Principal Component Analysis

Course of Machine Learning
Master Degree in Computer Science
University of Rome “Tor Vergata”
a.a. 2024-2025

Giorgio Gambosi

Curse of dimensionality

In general, many features: high-dimensional spaces.

- sparseness of data

High dimensions lead to difficulties in machine learning algorithms (lower reliability or need of large number of coefficients) this is denoted as **curse of dimensionality**

Dimensionality reduction

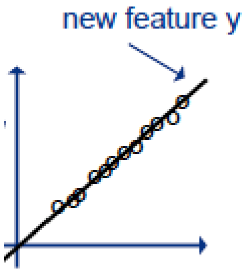
For any given classifier, the training set size required to obtain a certain accuracy grows exponentially wrt the number of features. It is then important to bound the number of features, identifying the less discriminant ones

- Feature selection: identify a subset of features which are still discriminant, or, in general, still represent most dataset variance

- Feature extraction: identify a projection of the dataset onto a lower-dimensional space, in such a way to still represent most dataset variance
  - Linear projection: principal component analysis, probabilistic PCA, factor analysis
  - Non linear projection: manifold learning, autoencoders

Searching hyperplanes for the dataset

Approach: verify whether training set elements lie on a hyperplane (a space of lower dimensionality), apart from a limited variability (which could be seen as noise)



Principal component analysis looks for a d'-dimensional subspace ($d' < d$) such that the projection of elements onto such subspace is a “faithful” representation of the original dataset. By “faithful” representation we mean that distances between elements and their projections are small, even minimal

PCA for $d' = 0$

Objective: represent all d-dimensional vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ by means of a unique vector $\mathbf{x}_0$, in the most faithful way, that is so that

$$J(\mathbf{x}_0) = \sum_{i=1}^{n} ||\mathbf{x}_0 - \mathbf{x}_i||^2$$

is minimum. It is easy to show that

$$\mathbf{x}_0 = \mathbf{m} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

In fact,

$$\begin{aligned}
J(\mathbf{x}_0) &= \sum_{i=1}^{n} ||(\mathbf{x}_0 - \mathbf{m}) - (\mathbf{x}_i - \mathbf{m})||^2 \\
&= \sum_{i=1}^{n} ||\mathbf{x}_0 - \mathbf{m}||^2 - 2\sum_{i=1}^{n} (\mathbf{x}_0 - \mathbf{m})^T (\mathbf{x}_i - \mathbf{m}) + \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{m}||^2 \\
&= \sum_{i=1}^{n} ||\mathbf{x}_0 - \mathbf{m}||^2 - 2(\mathbf{x}_0 - \mathbf{m})^T \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m}) + \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{m}||^2 \\
&= \sum_{i=1}^{n} ||\mathbf{x}_0 - \mathbf{m}||^2 + \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{m}||^2
\end{aligned}$$

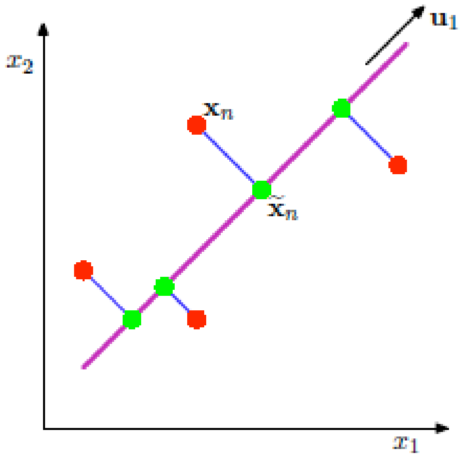
Since

$$\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m}) = \sum_{i=1}^{n} \mathbf{x}_i - n \cdot \mathbf{m} = n \cdot \mathbf{m} - n \cdot \mathbf{m} = 0$$

the second term is independent from $\mathbf{x}_0$, while the first one is equal to zero for $\mathbf{x}_0 = \mathbf{m}$

PCA for $d' = 1$

A single vector is too concise a representation of the dataset: anything related to data variability gets lost: a more interesting case is the one when vectors are projected onto a line passing through $\mathbf{m}$.



Let $\mathbf{u}_1$ be unit vector ($||\mathbf{u}_1|| = 1$) in the line direction: the line equation is then

$$\mathbf{x} = \alpha \mathbf{u}_1 + \mathbf{m}$$

where α is the distance of $\mathbf{x}$ from $\mathbf{m}$ along the line.

Also, let $\tilde{\mathbf{x}}_i = \alpha_i \mathbf{u}_1 + \mathbf{m}$ be the projection of $\mathbf{x}_i$ ($i = 1, \ldots, n$) onto the line: given $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we wish to find the set of projections minimizing the quadratic error

The quadratic error is defined as

$$
\begin{aligned}
J(\alpha_1, \ldots, \alpha_n, \mathbf{u}_1) &= \sum_{i=1}^{n} ||\tilde{\mathbf{x}}_i - \mathbf{x}_i||^2 \\
&= \sum_{i=1}^{n} ||(\mathbf{m} + \alpha_i \mathbf{u}_1) - \mathbf{x}_i||^2 \\
&= \sum_{i=1}^{n} ||\alpha_i \mathbf{u}_1 - (\mathbf{x}_i - \mathbf{m})||^2 \\
&= \sum_{i=1}^{n} +\alpha_i^2 ||\mathbf{u}_1||^2 + \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{m}||^2 - 2 \sum_{i=1}^{n} \alpha_i \mathbf{u}_1^T (\mathbf{x}_i - \mathbf{m}) \\
&= \sum_{i=1}^{n} \alpha_i^2 + \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{m}||^2 - 2 \sum_{i=1}^{n} \alpha_i \mathbf{u}_1^T (\mathbf{x}_i - \mathbf{m})
\end{aligned}
$$

Its derivative wrt α_k is

$$
\frac{\partial}{\partial \alpha_k} J(\alpha_1, \ldots, \alpha_n, \mathbf{u}_1) = 2\alpha_k - 2\mathbf{u}_1^T (\mathbf{x}_k - \mathbf{m})
$$

which is zero when $\alpha_k = \mathbf{u}_1^T (\mathbf{x}_k - \mathbf{m})$ (the orthogonal projection of $\mathbf{x}_k$ onto the line).

The second derivative turns out to be positive

$$
\frac{\partial}{\partial \alpha_k^2} J(\alpha_1, \ldots, \alpha_n, \mathbf{u}_1) = 2
$$

showing that what we have found is indeed a minimum.

To derive the best direction $\mathbf{u}_1$ of the line, we consider the covariance matrix of the dataset

$$
\mathbf{S} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T
$$

By plugging the values computed for α_i into the definition of $J(\alpha_1, \ldots, \alpha_n, \mathbf{u}_1)$, we get

$$
\begin{aligned}
J(\mathbf{u}_1) &= \sum_{i=1}^{n} \alpha_i^2 + \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{m}||^2 - 2 \sum_{i=1}^{n} \alpha_i^2 \\
&= - \sum_{i=1}^{n} [\mathbf{u}_1^T (\mathbf{x}_i - \mathbf{m})]^2 + \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{m}||^2 \\
&= - \sum_{i=1}^{n} \mathbf{u}_1^T (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T \mathbf{u}_1 + \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{m}||^2 \\
&= -n \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 + \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{m}||^2
\end{aligned}
$$

Since $\mathbf{u}_1^T (\mathbf{x}_i - \mathbf{m})$ is the projection of $\mathbf{x}_i$ onto the line, the product

$$
\mathbf{u}_1^T (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T \mathbf{u}_1 = (\mathbf{u}_1^T (\mathbf{x}_i - \mathbf{m}))(\mathbf{u}_1^T (\mathbf{x}_i - \mathbf{m}))^T = ||\mathbf{u}_1^T (\mathbf{x}_i - \mathbf{m})||^2
$$

is the squared length of the projection of $\mathbf{x}_i - \mathbf{m}$ on the line, that is the squared distance, along the line, of the projection of $\mathbf{x}_i$ from the mean $\mathbf{m}$, and the sum

$$\sum_{i=1}^{n} \mathbf{u}_1^T(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T \mathbf{u}_1 = n\mathbf{u}_1^T \mathbf{S}\mathbf{u}_1$$

is the sum of such squared distances, which is proportional (by a factor n) to the overall variance of the projections of vectors $\mathbf{x}_i$ wrt the mean $\mathbf{m}$.

Minimizing $J(\mathbf{u}_1)$ is equivalent to maximizing $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$. That is, $J(\mathbf{u}_1)$ is minimum if $\mathbf{u}_1$ is the direction which keeps the maximum amount of variance in the dataset. Hence, we wish to maximize $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$ (wrt $\mathbf{u}_1$), with the constraint $||\mathbf{u}_1|| = 1$.

By applying Lagrange multipliers this results equivalent to maximizing

$$u = \mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 - \lambda_1(\mathbf{u}_1^T\mathbf{u}_1 - 1)$$

This can be done by setting the first derivative wrt $\mathbf{u}_1$ to 0:

$$\frac{\partial u}{\partial \mathbf{u}_1} = 2\mathbf{S}\mathbf{u}_1 - 2\lambda_1\mathbf{u}_1 = 0$$

obtaining

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1$$

Note that:

- u is maximized if $\mathbf{u}_1$ is an eigenvector of $\mathbf{S}$
- the overall variance of the projections is then equal to the corresponding eigenvalue

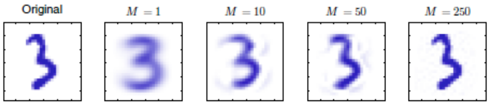
$$\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 = \mathbf{u}_1^T\lambda_1\mathbf{u}_1 = \lambda_1\mathbf{u}_1^T\mathbf{u}_1 = \lambda_1$$

- the variance of the projections is then maximized (and the error minimized) if $\mathbf{u}_1$ is the eigenvector of $\mathbf{S}$ corresponding to the maximum eigenvalue λ_1

As a consequence, the quadratic error is minimized by projecting vectors onto a hyperplane defined by the directions associated to the d' eigenvectors corresponding to the d' largest eigenvalues of $\mathbf{S}$. If we assume data are modeled by a d-dimensional gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ, PCA returns a d'-dimensional subspace corresponding to the hyperplane defined by the eigenvectors associated to the d' largest eigenvalues of Σ: the projections of vectors onto that hyperplane are distributed as a d'-dimensional distribution which keeps the maximum possible amount of data variability.

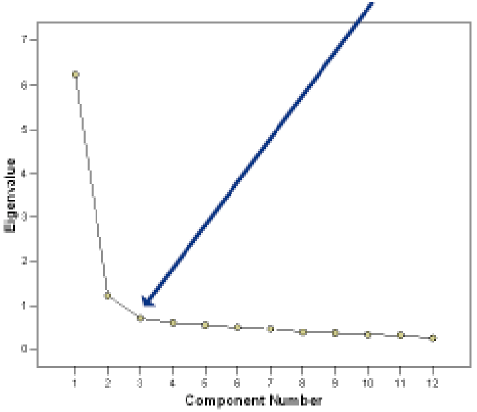
An example of PCA

Digit recognition ($D = 28 \times 28 = 784$)



Choosing d'

Eigenvalue size distribution is usually characterized by a fast initial decrease followed by a small decrease



This makes it possible to identify the number of eigenvalues to keep, and thus the dimensionality of the projections. Eigenvalues measure the amount of distribution variance kept in the projection.

Let us consider, for each $k < d$, the value

$$r_k = \frac{\sum_{i=1}^{k} \lambda_i^2}{\sum_{i=1}^{n} \lambda_i^2}$$

which provides a measure of the variance fraction associated to the k largest eigenvalues.

When $r_1 < \ldots < r_d$ are known, a certain amount p of variance can be kept by setting

$$d' = \underset{i \in \{1,\ldots,d\}}{\text{argmin}} \; r_i > p$$

Dimensionality reduction in co-occurrence data

Models referring to co-occurrence data: terms in documents, customer choices vs. items, people interacting each other,. . .

They consider, given two collections $\mathbf{V}, \mathbf{D}$ (for example, terms and documents, customers and items, people) a sequence of observations $\mathbb{W} = \{(w_1, d_1), \ldots, (w_N, d_N)\}$, with $w_i \in \mathbf{V}, d_i \in \mathbf{D}$ (for example, occurrences of terms in documents, customers accessing item description, pairs of people interacting, etc.)

Introduction to LSA

Basic assumptions

The approach of LSA (Latent Semantic Analysis) refers to three assumptions:

- ”semantic” information can be derived from the $\mathbf{V}, \mathbf{D}$ matrix
- dimensionality reduction is a key aspect for such derivation
- ”terms” and ”documents” can be modeled as points (vectors) in a euclidean space

Framework

1. Dictionary $\mathbf{V}$ of $V = |\mathbf{V}|$ terms $t_1, t_2, \ldots, t_V$
2. Corpus $\mathbf{D}$ of $D = |\mathbf{D}|$ documents $d_1, d_2, \ldots, d_D$
3. Each ”document” d_i is a sequence of N_i occurrences of ”terms” from $\mathbf{V}$

Model

1. A "document" d_i can be seen as a multiset of N_i "terms" in $\mathbf{V}$ (bag of words hypothesis in information retrieval)

2. There exists a correspondence between $\mathbf{V}$ and $\mathbf{D}$, and a vector space $\mathcal{S}$. Each term t_i has an associated vector $\mathbf{u}_i$. Also, a vector $\mathbf{v}_j$ in $\mathcal{S}$ is associated to each document d_j

Let us define the Occurrence matrix $\mathbf{W} \in \mathbb{R}^{V \times D}$, where $w_{i,j}$ is associated to the occurrences of term t_i into document d_j. The value $w_{i,j}$ derives from some measure of the number of occurrences of t_i into d_j (binary, count, tf, tf-idf, entropy, etc.).

- Terms corresponds to row vectors (size D): a "term" is defined only on the basis of the "documents" in which it occurs

- Documents correspond to column vector (size V): a "document" is defined only on the basis of the occurring "terms"

This representation has some problems:

1. The values V, D are usually quite large

2. Vectors corresponding to t_i and d_j are very sparse (no relation for most t_i, d_j pairs)

3. Terms and documents are modeled as vectors defined on different spaces ($\mathbb{R}^D$ and $\mathbb{R}^V$, respectively)

A more compact and uniform representation can be obtained by applying singular value decomposition.

Let $\mathbf{W} \in \mathbb{R}^{n \times m}$ be a matrix of rank $r \leq \min(n, m)$, and let $n > m$. Then, there exist

- $\mathbf{U} \in \mathbb{R}^{n \times r}$ orthonormal (that is $\mathbf{U}^T\mathbf{U} = \mathbf{I}_r$)

- $\mathbf{V} \in \mathbb{R}^{m \times r}$ orthonormal (that is $\mathbf{V}\mathbf{V}^T = \mathbf{I}_r$)

- $\Sigma \in \mathbb{R}^{r \times r}$ diagonal

such that $\mathbf{W} = \mathbf{U}\Sigma\mathbf{V}^T$.

Let us consider a matrix $\mathbf{W} \in \mathbb{R}^{n \times m}$, and let us denote as r the rank (number of linearly independent rows or columns) of $\mathbf{W}^T\mathbf{W} \in \mathbb{R}^{m \times m}$.

Several properties hold:

1. $\mathbf{W}^T\mathbf{W}$ is symmetric and semidefinite positive
   - a matrix $\mathbf{A}$ is symmetric iff $a_{ij} = a_{ji}$ for all i, j
   - a matrix $\mathbf{A}$ is semidefinite positive iff $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for all non null $\mathbf{x} \in \mathbb{R}^n$

2. All eigenvalues $\lambda_1, \ldots, \lambda_r$ of $\mathbf{W}^T\mathbf{W}$ are real and positive

3. The corresponding eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are orthonormal (they are orthogonal and have unitary length)

Let us define the singular values $\sigma_i = \sqrt{\lambda_i}$, $i = 1, \ldots, r$ and let us also consider vectors $\mathbf{u}_i = \dfrac{1}{\sigma_i}\mathbf{W}\mathbf{v}_i$, $i = 1, \ldots, r$. It is easy to show that $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are orthonormal

Let us also consider the following matrices

- $\mathbf{V} \in \mathbb{R}^{m \times r}$ having vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ as columns

$$\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r]$$

- $\mathbf{U} \in \mathbb{R}^{n \times r}$ having vectors $\mathbf{u}_1, \dots, \mathbf{u}_r$ as columns

$$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r]$$

- $\Sigma \in \mathbb{R}^{r \times r}$ having singular values on the diagonal

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix}$$

It is easy to verify that

$$\mathbf{W}\mathbf{V} = \mathbf{U}\Sigma$$

Moreover, since $\mathbf{V}$ is orthogonal, its is $\mathbf{V}^{-1} = \mathbf{V}^T$ and, as a consequence,

$$\mathbf{W} = \mathbf{U}\Sigma\mathbf{V}^T$$

Reassuming,

- The occurrences matrix $\mathbf{W}$ is decomposed in the product of three matrices.
- The term matrix $\mathbf{U}$, whose rows correspond to "terms"
- The document matrix $\mathbf{V}^T$, whose columns correspond to "documents"
- The singular values matrix Σ, which specifies the relevance of each dimension

Application of SVD: LSA

Key property: Each singular value tells us how important its dimension is. By setting less important dimensions to zero, we keep the important information, but get rid of the "details": this is equivalent to deleting rows (in $\mathbf{U}$), columns (in $\mathbf{V}$) and rows and columns (in Σ), corresponding to such less important dimensions (i.e. smaller singular values).

The dimension d of the new space may be predefined, and less than the rank of $\mathbf{W}$. In this case,

$$\mathbf{W} \approx \overline{\mathbf{W}} = \mathbf{U}\Sigma\mathbf{V}^T$$

The property

$$\min_{\mathbf{A}:\mathrm{rank}(\mathbf{A})=d} ||\mathbf{W} - \mathbf{A}||_2 = ||\mathbf{W} - \overline{\mathbf{W}}||_2$$

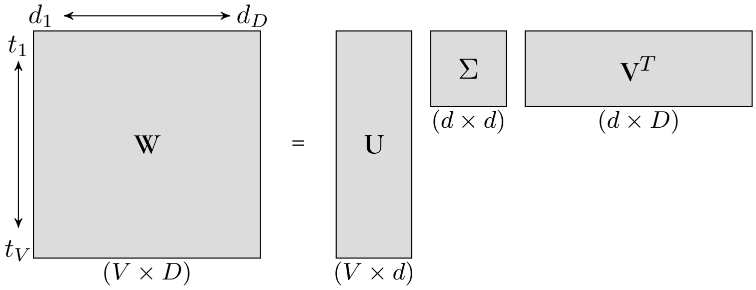
holds. The matrix $\overline{\mathbf{W}}$ is the matrix that best approximates $\mathbf{W}$ among all matrices of rank d according to the norm L_2 or of Frobenius

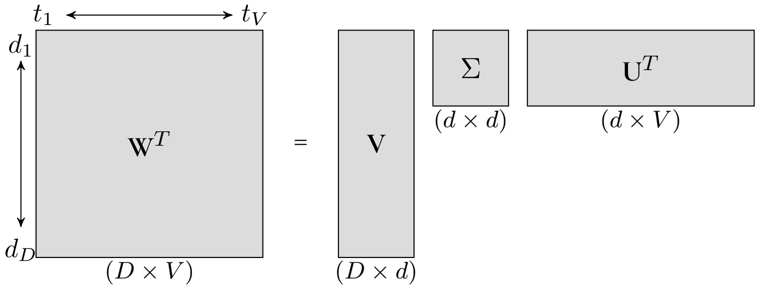
$$||\mathbf{A}||_2 = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}$$

SVD defines a transformation from two discrete vector spaces $\mathcal{V} \in \mathbb{Z}^D$ and $\mathcal{D} \in \mathbb{Z}^V$, to one smaller continuous vector space, $\mathcal{T} \in \mathbb{R}^d$.

The dimension of $\mathcal{T}$ is less than or equal to the rank (unknown) of $\mathbf{W}$, and it is lower bounded from the amount of distortion acceptable in the projection.

The rows of $\mathbf{W}$ (terms) are projected on a d-dimensional subspace of $\mathbb{R}^D$ having the set of columns of $\mathbf{V}$ as basis: this defines for each term a new representation (row of $\mathbf{U}\Sigma \in \mathbb{R}^d$) as a vector of the coordinates with respect to





this basis. In particular, each term is a vector wrt to that base, with set of coordinates given by $\mathbf{U}\Sigma \in \mathbb{R}^r$: value $u_{ik}\sigma_k$ provides a measure of the relevance of term t_i in the k-th "topic".

The rows of $\mathbf{W}^T$ (documents) are projected on a d-dimensional subspace of $\mathbb{R}^V$ having the set of columns of $\mathbf{U}$ as basis: this defines for each document a new representation (row of $\mathbf{V}\Sigma \in \mathbb{R}^d$) as a vector of the coordinates with respect to this basis. In particular, each document is a vector wrt to that base, with set of coordinates given by $\mathbf{V}\Sigma \in \mathbb{R}^r$: value $v_{jk}\sigma_k$ provides a measure of the presence of the k-th topic in document d_j.

Interpretation

$\hat{\mathbf{W}}$ captures the largest part of the associations between terms and documents $\mathbf{W}$, neglecting the least significative relations.

- Each term is represented as a (linear) combinations of hidden topics, corresponding to the columns of $\mathbf{V}$: terms with projections near to each other tend to appear in the same documents (or in semantically similar documents)
- Each document is represented as a (linear) combinations of hidden topics, corresponding to the columns of $\mathbf{U}$: documents with projections near to each other tend to include the same terms (or semantically similar terms)

LSA and clustering

Co-occurrences

- $\mathbf{W}\mathbf{W}^T \in \mathbb{Z}^{V\times V}$ represents the co-occurrences between terms in $\mathbf{V}$ (number of documents where the two terms both occur)
- $\mathbf{W}^T\mathbf{W} \in \mathbb{Z}^{D\times D}$ represents the co-occurrences between documents in $\mathbf{D}$ (number of terms in common between them)

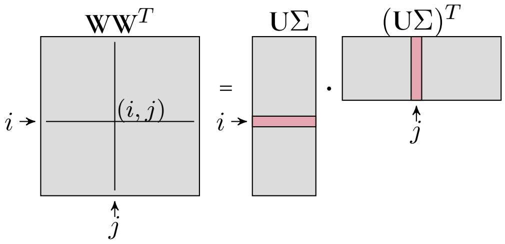
By applying SVD,

$$\mathbf{W}\mathbf{W}^T = \mathbf{U}\Sigma\mathbf{V}^T\mathbf{V}\Sigma\mathbf{U}^T = \mathbf{U}\Sigma^2\mathbf{U}^T$$

and

$$\mathbf{W}^T\mathbf{W} = \mathbf{V}\Sigma\mathbf{U}^T\mathbf{U}\Sigma\mathbf{V}^T = \mathbf{V}\Sigma^2\mathbf{V}^T$$

a term clustering is obtained.

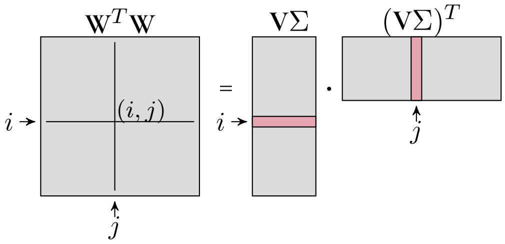


A reasonable measure of the proximity between terms t_i and t_j is given by the value of item (i,j) of $\mathbf{W}\mathbf{W}^T$, hence of the inner product between $\mathbf{u}_i$ (i-th row of $\mathbf{U}\Sigma$) and $\mathbf{u}_j$ (j-th row of $\mathbf{U}\Sigma$). In particolar,

$$\mathcal{D}(t_i,t_j) = \frac{1}{\cos(\mathbf{u}_i,\mathbf{u}_j)} = \frac{||\mathbf{u}_i|| \cdot ||\mathbf{u}_j||}{\mathbf{u}_i\mathbf{u}_j^T}$$

can be assumed as a measure of the distance between terms.

Documents can also be clustered.



