

$$\underline{y} = c_1 \underline{x}_1 + \cdots + c_n \underline{x}_n$$

or

$$|\underline{y}\rangle = c_1 |\underline{x}_1\rangle + \cdots + c_n |\underline{x}_n\rangle$$

Vector Operations

Sum of two vector $\underline{x}$ and $\underline{y}$

$$\underline{x}+\underline{y}=|\underline{x}\rangle+|\underline{y}\rangle=\begin{pmatrix} x_1+y_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n+y_n \end{pmatrix}$$

Multiplication by scalar α

$$\alpha\underline{x}=\alpha|\underline{x}\rangle=\begin{pmatrix} \alpha x_1 \\ \cdot \\ \cdot \\ \cdot \\ \alpha x_n \end{pmatrix}$$

Linear combination

$$\underline{y}=c_1\underline{x}_1+\cdots+c_n\underline{x}_n$$

or

$$|\underline{y}\rangle=c_1|\underline{x}_1\rangle+\cdots+c_n|\underline{x}_n\rangle$$

Linear dependence

Conditions for linear dependence

A set o vectors $\{\underline{x}_1, \ldots, \underline{x}_n\}$ are *linearly dependent* if there a set constant scalars $c_1, \ldots, c_n$ exists, not all 0, such that:

$$c_1\underline{x}_1 + \cdots + c_n\underline{x}_n = \underline{0}$$

Linear dependence

Conditions for linear dependence

A set o vectors $\{\underline{x}_1, \ldots, \underline{x}_n\}$ are *linearly dependent* if there a set constant scalars $c_1, \ldots, c_n$ exists, not all 0, such that:

$$c_1\underline{x}_1 + \cdots + c_n\underline{x}_n = \underline{0}$$

Conditions for linear independence

A set o vectors $\{\underline{x}_1, \ldots, \underline{x}_n\}$ are *linearly independent* if and only if the *linear condition* $c_1\underline{x}_1 + \cdots + c_n\underline{x}_n = \underline{0}$ is satisfied only when $c_1 = c_2 = \cdots = c_n = 0$

Basis

Definition:

A *basis* for a space is a set of n linearly independent vectors in a n-dimensional vector space V_n.

Basis

Definition:

A *basis* for a space is a set of n linearly independent vectors in a n-dimensional vector space V_n.

This means that every arbitrary vector $\underline{x} \in V$ can be expressed as linear combination of the *basis* vectors,

$$\underline{x} = c_1\underline{x}_1 + \cdots + c_n\underline{x}_n$$

where the c_i are called the co-ordinates of $\underline{x}$ wrt. the basis set $\{\underline{x}_1, \ldots, \underline{x}_n\}$

Inner Product

Definition:

Is a real-valued function on the cross product $V_n \times V_n$ associating with each pair of vectors $(\underline{x}, \underline{y})$ a unique real number.
The function $(.,.)$ has the following properties:

1. $(\underline{x}, \underline{y}) = (\underline{y}, \underline{x})$
2. $(\underline{x}, \lambda \underline{y}) = \lambda (\underline{x}, \underline{y})$
3. $(\underline{x}_1 + \underline{x}_2, \underline{y}) = (\underline{x}_1, \underline{y}) + (\underline{x}_2, \underline{y})$
4. $(\underline{x}, \underline{x}) \geq 0$ and $(\underline{x}, \underline{x}) = 0$ **iff** $\underline{x} = \underline{0}$

Inner Product

Definition:

Is a real-valued function on the cross product $V_n \times V_n$ associating with each pair of vectors $(\underline{x}, \underline{y})$ a unique real number.
The function $(.,.)$ has the following properties:

1. $(\underline{x}, \underline{y}) = (\underline{y}, \underline{x})$
2. $(\underline{x}, \lambda \underline{y}) = \lambda (\underline{x}, \underline{y})$
3. $(\underline{x}_1 + \underline{x}_2, \underline{y}) = (\underline{x}_1, \underline{y}) + (\underline{x}_2, \underline{y})$
4. $(\underline{x}, \underline{x}) \geq 0$ and $(\underline{x}, \underline{x}) = 0$ **iff** $\underline{x} = \underline{0}$

