Spazi vettoriali e misure di similarità

R. Basili

Corso di Web Mining e Retrieval
a.a. 2008-9

April 10, 2009
Outline

- Spazi vettoriali a valori reali
- Operazioni tra vettori
- Indipendenza Lineare
- Basi
- Prodotto Interno
- Norma di un vettore e Proprietá
- Vettori unitari
- Ortogonalità
- Similarità
- Norme e similarità
Real-valued Vector Space

Vector Space definition:

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

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:

**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 \( \Phi \) (called the origin) such that
   \[ \underline{x} + \Phi = \underline{x} \quad \forall \underline{x} \in V \]
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 \( \alpha \) and \( \underline{x} \), where \( \alpha \) is a scalar and \( \underline{x} \in V \), there corresponds a vector \( \alpha \underline{x} \), called the product of \( \alpha \) and \( \underline{x} \), in such a way that:

1. associativity \( \alpha (\beta \underline{x}) = (\alpha \beta) \underline{x} \)
2. \( 1 \underline{x} = \underline{x} \quad \forall \underline{x} \in V \)
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 \( \mathbf{x} \) and \( \mathbf{y} \)**

\[
\mathbf{x} + \mathbf{y} = |\mathbf{x}\rangle + |\mathbf{y}\rangle = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ \vdots \\ x_n + y_n \end{pmatrix}
\]

**Linear combination**

\[
y = c_1x_1 + \cdots + c_nx_n
\]

or

\[
|y\rangle = c_1|x_1\rangle + \cdots + c_n|x_n\rangle
\]

**Multiplication by scalar \( \alpha \)**

\[
\alpha \mathbf{x} = \alpha |\mathbf{x}\rangle = \begin{pmatrix} \alpha x_1 \\ \vdots \\ \vdots \\ \alpha x_n \end{pmatrix}
\]
Linear dependence

**Conditions for linear dependence**

A set of vectors \( \{x_1, \ldots, x_n\} \) are *linearly dependent* if there a set constant scalars \( c_1, \ldots, c_n \) exists, not all 0, such that:

\[
c_1 x_1 + \cdots + c_n x_n = 0
\]

**Conditions for linear independence**

A set of vectors \( \{x_1, \ldots, x_n\} \) are *linearly independent* if and only if the linear condition \( c_1 x_1 + \cdots + c_n x_n = 0 \) is satisfied only when \( c_1 = c_2 = \cdots = c_n = 0 \)
**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. $(x, y) = (y, x)$
2. $(x, \lambda y) = \lambda (x, y)$
3. $(x_1 + x_2, y) = (x_1, y) + (x_2, y)$
4. $(x, x) \geq 0$ and $(x, x) = 0 \text{ iff } x = 0$

**Standard Inner Product**

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

**Other notations**

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

**Geometric interpretation**
Geometrically the *norm* represent the length of the vector

**Definition**
The *norm* is a function $\| \cdot \| : V_n \to \mathbb{R}$

---

Euclidean Norm:

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

**Properties**

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

A vector $x \in V_n$ is a *unit vector*, or *normalsized*, when $\| x \| = 1$
From Norm to distance

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

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

These measure, noted sometimes as $||\underline{x} - \underline{y}||^2$, 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:

\[-1 \leq \frac{(x,y)}{||x|| \cdot ||y||} \leq 1\]

and therefore we can express it as:

\[(x, y) = ||x|| \cdot ||y|| \cos \varphi \quad 0 \leq \varphi \leq \pi\]

where \(\varphi\) is the angle between the two vectors \(x\) and \(y\)

Cosine distance

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

If the vectors \(x, y\) have the norm equal to 1 then:

\[\cos \varphi = \sum_{i=1}^{n} x_i y_i = (x, y)\]
**Ortgonality**

**Definition**

$x$ and $y$ are orthogonal if and only if $(x, y) = 0$

**Orthonormal basis**

A set of linearly independent vectors $\{x_1, \ldots, x_n\}$ constitutes an orthonormal basis for the space $V_n$ if and only if

$$x_i, x_j = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$
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.
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.
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 $\mathbf{X} \in \mathcal{F}$ from the raw objects, a suitable object representation has to be found. Given an object $O \in \mathcal{D}$, we will refer to such a representation as the feature vector $\mathbf{x}$ of $X$.

