Principal component analysis

Course of Machine Learning
Master Degree in Computer Science
University of Rome ``Tor Vergata''

Giorgio Gambosi

a.a. 2017-2018

In general, many features: high-dimensional spaces.

- sparseness of data
- increase in the number of coefficients, for example for dimension D and order 3 of the polynomial,

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k$$

  number of coefficients is $O(D^M)$

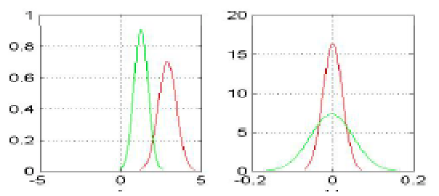
High dimensions lead to difficulties in machine learning algorithms (lower reliability or need of large number of coefficients) this is denoted as curse of dimensionality

Dimensionality reduction

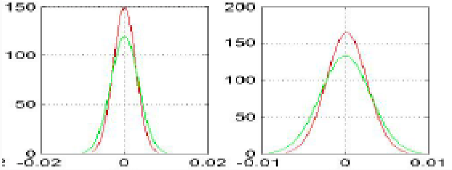
- for any given classifier, the training set size required to obtain a certain accuracy grows exponentially wrt the number of features (*curse of dimensionality*)
- it is important to bound the number of features, identifying the less discriminant ones

Discriminant features

- Discriminant feature: makes it possible to distinguish between two classes

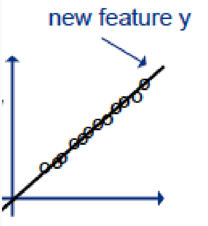


- Non discriminant feature: does not allow classes to be distinguished



Searching hyperplanes for the dataset

- verifying 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 suspace is a ``faithful'' representation of the original dataset
- as ``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 $\mathbf{x}_1, \ldots, \mathbf{x}_n$ 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,

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

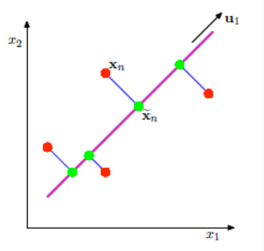
- since

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

- 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 $\mathbf{m}$



PCA for $d' = 1$

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

PCA for $d' = 1$

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

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

- $\mathbf{u}_1^T(\mathbf{x}_i - \mathbf{m})$ is the projection of $\mathbf{x}_i$ onto the line
- the product
$$\mathbf{u}_1^T(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T\mathbf{u}_1$$
is then the variance of the projection of $\mathbf{x}_i$ wrt the mean $\mathbf{m}$
- 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 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 $\mathbf{u}_1$:

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

to 0, obtaining

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

PCA for $d' = 1$

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

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

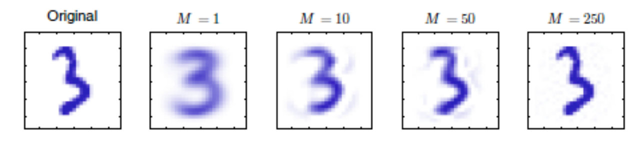
- the variance of the projections is then maximized (and the error minimized) if $\mathbf{u}_1$ is the eigenvector of $\mathbf{S}$ corresponding to the maximum eigenvalue λ_1

PCA for $d' > 1$

- 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 $\mathbf{\Sigma}$, PCA returns a d'-dimensional subspace corresponding to the hyperplane defined by the eigenvectors associated to the d' largest eigenvalues of $\mathbf{\Sigma}$
- The projections of vectors onto that hyperplane are distributed as a d'-dimensional distribution which keeps the maximum possible amount of data variability

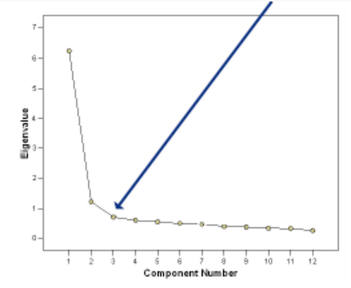
An example of PCA

- 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.