Standard Inner Product

$$(\underline{x}, \underline{y}) = \sum_{i=1}^{n} x_i y_i$$

Inner Product

Definition:

Is a real-valued function on the cross product $V_n \times V_n$ associating with each pair of vectors $(\underline{x}, \underline{y})$ a unique real number.
The function $(.,.)$ has the following properties:

1. $(\underline{x}, \underline{y}) = (\underline{y}, \underline{x})$
2. $(\underline{x}, \lambda \underline{y}) = \lambda (\underline{x}, \underline{y})$
3. $(\underline{x}_1 + \underline{x}_2, \underline{y}) = (\underline{x}_1, \underline{y}) + (\underline{x}_2, \underline{y})$
4. $(\underline{x}, \underline{x}) \geq 0$ and $(\underline{x}, \underline{x}) = 0$ **iff** $\underline{x} = \underline{0}$

Standard Inner Product

$$(\underline{x}, \underline{y}) = \sum_{i=1}^{n} x_i y_i$$

Other notations

- $\underline{x}^T \underline{y}$ where $\underline{x}^T$ is the transpose of $\underline{x}$
- $\langle \underline{x} | \underline{y} \rangle$ or sometimes $\langle \underline{x} || \underline{y} \rangle$ in Dirac notation

Norm

Geometric interpretation

Geometrically the *norm* represent the length of the vector

Norm

Geometric interpretation

Geometrically the *norm* represent the length of the vector

Definition

The *norm* id a function $||.||$ from V_n to $\mathbb{R}$

Norm

Geometric interpretation

Geometrically the *norm* represent the length of the vector

Definition

The *norm* id a function $||.||$ from V_n to $\mathbb{R}$

Euclidean Norm:

$$||\underline{x}|| = \sqrt{(\underline{x}, \underline{x})} = \sqrt{\sum_{i=1}^{n} x_i^2} = \left(x_1^2 + \cdots + x_n^2\right)^{1/2}$$

Norm

Geometric interpretation

Geometrically the *norm* represent the length of the vector

Definition

The *norm* id a function $||.||$ from V_n to $\mathbb{R}$

Euclidean Norm:

$$||\underline{x}|| = \sqrt{(\underline{x},\underline{x})} = \sqrt{\sum_{i=1}^{n} x_i^2} = \left(x_1^2 + \cdots + x_n^2\right)^{1/2}$$

Properties

1. $||\underline{x}|| \geq 0$ and $||\underline{x}|| = 0$ if and only if $\underline{x} = 0$
2. $||\alpha \underline{x}|| = |\alpha| \, ||\underline{x}||$ for all α and $\underline{x}$
3. $\forall \underline{x}, \underline{y}, ||(\underline{x}, \underline{y})|| \leq ||\underline{x}|| \, ||\underline{y}||$ (Cauchy-Schwartz)

Norm

Geometric interpretation

Geometrically the *norm* represent the length of the vector

Definition

The *norm* id a function $||.||$ from V_n to $\mathbb{R}$

Euclidean Norm:

$$||\underline{x}|| = \sqrt{(\underline{x},\underline{x})} = \sqrt{\sum_{i=1}^{n} x_i^2} = \left(x_1^2 + \cdots + x_n^2\right)^{1/2}$$

Properties

1. $||\underline{x}|| \geq 0$ and $||\underline{x}|| = 0$ if and only if $\underline{x} = 0$
2. $||\alpha \underline{x}|| = |\alpha| \, ||\underline{x}||$ for all α and $\underline{x}$
3. $\forall \underline{x}, \underline{y}, ||(\underline{x},\underline{y})|| \leq ||\underline{x}|| \, ||\underline{y}||$ (Cauchy-Schwartz)

Norm

Geometric interpretation

Geometrically the *norm* represent the length of the vector

Definition

The *norm* id a function $||.||$ from V_n to $\mathbb{R}$

Euclidean Norm:

$$||\underline{x}|| = \sqrt{(\underline{x},\underline{x})} = \sqrt{\sum_{i=1}^{n} x_i^2} = \left(x_1^2 + \cdots + x_n^2\right)^{1/2}$$