- In the second step, a measure of proximity $S \in \mathcal{S}$ has to be defined between objects, i.e. $S : \mathcal{D}^2 \rightarrow \mathbb{R}$. The choice of similarity or distance can have a deep impact on clustering quality.
**Minkowski distances**

The *Minkowski distances* $L_p(x, y)$ defined as:

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

are the standard metrics for geometrical problems.

**Euclidean Distance**

For $p = 2$ we obtain the Euclidean distance, $\|x, y\|_2^2$. 
Minkowski distances

There are several possibilities for converting an $L_p(x, y)$ distance metric (in $[0, \text{inf})$, 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:

\[ s^{(E)}(x, y) = e^{-\|x-y\|^2_2} \]
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:

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

where \(\bar{x}\) denotes the average feature value of \(x\) over all dimensions.
Pearson Correlation

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

\[
s^{(P)}(\underline{x}, \underline{y}) = r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)sxsy},
\]

where \(\bar{x}\) denotes the average feature value of \(x\) over all dimensions, and \(s_x\) and \(s_y\) are the standard deviations of \(x\) and \(y\), 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 \( x \) \( \text{AND} \) \( y \) to the number possessed by \( x \) \( \text{OR} \) \( y \).

**Example**

For example, given two sets’ binary indicator vectors \( x = (0, 1, 1, 0)^T \) and \( y = (1, 1, 0, 0)^T \), 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

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

$$s^{(J)}(x, y) = \frac{x^T y}{\|x\|_2^2 + \|y\|_2^2 - x^T y}$$
Another similarity measure highly related to the extended Jaccard is the Dice coefficient:

\[ s^{(D)}(x, y) = \frac{2x^T y}{\|x\|_2^2 + \|y\|_2^2} \]

The Dice coefficient can be obtained from the extended Jaccard coefficient by adding \( x^T y \) to both the numerator and denominator.
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.

**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 $s^{(J)} \to 0$, extended Jaccard behaves like the cosine measure, and for $s^{(J)} \to 1$, it behaves like the Euclidean distance.
Similarity in Clustering

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

$$c_\ell = \arg \min_{\bar{z} \in \mathcal{F}} \sum_{x_j \in C_\ell} \|x_j - \bar{z}\|_2^2$$

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

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

\[
f(\mathcal{C}_\ell, \bar{z}) = \sum_{x_j \in \mathcal{C}_\ell} d(x_j, \bar{z})^2 = \sum_{x_j \in \mathcal{C}_\ell} \|x - \bar{z}\|_2^2.
\]

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

\[
f(\mathcal{C}_\ell, \bar{z}) = \sum_{x_j \in \mathcal{C}_\ell} -\log(s(x_j, \bar{z}))
\]
Switching from distances to similarity

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

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

Thus, in the similarity space, the least squared error representative $c_\ell \in \mathcal{F}$ for a cluster $C_\ell$ satisfies:

$$c_\ell = \arg \max_{\bar{z} \in \mathcal{F}} \prod_{x_j \in C_\ell} s(x_j, \bar{z})$$

Using the concave evaluation function $f'$, we can obtain optimal representatives for non-Euclidean similarity spaces $\mathcal{S}$. 
To illustrate the values of the evaluation function $f'(\{x_1, x_2\}, z)$ are used to shade the background in the figure below.

![Figure 4.2](image)

**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).

The maximum likelihood representative of $x_1$ and $x_2$ is marked with a ⋆.
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

$$p(\bar{z}|c_\ell) \sim s(\bar{z}, c_\ell)$$

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

**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).
Further similarity measures

Vector similarities

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

Distributional (Probabilistic) similarities

▶ Lin similarity (commonalities) (Dice like)

\[
\text{sim}(x, y) = \frac{\log P(\text{common}(x, y))}{\log(P(\text{desc}(x, y)))}
\]

▶ Jensen-Shannon total divergence to the mean:

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

▶ \(\alpha\)-skewed divergence (Lee, 1999): \(s_\alpha(p, q) = D(p\|\alpha p + (1 - \alpha)q)\)

\((\alpha = 0, 1 \text{ or } 0.01)\)
References

**Vectors, Operations, Norms and Distances**

**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: