

Text Classification: the Geometrical approach. Vector models, and similarity

R. Basili

Corso di *Deep Learning*
a.a. 2024-25

March 13, 2025

Outline

Outline

- 1 Overview
- 2 Vector Spaces
 - Inner Product, Norms and Distances
- 3 Distance, similarity and classification
 - The Rocchio TC model
 - Memory Based Learning
 - Distances and similarities
 - Distances and similarities: Discussion
 - Other Distance Metrics
 - Discussion
- 4 A digression: IT
- 5 Probabilistic Norms
 - Mutual Information
 - Probabilistic Norms
- 6 References

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
 $\underline{x} + \Phi = \underline{x} \quad \forall \underline{x} \in V$
- 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 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 $1 \underline{x} = \underline{x} \quad \forall \underline{x} \in V$
- 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 \underline{x} and \underline{y}

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

Basis

Definition:

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

Inner Product

Inner Product

Definition:

Is a real-valued function on the cross product $V_n \times V_n$ associating with each pair of vectors (x, y) a unique real number.

The function (\cdot, \cdot) 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** $\underline{x} = 0$

Standard Inner Product

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

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* is a function $||\cdot||$ from V_n to \mathbb{R}

Norm

Geometric interpretation

Geometrically the *norm* represent the length of the vector

Definition

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

Euclidean Norm:

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

Norm

Geometric interpretation

Geometrically the *norm* represent the length of the vector

Definition

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

Euclidean Norm:

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

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 $\forall \underline{x}, \underline{y}, |(x, y)| \leq \|\underline{x}\| \|\underline{y}\|$ (Cauchy-Schwartz)

Norm

Geometric interpretation

Geometrically the *norm* represent the length of the vector

Definition

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

Euclidean Norm:

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

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 $\forall \underline{x}, \underline{y}, |(x, y)| \leq \|\underline{x}\| \|\underline{y}\|$ (Cauchy-Schwartz)

Norm

Geometric interpretation

Geometrically the *norm* represent the length of the vector

Definition

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

Euclidean Norm:

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

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 $\forall \underline{x}, \underline{y}, |(x, y)| \leq \|\underline{x}\| \|\underline{y}\|$ (Cauchy-Schwartz)

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

From Norm to distance

In V_n we can define the distance between two vectors x and 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 $\|\underline{x} - \underline{y}\|_2^2$, 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:

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

where φ is the angle between the two vectors x and y

Cosine distance

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

Orthogonality

Definition

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

Orthonormal basis

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

$$(x_i, x_j) = \delta_{ij} = \begin{pmatrix} 1 & \text{if } i=j \\ 0 & \text{if } i \neq j \end{pmatrix}$$

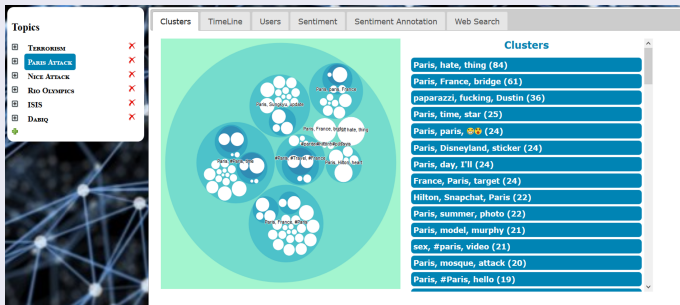
Similarity

Looking to texts as points a n -dimensional space

A structure for organizing large bodies of texts for efficient searching and browsing can be the notion of metric space.

Internet search engines may suitably exploit cluster analysis to documents in order to organize them visually.

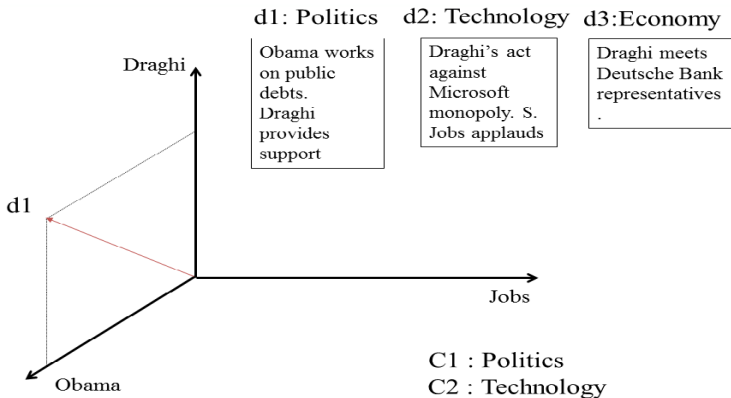
Clustering of texts for browsing



◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻

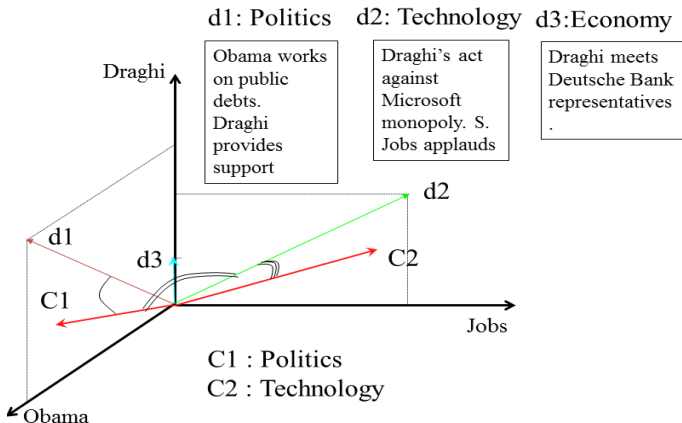
Text Classification: Vector Space Modeling

In Vector Space Model documents words corresponds to the space (orthonormal) basis, and individual texts are mapped into vectors ...



Text Classification: Classification Inference

Categories are also vectors and cosine similarity measures can support the final inference about category membership, e.g. $d1 \in C1$ and $d2 \in C2$:



T

T

T

T

T

T

- T

T

T

T

T

T

T

Representing Categories: the Rocchio model

The last step in providing a geometric account of text categorization is related to the representation of a category C_i .

Definition: Category Profile

A word w has a weight $\Omega(w, C_i)$ in a document category vector \underline{C}_i defined as:

$$\Omega(w, C_i) = \Omega_w^i = \max \left\{ 0, \frac{\beta}{|T_i|} \sum_{d \in T_i} \omega_w^d - \frac{\gamma}{|\bar{T}_i|} \sum_{d \in \bar{T}_i} \omega_w^d \right\}$$

where T_i is the set of training documents classified in C_i and $\overline{T_i}$ are the set of training document not classified in C_i

Representing Categories: the Rocchio model

The last step in providing a geometric account of text categorization is related to the representation of a category C_i .

Definition: Category Profile

A word w has a weight $\Omega(w, C_i)$ in a document category vector \underline{C}_i defined as:

$$\Omega(w, C_i) = \Omega_w^i = \max \left\{ 0, \frac{\beta}{|T_i|} \sum_{d \in T_i} \omega_w^d - \frac{\gamma}{|\bar{T}_i|} \sum_{d \in \bar{T}_i} \omega_w^d \right\}$$

where T_i is the set of training documents classified in C_i and $\overline{T_i}$ are the set of training document not classified in C_i

Rocchio: document and category vectors

Rocchio: document and category vectors

Rocchio: document and category vectors

Document and Category vectors are derived from the weights assigned to all the words in the vocabulary of a given collection.

A word is added to the vocabulary V whenever it appears in at least one document, although several feature selection methods can be applied.

Category Profile, \underline{C}_i

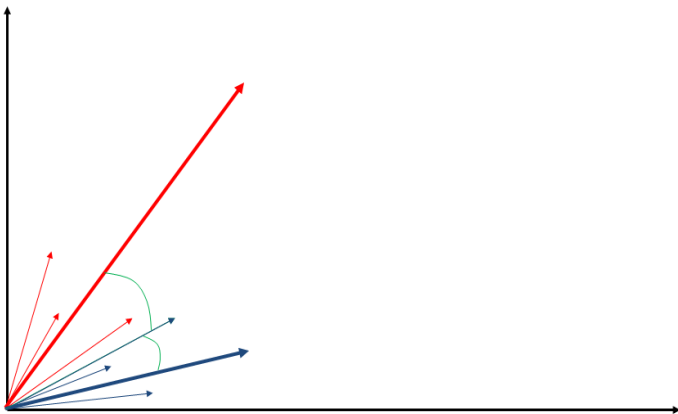
$$\underline{C}_i = \begin{pmatrix} \Omega_1^i \\ \vdots \\ \Omega_M^i \end{pmatrix}$$