Properties

1. $||\underline{x}|| \geq 0$ and $||\underline{x}|| = 0$ if and only if $\underline{x} = 0$
2. $||\alpha \underline{x}|| = |\alpha|\,||\underline{x}||$ for all α and $\underline{x}$
3. $\forall \underline{x}, \underline{y}, ||(\underline{x},\underline{y})|| \leq ||\underline{x}||\,||\underline{y}||$ (Cauchy-Schwartz)

A vector $\underline{x} \in V_n$ is a *unit vector*, or *normalsized*, when $||\underline{x}|| = 1$

From Norm to distance

In V_n we can define the distance between two vectors $\underline{x}$ and $\underline{y}$ as:

$$d(\underline{x},\underline{y}) = ||\underline{x}-\underline{y}|| = \sqrt{(\underline{x}-\underline{y},\underline{x}-\underline{y})} = \left((x_1-y_1)^2+\cdots+(x_n-y_n)^2\right)^{1/2}$$

These measure, noted sometimes as $||\underline{x}-\underline{y}||_2^2$, is also named *Euclidean distance*.

From Norm to distance

In V_n we can define the distance between two vectors $\underline{x}$ and $\underline{y}$ as:

$$d(\underline{x},\underline{y}) = ||\underline{x}-\underline{y}|| = \sqrt{(\underline{x}-\underline{y},\underline{x}-\underline{y})} = \left((x_1-y_1)^2+\cdots+(x_n-y_n)^2\right)^{1/2}$$

These measure, noted sometimes as $||\underline{x}-\underline{y}||_2^2$, is also named *Euclidean distance*.

Properties:

- $d(\underline{x},\underline{y}) \geq 0$ and $d(\underline{x},\underline{y}) = 0$ if and only if $\underline{x}=\underline{y}$
- $d(\underline{x},\underline{y}) = d(\underline{y},\underline{x})$ symmetry
- $d(\underline{x},\underline{y}) =\leq d(\underline{x},\underline{z}) + d(\underline{z},\underline{y})$ triangle inequality

From Norm to distance

An immediate consequence of Cauchy-Schwartz property is that:

$$-1 \leq \frac{(\underline{x}, \underline{y})}{||\underline{x}||\,||\underline{y}||} \leq 1$$

and therefore we can express it as:

$$(\underline{x}, \underline{y}) = ||\underline{x}||\,||\underline{y}|| \cos\varphi \qquad 0 \leq \varphi \leq \pi$$

where φ is the angle between the two vectors $\underline{x}$ and $\underline{y}$

From Norm to distance

An immediate consequence of Cauchy-Schwartz property is that:

$$-1 \leq \frac{(\underline{x},\underline{y})}{||\underline{x}||\,||\underline{y}||} \leq 1$$

and therefore we can express it as:

$$(\underline{x},\underline{y}) = ||\underline{x}||\,||\underline{y}|| \cos\varphi \qquad 0 \leq \varphi \leq \pi$$

where φ is the angle between the two vectors $\underline{x}$ and $\underline{y}$

Cosine distance

$$\cos\varphi = \frac{(\underline{x},\underline{y})}{||\underline{x}||\,||\underline{y}||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$

From Norm to distance

An immediate consequence of Cauchy-Schwartz property is that:

$$-1 \leq \frac{(\underline{x},\underline{y})}{||\underline{x}||\,||\underline{y}||} \leq 1$$

and therefore we can express it as:

$$(\underline{x},\underline{y}) = ||\underline{x}||\,||\underline{y}||\cos\varphi \qquad 0 \leq \varphi \leq \pi$$

where φ is the angle between the two vectors $\underline{x}$ and $\underline{y}$

Cosine distance

$$\cos\varphi = \frac{(\underline{x},\underline{y})}{||\underline{x}||\,||\underline{y}||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$

If the vectors $\underline{x}$, $\underline{y}$ have the norm equal to 1 then:

$$\cos\varphi = \sum_{i=1}^{n} x_i y_i = (\underline{x},\underline{y})$$

Ortogonality

Definition

$\underline{x}$ and $\underline{y}$ are ortogonal if and only if $(\underline{x}, \underline{y}) = 0$

Orthonormal basis

A set of linearly independent vectors $\{\underline{x}_1, \ldots, \underline{x}_n\}$ constitutes an orthonormal basis for the space V_n if and only if

$$\underline{x}_i, \underline{x}_j = \delta_{ij} = \left(\begin{array}{lll} 1 & \text{if} & i = j \\ 0 & \text{if} & i \neq j \end{array} \right)$$

Similarity

Applications

Document clusters provide often a structure for organizing large bodies of texts for efficient searching and browsing.
For example, recent advances in Internet search engines (e.g., http://vivisimo.com/, http://metacrawler.com/) exploit document cluster analysis.