A reasonable measure of the proximity between documents d_i and d_j is the value of item (i,j) of $\mathbf{W}^T\mathbf{W}$, hence of the inner product between $\mathbf{v}_i$ (i-th row of $\mathbf{V}\Sigma$) and $\mathbf{v}_j$ (j-th row of $\mathbf{V}\Sigma$). In particolar,

$$\mathcal{D}(d_i,d_j) = \frac{1}{\cos(\mathbf{v}_i,\mathbf{v}_j)} = \frac{||\mathbf{v}_i|| \cdot ||\mathbf{v}_j||}{\mathbf{v}_i\mathbf{v}_j^T}$$

can be assumed as a measure of the distance between documents

Classification

Determining, given a document, to which topic (in a predefined collection) its content is most related.

Approach: construction of a vector of (possibly weighted) terms, to describe the class. It can be seen as an additional document $\overline{d}$ (template of the class)

$\mathbf{W}$ can be extended by appending $\overline{d}$ as the $D+1$-th column of $\mathbf{W}$ (thus obtaining $\overline{\mathbf{W}} \in \mathbb{Z}^{V\times(D+1)}$)

SVD introduces an additional vector $\overline{\mathbf{v}} \in \mathbb{R}^d$ as $D+1$-th row of $\mathbf{V}$, where $\overline{d} = \mathbf{U}\Sigma\overline{\mathbf{v}}^T$

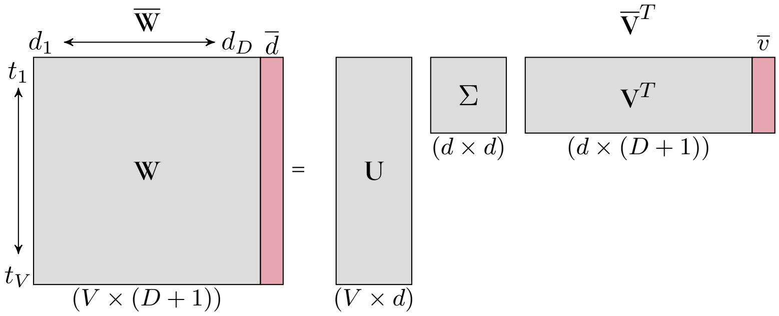
Proximity of a document to a topic

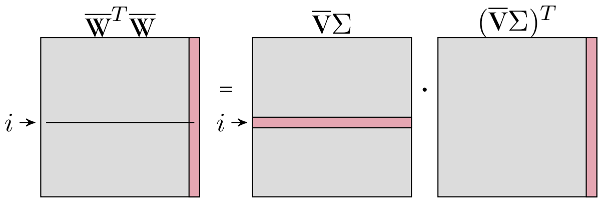
A reasonable measure of the proximity between a document d_i and a class $\overline{d}$ is given by the value of item $(i, D+1)$ of $\overline{\mathbf{W}}^T\overline{\mathbf{W}}$, hence of the inner product between $\mathbf{v}_i$ (i-th row ofi $\overline{\mathbf{V}}\Sigma$) and $\overline{\mathbf{v}}$ ($(D+1)$-th row of $\overline{\mathbf{V}}\Sigma$).

In particolar,

$$\mathcal{D}(d_i,\overline{d}) = \frac{1}{\cos(\mathbf{v}_i,\overline{\mathbf{v}})} = \frac{||\mathbf{v}_i|| \cdot ||\overline{\mathbf{v}}||}{\mathbf{v}_i\overline{\mathbf{v}}^T}$$

can be assumed as a measure of the distance between a document and a class


$\overline{\mathbf{W}}$
d_1
d_D
$\overline{d}$
t_1
t_V
$\mathbf{W}$
$(V \times (D+1))$
=
$\mathbf{U}$
$(V \times d)$
Σ
$(d \times d)$
$\overline{\mathbf{V}}^T$
$\overline{v}$
$\mathbf{V}^T$
$(d \times (D+1))$


$\overline{\mathbf{W}}^T\overline{\mathbf{W}}$
i
=
$\overline{\mathbf{V}}\Sigma$
i
$\cdot$
$(\overline{\mathbf{V}}\Sigma)^T$