Document Vector, \underline{d}

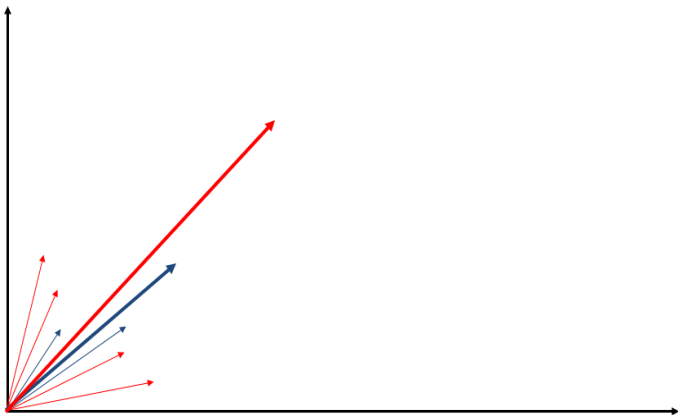
$$\underline{d} = \begin{pmatrix} \omega_1^d \\ \vdots \\ \omega_M^d \end{pmatrix}$$

Bidimensional View of Rocchio: classifying

As $(\underline{d}, \underline{C}_{red}) < (\underline{d}, \underline{C}_{blue})$ the new document d is lastly classified in the class of blue instances.



Prototype-based models have problems with polymorphic (i.e. disjunctive) categories.



Memory-based Learning

Memory-based learning: learning is just storing the representations of the training examples in the collection T .

Overview of MBL

The task is again:

- Testing instance x :
- Compute similarity between x and all examples in D .
- Assign x the **category of the most similar examples in D** .

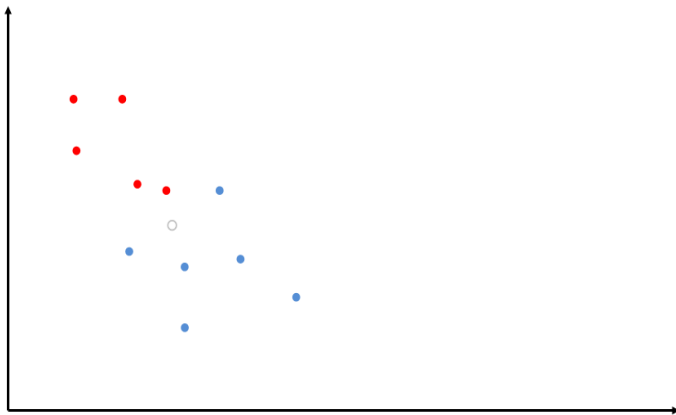
Does not explicitly compute a generalization or category prototypes.

Variants of MBL

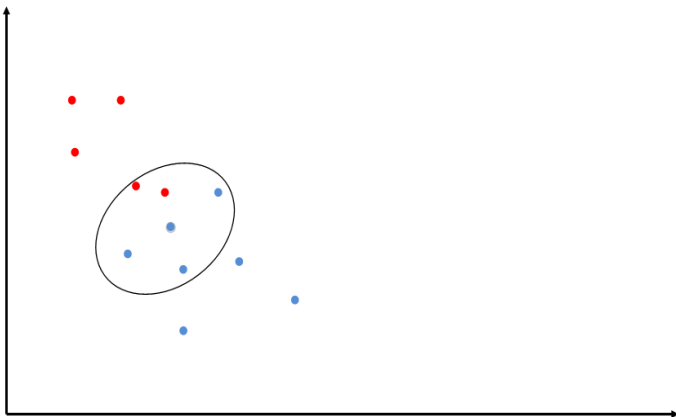
The general perspective of MBL is also called:

- Case-based (reasoning as retrieval of most similar cases)
- Memory-based (*memory* as examples are stored for later use)
- Lazy learning (*Lazy* as no model is built, so no generalization is attempted)

Labeled instances provides a rich description of a newly incoming instance within the space region close enough to the new example.