Similarity

Applications

Document clusters provide often a structure for organizing large bodies of texts for efficient searching and browsing.
For example, recent advances in Internet search engines (e.g., http://vivisimo.com/, http://metacrawler.com/) exploit document cluster analysis.

Document and vectors

For this purpose, a document is commonly represented as a *vector* consisting of the suitably normalized frequency counts of words or terms.
Each document typically contains only a small percentage of all the words ever used. If we consider each document as a multi-dimensional vector and then try to cluster documents based on their word contents, the problem differs from classic clustering scenarios in several ways.

Similarity

The role of similarity among vectors

Document data is high-dimensional, characterized by a very sparse term-document matrix with positive ordinal attribute values and a significant amount of outliers.

Similarity

The role of similarity among vectors

Document data is high-dimensional, characterized by a very sparse term-document matrix with positive ordinal attribute values and a significant amount of outliers. In such situations, one is truly faced with the ‘curse of dimensionality’ issue since, even after feature reduction, one is left with **hundreds of dimensions** per object.

Similarity

In the relationship-based clustering process, key cluster analysis activities can be associated with each step:

Clustering steps

- *Representation of raw objects* (i.e. documents) into *vectors* of properties with real-valued scores (weights)

Similarity

In the relationship-based clustering process, key cluster analysis activities can be associated with each step:

Clustering steps

- *Representation of raw objects* (i.e. documents) into *vectors* of properties with real-valued scores (weights)
- Definition of a *proximity measure*

Similarity

In the relationship-based clustering process, key cluster analysis activities can be associated with each step:

Clustering steps

- *Representation of raw objects* (i.e. documents) into *vectors* of properties with real-valued scores (weights)
- Definition of a *proximity measure*
- Clustering algorithm

Similarity

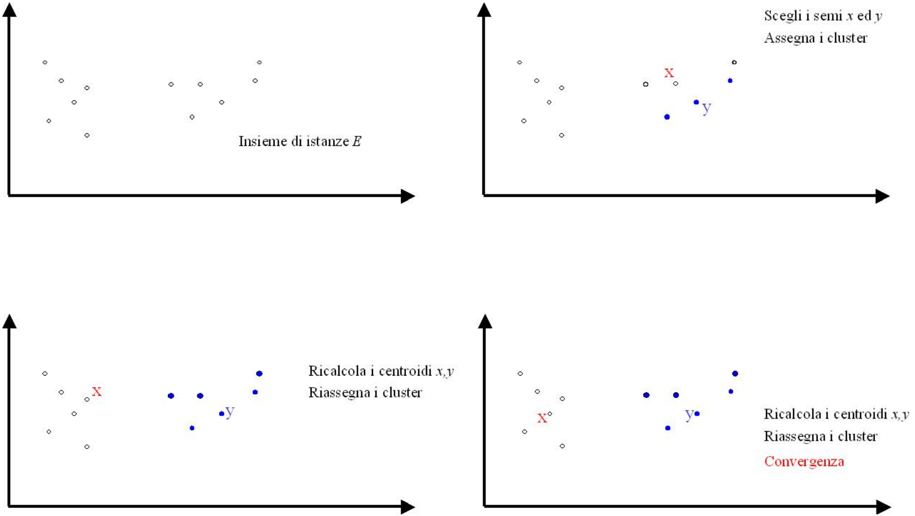
In the relationship-based clustering process, key cluster analysis activities can be associated with each step:

Clustering steps

- *Representation of raw objects* (i.e. documents) into *vectors* of properties with real-valued scores (weights)
- Definition of a *proximity measure*
- Clustering algorithm
- Evaluation

Similarity and Clustering

A well-known example of clustering algorithm is k-mean.



Similarity

Clustering steps

- To obtain features $\mathbf{X} \in \mathscr{F}$ from the raw objects, a suitable object representation has to be found.

Similarity

Clustering steps

- To obtain features $\mathbf{X} \in \mathscr{F}$ from the raw objects, a suitable object representation has to be found.
- Given an objext $O \in \mathscr{D}$, we will refer to such a representation as the feature vector $\underline{x}$ of X.

Similarity

Clustering steps

- To obtain features $\mathbf{X} \in \mathscr{F}$ from the raw objects, a suitable object representation has to be found.
- Given an objext $O \in \mathscr{D}$, we will refer to such a representation as the feature vector $\underline{x}$ of X.
- In the second step, a measure of proximity $\mathbf{S} \in \mathscr{S}$ has to be defined between objects, i.e. $\mathbf{S} : \mathscr{D}^2 \rightarrow \mathbb{R}$.

Similarity

Clustering steps

- To obtain features $\mathbf{X} \in \mathscr{F}$ from the raw objects, a suitable object representation has to be found.
- Given an objext $O \in \mathscr{D}$, we will refer to such a representation as the feature vector $\underline{x}$ of X.
- In the second step, a measure of proximity $\mathbf{S} \in \mathscr{S}$ has to be defined between objects, i.e. $\mathbf{S} : \mathscr{D}^2 \to \mathbb{R}$. The choice of similarity or distance can have a deep impact on clustering quality.

Minkowski distances

Minkowski distances

The *Minkowski distances* $L_p(\underline{x}, \underline{y})$ defined as:

$$L_p(\underline{x}, \underline{y}) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$$

are the standard metrics for geometrical problems.

Minkowski distances

Minkowski distances

The *Minkowski distances* $L_p(\underline{x},\underline{y})$ defined as:

$$L_p(\underline{x},\underline{y}) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$$

are the standard metrics for geometrical problems.

Euclidean Distance

For $p = 2$ we obtain the Euclidean distance, $d(\underline{x},\underline{y}) = \|\underline{x} - \underline{y}\|_2^2$.

Minkowski distances

There are several possibilities for converting an $L_p(\underline{x}, \underline{y})$ distance metric (in $[0, \inf)$, with 0 closest) into a *similarity measure* (in $[0, 1]$, with 1 closest) by a monotonic decreasing function.

Minkowski distances

There are several possibilities for converting an $L_p(\underline{x}, \underline{y})$ distance metric (in $[0, \inf)$, with 0 closest) into a *similarity measure* (in $[0, 1]$, with 1 closest) by a monotonic decreasing function.

Relation between distances and similarities

For Euclidean space, we chose to relate distances d and similarities s using

$$s = e^{-d^2}$$

Minkowski distances

There are several possibilities for converting an $L_p(\underline{x}, \underline{y})$ distance metric (in $[0, \inf)$, with 0 closest) into a *similarity measure* (in $[0, 1]$, with 1 closest) by a monotonic decreasing function.

Relation between distances and similarities

For Euclidean space, we chose to relate distances d and similarities s using

$$s = e^{-d^2}$$

Consequently, the *Euclidean* [0,1]-*normalized similarity* is defined as:

$$s^{(\mathrm{E})}(\underline{x}, \underline{y}) = e^{-\|\underline{x}-\underline{y}\|_2^2}$$

Pearson Correlation

Pearson Correlation

In collaborative filtering, correlation is often used to predict a feature from a highly similar mentor group of objects whose features are known.
The [0,1]-*normalized Pearson correlation* is defined as:

$$s^{(\mathrm{P})}(\underline{x},\underline{y}) = \frac{1}{2}\left(\frac{(\underline{x}-\bar{x})^T(\underline{y}-\bar{y})}{\|\underline{x}-\bar{x}\|_2 \cdot \|\underline{y}-\bar{y}\|_2} + 1\right),$$

where $\bar{x}$ denotes the average feature value of $\underline{x}$ over all dimensions.

Pearson Correlation

Pearson Correlation

The [0,1]-*normalized Pearson correlation* can also be seen as a probabilistic measure as in:

$$s^{(\mathrm{P})}(\underline{x}, \underline{y}) = r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y},$$

where $\bar{x}$ denotes the average feature value of $\underline{x}$ over all dimensions, and s_x and s_y are the standard deviations of $\underline{x}$ and $\underline{y}$, respectively.