Choosing d'

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

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

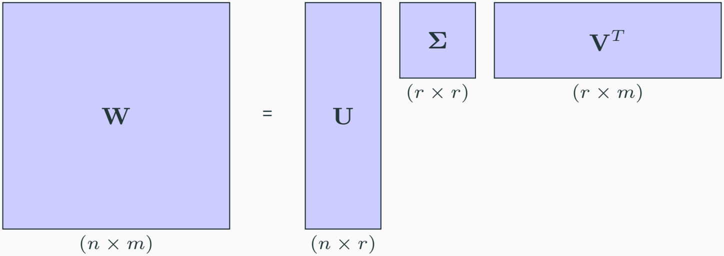
Singular value decomposition

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, $\mathbf{V}\mathbf{V}^T = \mathbf{I}_r$)
- $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ diagonal

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



SVD in greater detail

Let us consider the matrix $\mathbf{A} = \mathbf{W}^T\mathbf{W} \in \mathbb{R}^{m \times m}$. Observe that

- by definition, $\mathbf{A}$ has the same rank of $\mathbf{W}$, that is r
- $\mathbf{A}$ is symmetric: in fact, $a_{ij} = \mathbf{w}_i^T\mathbf{w}_j$ by definition, where $\mathbf{w}_k$ is the k-th column of $\mathbf{W}$; by the commutativity of vector product, $a_{ij} = \mathbf{w}_i^T\mathbf{w}_j = \mathbf{w}_j^T\mathbf{w}_i = a_{ji}$
- $\mathbf{A}$ is semidefinite positive, that is $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for all non null $\mathbf{x} \in \mathbb{R}^m$: this derives from

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T(\mathbf{W}^T\mathbf{W})\mathbf{x} = (\mathbf{W}\mathbf{x})^T(\mathbf{W}\mathbf{x}) = ||\mathbf{W}\mathbf{x}||_2 \geq 0$$

All eigenvalues of $\mathbf{A}$ are real. In fact,

- let $\lambda \in \mathbb{C}$ be an eigenvalue of $\mathbf{A}$, and let $\mathbf{v} \in \mathbb{C}^n$ be a corresponding eigenvector: then, $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ and $\overline{\mathbf{v}}^T\mathbf{A}\mathbf{v} = \overline{\mathbf{v}}^T\lambda\mathbf{v} = \lambda\overline{\mathbf{v}}^T\mathbf{v}$
- observe that, in general, it must also be that the complex conjugates $\overline{\lambda}$ and $\overline{\mathbf{v}}$ are themselves an eigenvalue-eigenvector pair for $\mathbf{A}$: then, $\mathbf{A}\overline{\mathbf{v}} = \overline{\lambda}\overline{\mathbf{v}}$. Since $\overline{\lambda}\overline{\mathbf{v}}^T = (\overline{\lambda}\overline{\mathbf{v}})^T = (\mathbf{A}\overline{\mathbf{v}})^T = \overline{\mathbf{v}}^T\mathbf{A}^T = \overline{\mathbf{v}}^T\mathbf{A}$ by the simmetry of $\mathbf{A}$, it derives $\overline{\mathbf{v}}^T\mathbf{A}\mathbf{v} = \overline{\lambda}\overline{\mathbf{v}}^T\mathbf{v}$
- as a consequence, $\overline{\lambda}\overline{\mathbf{v}}^T\mathbf{v} = \lambda\overline{\mathbf{v}}^T\mathbf{v}$, that is $\overline{\lambda}||\mathbf{v}||^2 = \lambda||\mathbf{v}||^2$
- since $\mathbf{v} \neq \mathbf{0}$ (being an eigenvector), it must be $\overline{\lambda} = \lambda$, hence $\lambda \in \mathbb{R}$

The eigenvectors of $\mathbf{A}$ corresponding to different eigenvalues are orthogonal