Whenever only the k instances closest to the example are used the k -NN algorithm is obtained through the voting across k labeled instances.



k-NN: the algorithm

For each each training example $\langle x, c(x) \rangle \in D$

Compute the corresponding TF-IDF vector, x , for document x .

Test instance y :

Compute TF-IDF vector y for document y .

For each $\langle x, c(x) \rangle \in D$

$$s_x = \cos Sim(\underline{y}, \underline{x}) = \frac{(\underline{y}, \underline{x})}{||\underline{x}|| \cdot ||\underline{y}||}$$

Sort examples $x \in D$ by decreasing values of s_x .

Let kNN be the set of the closest (i.e. first) k examples in D .

RETURN the majority class of examples in kNN .

Similarity

The role of similarity among vectors

In most of the examples above, document data are expressed as high-dimensional vectors, characterized by very sparse term-by-document matrices with positive ordinal attribute values and a significant amount of outliers.

Similarity

The role of similarity among vectors

In most of the examples above, document data are expressed as high-dimensional vectors, characterized by very sparse term-by-document matrices 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 and dimensionality reduction

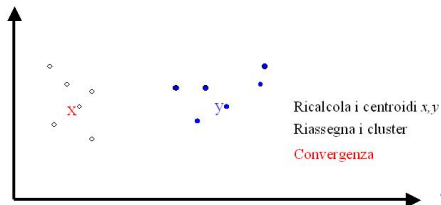
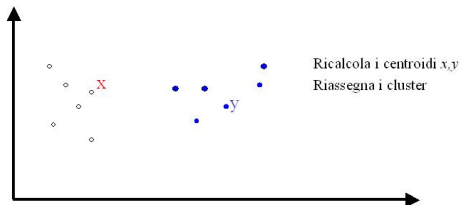
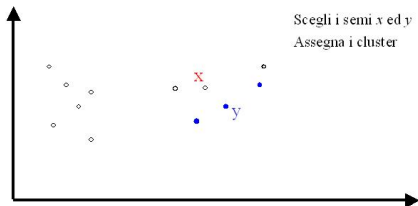
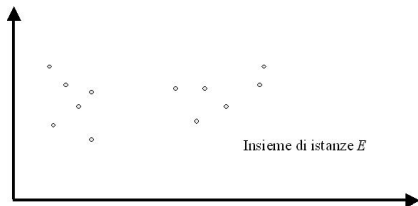
Clustering can be applied to documents to reduce the dimensions to take into account. Key cluster analysis activities can be thus devised:

Clustering steps

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

Similarity and Clustering

Clustering is a complex process as it requires a search within the set of all possible subsets. 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 \underline{x} of X .
- In the second step, a measure of proximity $\mathbf{S} \in \mathcal{S}$ has to be defined between objects, i.e. $\mathbf{S} : \mathcal{D}^2 \rightarrow \mathbb{R}$. **The choice of similarity or distance can have a deep impact on clustering quality.**

Minkowski distances

Minkowski distances

The *Minkowski distances* $L_p(\underline{x}, y)$ 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, $d(\underline{x}, y) = \|\underline{x} - y\|_2^2$.

Minkowski distances

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

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



Distance/similarity functions that have not a geometrical origin.

The role of probability

Very often objects in machine learning are described statistically, i.e. through the notion of distribution of probability that characterizes them: it serves to establish expectations about the values assumed by the object properties (e.g. how likely is 20 as the *age* of the instance of a “*young person*”).

Distances are this required to account for the likelihood that a value (e.g. 20) has with respect to others, and amplify (or decrease) the estimates according to such trends: this implies that non linear operators may arise and euclidean distances are not enough. Probability Theory and Information theory thus play a role in establishing some metrics that are useful in some Machine Learning tasks.

Distance/similarity functions that have not a geometrical origin.

Other evidence

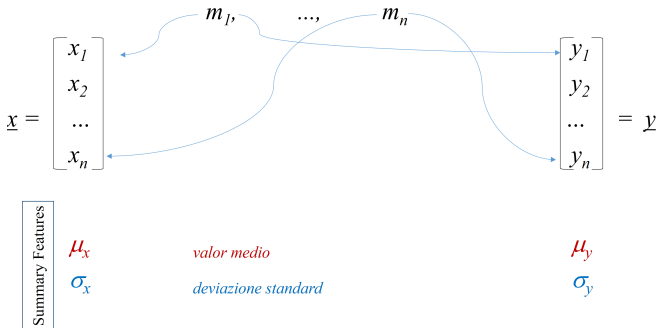
Other evidences also stem from extensions of the notion of standard set, such as the fuzzy sets. Fuzzy sets are usually characterized by smoothed membership functions that range not in the crisp set of $\{0, 1\}$ but in the full range of the $[0, 1]$ real valued interval.

In this cases, some definitions emerge from similarity operators deriving from standard set theory, such as the Dice and Jaccard measures.

Pearson Correlation:

objects ($\underline{x}, \underline{y}$), mentors (m_i) and features (x_i, y_i)

Vectors \underline{x} and \underline{y} are derived from mentor judgments as follows.



As a consequence, summary features are other useful descriptors of collective attitudes of mentors towards objects x and y .

Pearson Correlation

Pearson Correlation (2)

The $[0,1]$ -normalized Pearson correlation, $s^{(P)}$, is defined as:

$$s^{(\mathbf{P})}(\underline{x}, \underline{y}) \triangleq \frac{1}{2} \left(\frac{(\underline{x} - \underline{\mu}_x)^T (\underline{y} - \underline{\mu}_y)}{\|\underline{x} - \underline{\mu}_x\|_2 \cdot \|\underline{y} - \underline{\mu}_y\|_2} + 1 \right),$$

where $\underline{\mu}_c$ denotes the vector whose all components correspond to the average feature value μ_x of x , across all dimensions.

Pearson Correlation

Normalized Pearson Correlation

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

$$ns^{(P)}(\underline{x}, \underline{y}) \triangleq r_{xy} \triangleq \frac{\sum x_i y_i - n \mu_x \mu_y}{\sqrt{(\sum x_i^2 - n \mu_x^2)} \sqrt{(\sum y_i^2 - n \mu_y^2)}} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{(n-1) \sigma_x \sigma_y},$$

where μ_y denotes the average feature value of \underline{x} over all dimensions, and σ_x and σ_y are the standard deviations of x and y , respectively.

Normalized Pearson Correlation

The $[0,1]$ -normalized Pearson correlation:

$$r_{xy} \triangleq \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{(n-1)\sigma_x\sigma_y}$$

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.

Extended Jaccard Similarity

Extended Jaccard Similarity

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

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

Dice coefficient

Dice coefficient

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

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

The Dice coefficient can be obtained from the extended Jaccard coefficient by adding $x^T y$ 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.

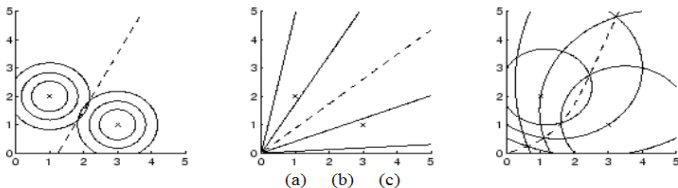


Figure 4.1: Properties of (a) Euclidean-based, (b) cosine, and (c) extended Jaccard similarity measures illustrated in 2 dimensions. Two points $(1, 2)^T$ and $(3, 1)^T$ are marked with \times s. For each point iso-similarity surfaces for $s = 0.25, 0.5$, and 0.75 are shown with solid lines. The surface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

Similarity: discussion

Similarity in Clustering

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

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

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

Similarity: discussion

Switching from distances to similarity

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

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

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

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

Similarity: discussion

Switching from distances to similarity

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

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

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

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

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

Information Theory

Let ξ be a discrete stochastic variable with a finite range $\Omega_\xi = \{x_1, \dots, x_M\}$ and let $p_i = p(x_i)$ be the corresponding probabilities.

How much information is there in knowing the outcome of ξ ?

Or equivalently:

How much uncertainty arises if the outcome ξ is unknown?

This is the information needed to specify which of the x_i has occurred. The problem is writing ξ .

Let us assume further that we only have a small set of symbols

$A = \{a_k : k = 1, \dots, D\}$, that is a *coding alphabet*.



Entropy

Uncertainty of ξ

The uncertainty introduced by the random variable ξ will be taken to be the *expectation value of the number of digits required to specify its outcome*. This is the expectation value of $-\log_2 P(\xi)$, i.e.

$$E[-\log_2 P(\xi)] = \sum_i -p_i \log_2 p_i$$

Entropy

The entropy $H[\xi]$ of ξ is precisely the amount of uncertainty introduced by the random variable ξ and it is more often referred to a natural logarithm $\ln(.)$, so that

$$H[\xi] = E[-\ln p(\xi)] = \sum_{x_i \in \Omega_\xi} -p(x_i) \ln p(x_i) = \sum_i^M -p_i \ln p_i$$

Entropy

Example 1: Rolling the dice

In the Die example, $\forall i = 1, \dots, 6$, it follows that $p_i = \frac{1}{6}$.

$$H[\xi] = E[-\ln p(\xi)] = \sum_{x_i \in \Omega_\xi} -p(x_i) \ln p(x_i) = 6 \cdot \frac{1}{6} \ln 6 = 1.792$$

Example 2: A loosing Die

A loosing Die: $p_1 = 1.00$, and $\forall i = 2, \dots, 6, p_i = 0$.

$$H[\xi] = E[-\ln p(\xi)] = \sum_{x_i \in \Omega_\xi} -p(x_i) \ln p(x_i) = 1 \ln 1 = 0$$

Entropy

Consequence

Given a distribution p_i ($i = 1, \dots, M$) for a discrete random variable ξ then for any other distribution q_i ($i = 1, \dots, M$) over the same sample space Ω_ξ it follows that:

$$H[\xi] = -\sum_i^M p_i \ln p_i \leq -\sum_i^M p_i \ln q_i$$

where equality holds **iff** the two distribution are the same, i.e.

$$\forall i = 1, \dots, M \quad p_i = q_i$$



Joint-Entropy

Given two random variable ξ and η :

Joint-Entropy

the *joint entropy* of ξ and η is defined as:

$$H[\xi, \eta] = - \sum_{i=1}^M \sum_{j=1}^L p(x_i, y_j) \ln p(x_i, y_j) = H[\eta, \xi]$$



Conditional-entropy

Conditional Entropy

the *conditional entropy* $H[\xi|\eta]$ of ξ and η is defined as:

$$\begin{aligned} H[\xi|\eta] &= - \sum_{j=1}^L p(y_j) \sum_{i=1}^M p(x_i|y_j) \ln p(x_i|y_j) = \\ &= - \sum_{j=1}^L \sum_{i=1}^M p(x_i, y_j) \ln p(x_i|y_j) \end{aligned}$$

Conditional and joint entropy

Conditional and Joint Entropy

The conditional and joint entropies are related just like the conditional and joint probabilities:

$$H[\xi, \eta] = H[\eta] + H[\xi | \eta]$$

Conveyed Information

The *information conveyed* by η , denoted $I[\xi|\eta]$, is the reduction in entropy of ξ by finding out the outcome of η . This is defined by:

$$I[\xi|\eta] = H[\xi] - H[\xi|\eta]$$

Mutual Information

Given two random variable ξ and η :

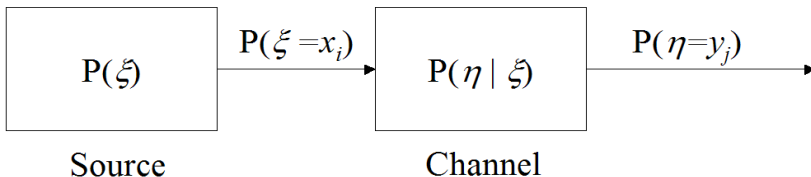
Mutual Information

The *mutual information* between ξ and η is defined as:

$$\begin{aligned} MI[\xi, \eta] &= E[\ln \frac{P(\xi, \eta)}{P(\xi) \cdot P(\eta)}] = \\ &= \sum_{(x,y) \in \Omega_{(\xi, \eta)}} f_{(\xi, \eta)}(x, y) \ln \frac{f_{(\xi, \eta)}(x, y)}{f_{\xi}(x) \cdot f_{\eta}(y)} \end{aligned}$$

Mutual Information

Mutual Information measures the amount of information about a random variable ξ an observer receives when the outcome of a random variable η is available.



How much information about the source output x_i does an observer gain by knowing the channel output y_i ?

Mutual Information

Mutual Information measures the amount of information about a random variable ξ an observer receives when the outcome of a random variable η is known, in fact:

Mutual Information

$$\begin{aligned} MI[\xi, \eta] &= H[\xi] - H[\xi|\eta] = \\ &= \sum_{(x,y) \in \Omega_{(\xi,\eta)}} f_{(\xi,\eta)}(x,y) \ln \frac{f_{(\xi,\eta)}(x,y)}{f_{\xi}(x) \cdot f_{\eta}(y)} \end{aligned}$$

Pointwise Mutual Information

Another way to look to mutual information is about the individual values (i.e. outcomes) $\xi = x_i$ and $\eta = y_j$.

Pointwise Mutual Information

Given the two random variable ξ and η : the *pointwise mutual information* between $\xi = x_i$ and $\eta = y_j$ is defined as:

$$MI[x_i, y_j] = f_{(\xi, \eta)}(x_i, y_j) \ln \frac{f_{(\xi, \eta)}(x_i, y_j)}{f_{\xi}(x_i) \cdot f_{\eta}(y_j)} = P(x_i, y_j) \ln \frac{P(x_i, y_j)}{P(x_i) \cdot P(y_j)}$$

Pointwise Mutual Information

Pointwise Mutual Information (pmi)

$$MI[x_i, y_j] = P(x_i, y_j) \ln \frac{P(x_i, y_j)}{P(x_i) \cdot P(y_j)}$$

Use of the pmi

If $MI[x_i, y_j] \gg 0$, there is a strong correlation between x_i and y_j

If $MI[x_i, y_j] \ll 0$, there is a strong negative correlation.

When $MI[x_i, y_j] \approx 0$ the two outcomes are almost independent.

Cross-entropy as a Norm

Cross-entropy

$$H_p[q] = - \sum_{x \in \Omega_\xi} p(x) \ln q(x)$$

Relative Entropy (or Kullback-Leibler distance)

$$D[p||q] = \sum_{x \in \Omega_{\xi}} p(x) \ln \frac{p(x)}{q(x)} = H_p[q] - H[p]$$

Cross-entropy and Norms

Relative Entropy (or Kullback-Leibler distance)

$$D[p||q] = \sum_{x \in \Omega_{\xi}} p(x) \ln \frac{p(x)}{q(x)} = H_p[q] - H[p]$$

KL distance: properties

$$D[p||q] \geq 0 \quad \forall q$$
$$D[p||q] = 0 \quad \text{iff } q = p$$

Norms, Similarity and Learning

Why ranking probability distributions is necessary?

- During a learning process we need to figure out the circumstances (i.e. the state of affairs of the world) under which a certain concept/class/property manifest.
- This make a direct reference to the probability of some (stochastic) event. Stochastic events are used to describe circumstances and properties.
- Moreover, learning proceeds from experience, i.e. known facts or previous classified examples, to rules, i.e. probability joint distributions over *decisions* and *circumstances*
- Learning in general means **to induce the proper probability distributions from the known examples**. There are several many ways to do it!!!

Norms, Similarity and Learning

Why ranking probability distributions is necessary?

- **Consequences.** In general, we need to compare different inductive hypothesis (IH), that are different probability distributions q_i of the same decision,
- In order to do it, we measure the agreement of our hypothesis with the observations (i.e. a pool of annotated data kept aside, the *held out*, to validate the different q_i)
- The result is an estimate of the similarity between the probability q_i induced at the i -th learning stage with the probability p characterizing the known examples.
- The KL divergence $D[p||q] = H_p(q) - H(p)$ can be the suitable dissimilarity function.
- The probability \hat{q} (such that \hat{q} minimizes $\forall i \quad D[p||q_i]$) is returned.

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

- Lin similarity (commonalities) (Dice like)

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

- *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)

Vector Space Modeling 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> or

<https://www.coli.uni-saarland.de/courses/LT1/2011/slides/string>

.

Probability and Information References

Elementary Information Theory

- in (Krenn & Samuelsson, 1997), Brigitte Krenn, Christer Samuelsson, *The Linguist's Guide to Statistics Don't Panic*, Univ. of Saarlandes, 1997.

URL: <http://nlp.stanford.edu/fsnlp/dontpanic.pdf>