Pearson Correlation

Pearson Correlation

The [0,1]-*normalized Pearson correlation* can also be seen as a probabilistic measure as in:

$$s^{(\mathrm{P})}(\underline{x}, \underline{y}) = r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1) s_x s_y},$$

where $\bar{x}$ denotes the average feature value of $\underline{x}$ over all dimensions, and s_x and s_y are the standard deviations of $\underline{x}$ and $\underline{y}$, respectively.

The correlation is defined only if both of the standard deviations are finite and both of them are nonzero. It is a corollary of the Cauchy-Schwarz inequality that the correlation cannot exceed 1 in absolute value.

Pearson Correlation

Pearson Correlation

The [0,1]-*normalized Pearson correlation* can also be seen as a probabilistic measure as in:

$$s^{(\mathrm{P})}(\underline{x}, \underline{y}) = r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y},$$

where $\bar{x}$ denotes the average feature value of $\underline{x}$ over all dimensions, and s_x and s_y are the standard deviations of $\underline{x}$ and $\underline{y}$, respectively.

The correlation is defined only if both of the standard deviations are finite and both of them are nonzero. It is a corollary of the Cauchy-Schwarz inequality that the correlation cannot exceed 1 in absolute value. The correlation is 1 in the case of an increasing linear relationship, -1 in the case of a decreasing linear relationship, and some value in between in all other cases, indicating the degree of linear dependence between the variables.

Jaccard Similarity

Binary Jaccard Similarity

The *binary Jaccard coefficient* measures the degree of overlap between two sets and is computed as the ratio of the number of shared features of $\underline{x}$ AND $\underline{y}$ to the number possessed by $\underline{x}$ OR $\underline{y}$.

Jaccard Similarity

Binary Jaccard Similarity

The *binary Jaccard coefficient* measures the degree of overlap between two sets and is computed as the ratio of the number of shared features of $\underline{x}$ AND $\underline{y}$ to the number possessed by $\underline{x}$ OR $\underline{y}$.

Example

For example, given two sets' binary indicator vectors $\underline{x} = (0,1,1,0)^T$ and $\underline{y} = (1,1,0,0)^T$, the cardinality of their intersect is 1 and the cardinality of their union is 3, rendering their Jaccard coefficient 1/3.

The binary Jaccard coefficient it is often used in retail market-basket applications.

Extended Jaccard Similarity

Extended Jaccard Similarity

The *extended Jaccard coefficient* is the generalized notion of the binary case and it is computed as:

$$s^{(\mathrm{J})}(\underline{x}, \underline{y}) = \frac{\underline{x}^T \underline{y}}{\|\underline{x}\|_2^2 + \|\underline{y}\|_2^2 - \underline{x}^T \underline{y}}$$

Dice coefficient

Dice coefficient

Another similarity measure highly related to the extended Jaccard is the *Dice coefficient*:

$$s^{(\mathrm{D})}(\underline{x}, \underline{y}) = \frac{2\underline{x}^T \underline{y}}{\|\underline{x}\|_2^2 + \|\underline{y}\|_2^2}$$

Dice coefficient

Dice coefficient

Another similarity measure highly related to the extended Jaccard is the *Dice coefficient*:

$$s^{(\mathrm{D})}(\underline{x}, \underline{y}) = \frac{2\underline{x}^T \underline{y}}{\|\underline{x}\|_2^2 + \|\underline{y}\|_2^2}$$

The Dice coefficient can be obtained from the extended Jaccard coefficient by adding $\underline{x}^T \underline{y}$ to both the numerator and denominator.

Similarity: discussion

Scale and Translation invariance

Euclidean similarity is *translation invariant* ...

Similarity: discussion

Scale and Translation invariance

Euclidean similarity is *translation invariant* ...
but *scale sensitive*

Similarity: discussion

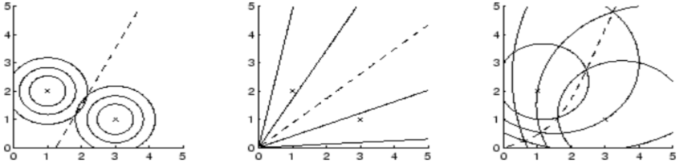
Scale and Translation invariance

Euclidean similarity is *translation invariant* ...
but *scale sensitive* while cosine is *translation sensitive* but *scale invariant*.