- Let $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{C}^n$ be two eigenvectors, with corresponding distinct eigenvalues λ_1, λ_2
- then, by the simmetry of $\mathbf{A}$, $\lambda_1(\mathbf{v}_1^T\mathbf{v}_2) = (\lambda_1\mathbf{v}_1)^T\mathbf{v}_2 = (\mathbf{A}\mathbf{v}_1)^T\mathbf{v}_2 = \mathbf{v}_1^T\mathbf{A}^T\mathbf{v}2 = \mathbf{v}_1^T\mathbf{A}\mathbf{v}_2 = \mathbf{v}_1^T\lambda_2\mathbf{v}_2 = \lambda_2(\mathbf{v}_1^T\mathbf{v}_2)$
- as a consequence, $(\lambda_1 - \lambda_2)\mathbf{v}_1^T\mathbf{v}_2 = 0$
- since $\lambda_1 \neq \lambda_2$, it must be $\mathbf{v}_1^T\mathbf{v}_2 = 0$, that is $\mathbf{v}_1, \mathbf{v}_2$ must be orthogonal

If an eigenvalue λ' has multiplicity $m > 1$, it is always possible to find a set of m orthonormal eigenvectors of λ'.

As a result, there exists a set of eigenvectors of $\mathbf{A}$ which provides an orthornormal base.

All eigenvalues of a $\mathbf{A}$ are greater than zero.

- $\mathbf{A}$ is real and symmetric, then for each eigenvalue λ it must be $\lambda \in \mathbb{R}$ and there must exist an eigenvector $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{Av} = \lambda\mathbf{v}$
- as a consequence, $\mathbf{v}^T(\mathbf{Av}) = \lambda\mathbf{v}^T\mathbf{v}$ and

$$\lambda = \frac{\mathbf{v}^T\mathbf{Av}}{\mathbf{v}^T\mathbf{v}} = \frac{\mathbf{v}^T\mathbf{Av}}{||\mathbf{v}||^2}$$

- $||\mathbf{v}||^2 > 0$ since $\mathbf{v}$ is an eigenvector and, since $\mathbf{A}$ is semidefinite positive, $\mathbf{v}^T\mathbf{Av} \geq 0$
- as a consequence, $\lambda \geq 0$

Overall,

- $\mathbf{A} = \mathbf{W}^T\mathbf{W}$ has r real and positive eigenvalues $\lambda_1, \dots, \lambda_r$
- the corresponding eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_r$ are orthonormal
- $\mathbf{A}\mathbf{v}_i = (\mathbf{W}^T\mathbf{W})\mathbf{v}_i = \lambda_i \mathbf{v}_i$, $i = 1, \dots, r$

Let us define r singular values

$$\sigma_i = \sqrt{\lambda_i} \qquad i = 1, \dots, r$$

and let us also consider the set of vectors

$$\mathbf{u}_i = \frac{1}{\sigma_i}\mathbf{W}\mathbf{v}_i \qquad i = 1, \dots, r$$

SVD in greater detail

- Observe that $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are orthogonal, in fact:

$$\mathbf{u}_i^T \mathbf{u}_j = \left(\frac{1}{\sigma_i}\mathbf{W}\mathbf{v}_i\right)^T \left(\frac{1}{\sigma_j}\mathbf{W}\mathbf{v}_j\right) = \frac{1}{\sigma_i \sigma_j}\mathbf{v}_i^T \mathbf{W}^T \mathbf{W}\mathbf{v}_j = \frac{1}{\sigma_i \sigma_j}\mathbf{v}_i^T (\lambda_j \mathbf{v}_j) = \frac{\sigma_j}{\sigma_i}\mathbf{v}_i^T$$

  Hence, $\mathbf{u}_i^T \mathbf{u}_j \neq 0$ iff $\mathbf{v}_i^T \mathbf{v}_j \neq 0$, that is iff $i \neq j$.