Similarity: discussion

Scale and Translation invariance

Euclidean similarity is *translation invariant* ...
but *scale sensitive* while cosine is *translation sensitive* but *scale invariant*.
The extended Jaccard has aspects of both properties as illustrated in figure.
Iso-similarity lines at $s = 0.25$, 0.5 and 0.75 for points $\underline{x} = (3,1)^T$ and $\underline{y} = (1,2)^T$ are shown for Euclidean, cosine, and the extended Jaccard.



(a) (b) (c)

Figure 4.1: Properties of (a) Euclidean-based, (b) cosine, and (c) extended Jaccard similarity measures illustrated in 2 dimensions. Two points $(1,2)^\dagger$ and $(3,1)^\dagger$ are marked with $\times$s. For each point iso-similarity surfaces for $s = 0.25$, 0.5, and 0.75 are shown with solid lines. The surface that is equi-similar to the two points is marked with a dashed line.

Similarity: discussion

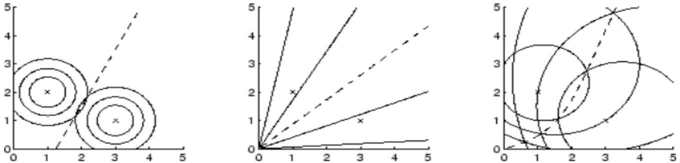


Figure 4.1: Properties of (a) Euclidean-based, (b) cosine, and (c) extended Jaccard similarity measures illustrated in 2 dimensions. Two points $(1,2)^\dagger$ and $(3,1)^\dagger$ are marked with $\times$s. For each point iso-similarity surfaces for $s = 0.25$, 0.5, and 0.75 are shown with solid lines. The surface that is equi-similar to the two points is marked with a dashed line.

Thus, for $s^{(\mathrm{J})} \to 0$, extended Jaccard behaves like the cosine measure, and for $s^{(\mathrm{J})} \to 1$, it behaves like the Euclidean distance

Similarity: discussion

Similarity in Clustering

In traditional Euclidean k-means clustering the optimal cluster representative $\mathbf{c}_\ell$ minimizes the sum of squared error criterion, i.e.,

$$\mathbf{c}_\ell = \arg\min_{\underline{z} \in \mathscr{F}} \sum_{\underline{x}_j \in \mathscr{C}_\ell} \|\underline{x}_j - \underline{z}\|_2^2$$

Any convex distance-based objective can be translated and extended to the similarity space.

Similarity: discussion

Swtiching from distances to similarity

Consider the generalized objective function $f(\mathscr{C}_\ell, \bar{z})$ given a cluster $\mathscr{C}_\ell$ and a representative $\bar{z}$:

$$f(\mathscr{C}_\ell, \bar{z}) = \sum_{\underline{x}_j \in \mathscr{C}_\ell} d(\underline{x}_j, \bar{z})^2 = \sum_{\underline{x}_j \in \mathscr{C}_\ell} \|\underline{x} - \bar{z}\|_2^2.$$

We use the transformation $s = e^{-d^2}$ to express the objective in terms of similarity rather than distance:

$$f(\mathscr{C}_\ell, \bar{z}) = \sum_{\underline{x}_j \in \mathscr{C}_\ell} -\log(s(\underline{x}_j, \bar{z}))$$

Similarity: discussion

Switching from distances to similarity

Finally, we simplify and transform the objective using a strictly monotonic decreasing function. Instead of minimizing $f(\mathscr{C}_\ell, \bar{z})$, we maximize

$$f'(\mathscr{C}_\ell, \bar{z}) = e^{-f(\mathscr{C}_\ell, \bar{z})}$$

Thus, in the similarity space, the least squared error representative $\mathbf{c}_\ell \in \mathscr{F}$ for a cluster $\mathscr{C}_\ell$ satisfies:

$$\mathbf{c}_\ell = \arg\max_{\bar{z} \in \mathscr{F}} \prod_{\underline{x}_j \in \mathscr{C}_\ell} s(\underline{x}_j, \bar{z})$$

Using the concave evaluation function f', we can obtain optimal representatives for non-Euclidean similarity spaces $\mathscr{S}$.