- Moreover, $\mathbf{u}_1, \ldots, \mathbf{u}_r$ have unitary norm, in fact:

$$\begin{aligned} ||\mathbf{u}_i||^2 &= \left|\left|\frac{1}{\sigma_i}\mathbf{W}\mathbf{v}_i\right|\right|^2 = \frac{1}{\lambda_i}(\mathbf{W}\mathbf{v}_i)^T(\mathbf{W}\mathbf{v}_i) = \frac{1}{\lambda_i}\mathbf{v}_i^T(\mathbf{W}^T\mathbf{W}\mathbf{v}_i) \\ &= \frac{1}{\lambda_i}\mathbf{v}_i^T(\lambda_i \mathbf{v}_i) = \frac{1}{\lambda_i}\lambda_i(\mathbf{v}_i^T\mathbf{v}_i) = 1 \end{aligned}$$

SVD in greater detail

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

$$\mathbf{V} = \begin{bmatrix} | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_r \\ | & | & & | \end{bmatrix}$$

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

$$\mathbf{U} = \begin{bmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_r \\ | & | & & | \end{bmatrix}$$

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

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

It is easy to verify that

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

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

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

$$\mathbf{W} = \begin{bmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_r \\ | & | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix} \begin{bmatrix} \text{--} & \mathbf{v}_1 & \text{--} \\ \text{--} & \mathbf{v}_2 & \text{--} \\ & \vdots & \\ \text{--} & \mathbf{v}_r & \text{--} \end{bmatrix}$$

PCA and SVD

PCA and SVD

- Given

$$\mathbf{X} = \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \\ | & | & & | \end{bmatrix}$$

- the mean of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is

$$\mathbf{m} = \frac{1}{n} \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \frac{1}{n}\mathbf{X}\mathbf{1}$$

- let $\tilde{\mathbf{X}}$ be the set of such vectors translated to have zero mean:

$$\tilde{\mathbf{X}} = \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \\ | & | & & | \end{bmatrix} - \begin{bmatrix} | & | & & | \\ \mathbf{m} & \mathbf{m} & \cdots & \mathbf{m} \\ | & | & & | \end{bmatrix} = \mathbf{X} - \mathbf{m}\mathbf{1}^T$$

$$= \mathbf{X} - \frac{1}{n}\mathbf{X}\mathbf{1}\mathbf{1}^T = \mathbf{X}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)$$

The correlation matrix of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is defined as:

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

where $\tilde{\mathbf{x}}_i$ is the i-th column of $\tilde{\mathbf{X}}$.

That is,

$$\mathbf{S} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$$

$\tilde{\mathbf{X}}$ has dimension $n \times d$: assuming $n > d$, we may consider its SVD

$$\tilde{\mathbf{X}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$$

where $\mathbf{U}\mathbf{U}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\boldsymbol{\Sigma}$ is a diagonal matrix.

By the properties of SVD, items on the diagonal of $\mathbf{\Sigma}$ are the eigenvalues of $\mathbf{S}$ and columns of $\mathbf{V}$ are the corresponding eigenvectors.

In summary:

- To perform a PCA on $\mathbf{X}$, it is sufficient to compute the SVD of matrix

$$\tilde{\mathbf{X}} = \mathbf{X}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

- The principal components of $\mathbf{X}$ are the columns of $\mathbf{V}$, with corresponding eigenvalues given by the diagonal elements of $\mathbf{\Sigma}^2$.

Latent semantic analysis

Definitions

Many models in text processing refer to co-occurrence data

Given two sets $\mathbf{V}, \mathbf{D}$ (for example, a set of terms and a collection of documents) a sequence of observations $\mathbf{W} = \{(w_1, d_1), \ldots, (w_N, d_N)\}$ is considered, with $w_i \in \mathbf{V}, d_i \in \mathbf{D}$ (for example, these are occurrences of terms in documents.

Fundamental hypotheses

The Latent Semantic Analysis (LSA) approach is based on the following three hypotheses:

- it is possible to derive semantic information from the matrix of occurrences of terms in documents
- the reduction of dimensionality is a key aspect of this derivation
- terms and documents can be modeled as points (vectors) in a euclidean space

Context

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

Idea

1. A document d_i can be seen as a multiset of N_i terms in $\mathbf{V}$ (bag of words hypotheses)
2. There exists a correspondance between $\mathbf{V}$ and $\mathbf{D}$, and a vector space $\mathcal{S}$.Each term t_i has an associated vector $\mathbf{u}_i$, also, to each document d_j a vector $\mathbf{v}_j$ in $\mathcal{S}$ is associated

Occurrence matrix

Let us define the 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)
- Documents correspond to column vector (size V)

Problem

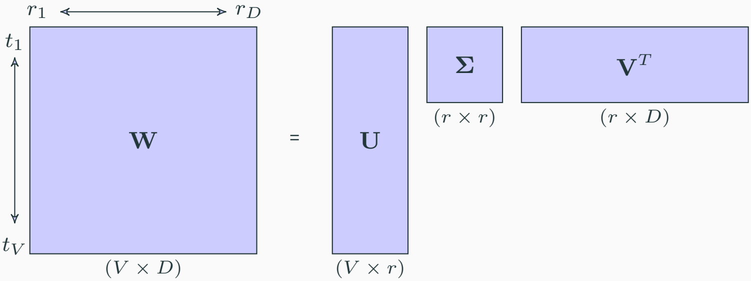
1. The values V, D are usually quite large
2. Vectors corresponding to t_i and d_j are very sparse
3. Terms and documents are modeled as vectors defined on different spaces ($\mathbb{R}^D$ and $\mathbb{R}^V$, respectively)

Exploit singular value decomposition.

In short

- The occurrence matrix $\mathbf{W}$ is decomposed in the product of three matrices.
- A term matrix $\mathbf{U}$, with rows corresponding to terms: each term spans over r dimensions
- A document matrix $\mathbf{V}^T$, with columns corresponding to documents: each document spans over r dimensions
- The matrix of singular values $\mathbf{\Sigma}$, whose diagonal elements provide a measure of the relevance of the corresponding dimensions

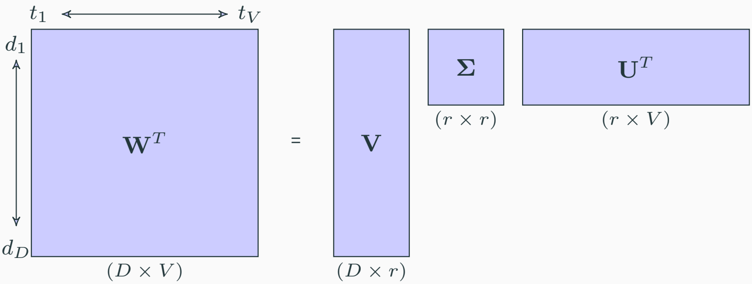
Use of SVD



Effect

Rows of $\mathbf{W}$ (terms) are projected onto an r-dimensional subspace of $\mathbb{R}^D$. The columns of $\mathbf{V}^T$ provide a basis of such subspace, hence each term is associated to a linear combination of these columns.

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.



Effect

Rows of $\mathbf{W}^T$ (documents) are projected onto an r-dimensional subspace of $\mathbb{R}^V$. The columns of $\mathbf{U}^T$ provide a basis of such subspace, hence each term is associated to a linear combination of these columns.

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.

Dimensionality reduction

The dimension d of the projection subspace can be predefined to be less than the rank of $\mathbf{W}$. In this case,

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

Approximation

The following property holds:

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

That is $\overline{\mathbf{W}}$ is the best approximation of $\mathbf{W}$among all matrices of rank d wrt the Frobenius norm

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

Effect

SVD provides a tranformation of two discrete vector spaces $\mathcal{V} \in \mathbb{Z}^D$ and $\mathcal{D} \in \mathbb{Z}^V$ into a unique continuous vector space with lower dimension $\mathcal{T} \in \mathbb{R}^d$.

The dimension of $\mathcal{T}$ is at most equal to the (unknown) rank of $\mathbf{W}$, and is determined by the acceptable amount of distortion induced by the projection

Interpretation

$\overline{\mathbf{W}}$ keeps most of the associations between terms and documents in $\mathbf{W}$: it only does not take into account the least significant relations

- Each term is now seen as a linear combination of unknown ``topics'': terms with similar projections tend to appear in the same documents (or in documents semantically similar, in which similar terms appear)
- Each document is also seen as a linear combination of the same unknown topics: documents with similar projections tend to contain the same terms (or terms semantically similar, which appear in similar documents)