Similarity: discussion

To illustrate the values of the evaluation function $f'(\{\mathbf{x}_1, \mathbf{x}_2\}, \mathbf{z})$ are used to shade the background in the figure below.

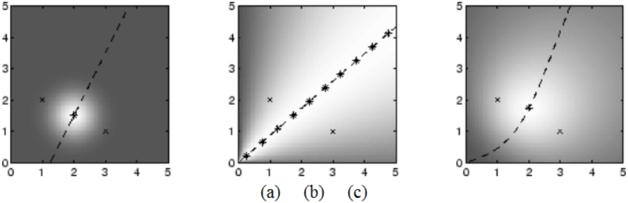


Figure 4.2: More similarity properties shown on the 2-dimensional example of figure 4.1. The goodness of a location as the common representative of the two points is indicated with brightness. The best representative is marked with a $\star$. The extended Jaccard (c) adopts the middle ground between Euclidean (a) and cosine-based similarity (b).

The maximum likelihood representative of $\underline{x}_1$ and $\underline{x}_2$ is marked with a $\star$.

Similarity: discussion

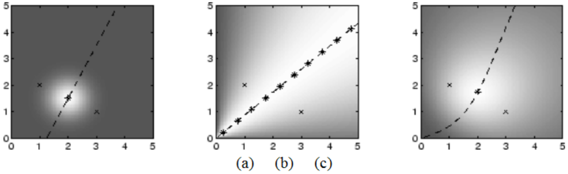


Figure 4.2: More similarity properties shown on the 2-dimensional example of figure 4.1. The goodness of a location as the common representative of the two points is indicated with brightness. The best representative is marked with a ⋆. The extended Jaccard (c) adopts the middle ground between Euclidean (a) and cosine-based similarity (b).

For cosine similarity all points on the equi-similarity are optimal representatives. In a maximum likelihood interpretation, we constructed the distance similarity transformation such that

$$p(\bar{z}|\mathbf{c}_\ell) \sim s(\bar{z}, \mathbf{c}_\ell)$$

Consequently, we can use the dual interpretations of probabilities in similarity space $\mathscr{S}$ and errors in distance space $\mathbb{R}$.

Further similarity measures

Vector similarities

- Grefenstette (fuzzy) set-oriented similarity for capturing dependency relations (head words)

Distributional (Probabilstic) similarities

- Lin similarity (commonalities) (Dice like)

$$sim(\underline{x},\underline{y}) = \frac{2 \cdot \log P(common(\underline{x},\underline{y}))}{\log P(\underline{x}) + \log P(\underline{y})}$$

- *Jensen-Shannon* total divergence to the mean:

$$A(p,q) = D(p\|\frac{p+q}{2}) + D(q\|\frac{p+q}{2})$$

- α-skewed divergence (Lee, 1999): $s_\alpha(p,q) = D(p\|\alpha p + (1-\alpha)q)$ ($\alpha = 0, 1$ or 0.01)

References

Vectors, Operations, Norms and Distances

K. Van Rijesbergen, The Geometry of Information Retrieval, CUP Press, 2004.

References

Vectors, Operations, Norms and Distances

K. Van Rijesbergen, The Geometry of Information Retrieval, CUP Press, 2004.

Distances and Similarities

Alexander Strehl, Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining, PhD Dissertation, University of Texas at Austin, 2002. URL: `http://www.lans.ece.utexas.edu/~strehl/diss/htdi.html`.

References

Vectors, Operations, Norms and Distances

K. Van Rijesbergen, The Geometry of Information Retrieval, CUP Press, 2004.

Distances and Similarities

Alexander Strehl, Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining, PhD Dissertation, University of Texas at Austin, 2002. URL: `http://www.lans.ece.utexas.edu/~strehl/diss/htdi.html`.

Nice collection of code and definitions

Sam- string metrics. URL: `http://www.dcs.shef.ac.uk/~sam/stringmetrics.html`.