Co-occurrences

- $\mathbf{W}\mathbf{W}^T \in \mathbb{Z}^{V \times V}$ provides co-occurrences of terms in $\mathbf{V}$ (number of documents in which both terms appear)
- $\mathbf{W}^T\mathbf{W} \in \mathbb{Z}^{D \times D}$ provides co-occurrences of documents in $\mathbf{D}$ (number of terms appearing in both documents)

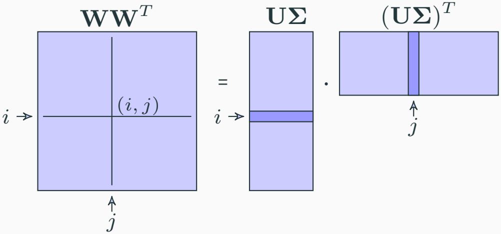
SVD and co-occurrence matrix

By applying SVD,

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

and

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

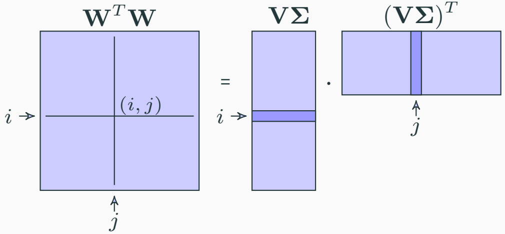


Proximity of terms

A reasonable measure of the proximity between two terms t_i, t_j is the number of documents in which they co-occur, that is the value of element (i, j) in $\mathbf{W}\mathbf{W}^T$. This corresponds to the dot product of vectors $\mathbf{u}_i\sigma_i$ (i-th row of $\mathbf{U}\mathbf{\Sigma}$) and $\mathbf{u}_j\sigma_j$ (j-th row of $\mathbf{U}\mathbf{\Sigma}$).

In particular, we may define

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



A reasonable measure of the proximity between two terms d_i, d_j is the number of terms co-occurring in then, that is the value of element (i, j) in $\mathbf{W}^T\mathbf{W}$. This corresponds to the dot product of vectors $\mathbf{v}_i\sigma_i$ (i-th row of $\mathbf{V\Sigma}$) and $\mathbf{v}_j\sigma_j$ (j-th row of $\mathbf{V\Sigma}$).

In particular, we may define

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

Proximity of a document to a topic

Objective

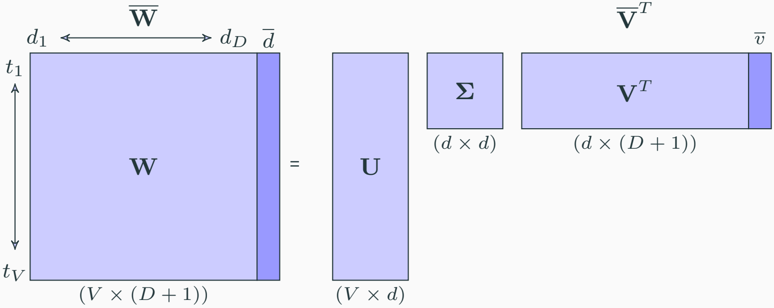
Determine, given a document, the topic (in a predefined collection) which is more related to its content.

Approach

Construction of a vector of weights associated to the topic: can be seen as a further document $\overline{d}$ (topic template)

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

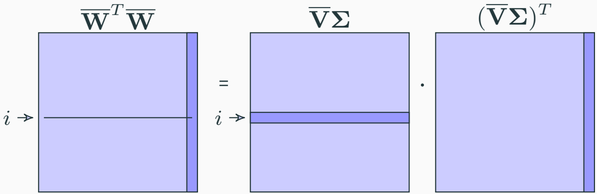
Proximity of a document to a topic



Effect

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

Proximity of a document to a topic



A reasonable measure of the proximity between a document d_i and a topic $\overline{d}$ corresponds to the dot product of vectors $\mathbf{v}_i\sigma_i$ (i-th row of $\mathbf{V\Sigma}$) and $\overline{\mathbf{v}}$ ($D+1$-th row of $\mathbf{V\Sigma}$).

In particular, we may define

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