Machine Learning for Networks: Dimensionality reduction

Andrea Araldo

February 6, 2023
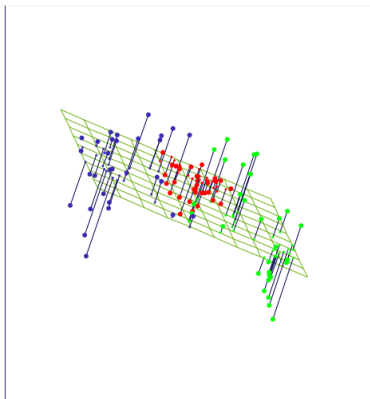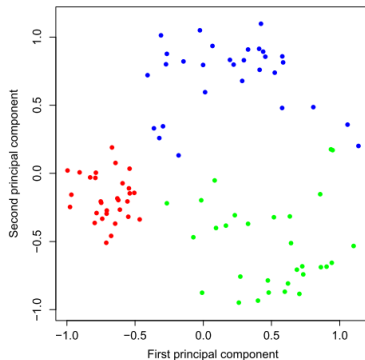
From Source:https://www.alimentipedia.it/come-si-fa-il-formaggio.html

Why dimensionality reduction

From [3]

Main idea:

- Sample = Vector of a vectorial space
  (each feature is a dimension)
- Find the "best" space for your data
- Keep few "most important" dimensions only.
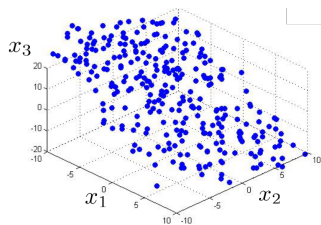- Project the samples in the new reduced space.

Advantages:

- Easier to visualize
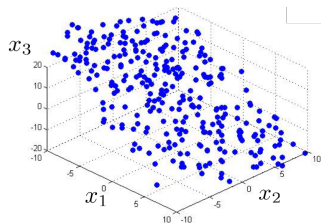- Smaller dataset
  ⇒ Faster model runs

- Which 2D space would you choose?
- Any mathematical argument for your choice?

From [5]

The "best space"

x_3

From [5]

- Which 2D space would you choose?
- Any mathematical argument for your choice?
- Hyperplane capturing most of the variation of data
- *Most important dimension*:
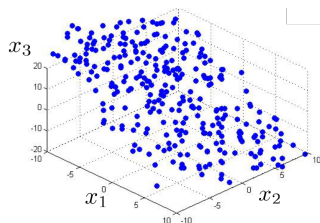 the one along which data have the largest variance.

The "best space"

From [5]

- Which 2D space would you choose?
- Any mathematical argument for your choice?
- Hyperplane capturing most of the variation of data
- *Most important dimension*:
  the one along which data have the largest variance.
- "Best space":
  Dimensions are ordered from the most to the least important.
- **Singular Value Decomposition** (**SVD**)
  finds this space!

- Change of basis
- Singular Value Decomposition
- Dimensionality Reduction
- Application to supervised learning
- Application to anomaly detection
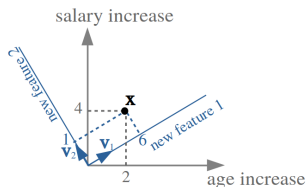
Section 1

Change of basis

Change of basis

Consider data matrix $\mathbf{X}$, already standardized. Sample $\mathbf{x}^{(i)}$ is a vector of space $\mathbb{R}^N$. The dimensions of such a space are feature $1, \ldots,$ feature $j, \ldots,$ feature N.

Take any set of N orthonormal vectors $\mathbf{v}_1, \ldots, \mathbf{v}_N$:

$$\mathbf{v}_z^T \cdot \mathbf{v}_z = 1;$$
$$\mathbf{v}_z^T \cdot \mathbf{v}_{z'} = 0 \quad \text{for } z' \neq z.$$

They are a new basis of space $\mathbb{R}^N$. Each new feature is a combination of the original features.

The projection of $\mathbf{x}^{(i)}$ over the j-th new feature is

$$x_j^{\widetilde{(i)}} = {\mathbf{x}^{(i)}}^T \cdot \mathbf{v}_j.$$

The projection of $\mathbf{x}^{(i)}$ over the new space is

$$\widetilde{\mathbf{x}^{(i)}}^T = {\mathbf{x}^{(i)}}^T \cdot \underbrace{[\mathbf{v}_1 | \ldots | \mathbf{v}_N]}_{\mathbf{V}} = {\mathbf{x}^{(i)}}^T \cdot \mathbf{V}.$$

The projections of all samples on the j-th new feature (the j-th column of the transformed dataset) is

$$\mathbf{X} \cdot \mathbf{v}_j.$$

The projection on all the new features is:

$$\mathbf{X} \cdot \mathbf{V}.$$

Properties of change of basis

Proposition

If all features of $\mathbf{X}$ have zero-mean, after changing bases, the new features have also zero-mean

Proof
We know that

$$(1, \ldots, 1) \cdot \mathbf{X} = (0, \ldots, 0).$$

Let us compute the mean of the new features:

$$(1, \ldots, 1) \cdot \mathbf{X} \cdot \mathbf{v}_j = (0, \ldots, 0).$$

Proposition

The total variance of $\mathbf{X}$ does not change when changing basis.

Proof
Denote with $\mathbf{x}_j$ the j-th column of the original dataset. The total variance is

$$\sum_{j=1}^{N} \text{Var}_j = \sum_{j=1}^{N} \mathbf{x}_j^T \cdot \mathbf{x}_j = \sum_{j=1}^{N} \sum_{i=1}^{M} x_j^{(i)} \cdot x_j^{(i)}$$
$$= \sum_{i=1}^{M} \sum_{j=1}^{N} x_j^{(i)} \cdot x_j^{(i)} = \sum_{i=1}^{M} \mathbf{x}^{(i)^T} \cdot \mathbf{x}^{(i)}$$

Similarly, the total variance of the transformed dataset is:

$$\sum_{j=1}^{N} \text{Var}'_j = \sum_{i=1}^{M} \mathbf{x}^{\widetilde{(i)}^T} \cdot \mathbf{x}^{\widetilde{(i)}} = \sum_{i=1}^{M} \mathbf{x}^{(i)^T} \cdot \mathbf{V} \cdot \mathbf{V}^T \mathbf{x}^{(i)}$$
$$\underbrace{=}_{\text{orthonormal}} \sum_{i=1}^{M} \mathbf{x}^{(i)^T} \cdot \mathbf{x}^{(i)}$$
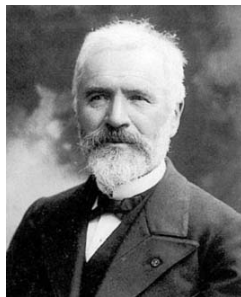
Section 2

Singular Value Decomposition

An old history

From [2]

The Singular Value Decomposition was discovered and developed independently by a number of mathematicians. Eugenio Beltrami and Camille Jordan were the first to do so, in 1873 and 1874, respectively

Eugenio Beltrami
Prof. at Università di Bologna.
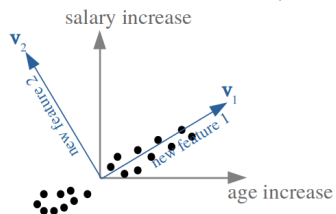

Camille Jordan
Prof. at École Polytechnique.

With the following dataset, can
you find the "best" basis $\mathbf{v}_1, \mathbf{v}_2$?

Best new feature space

With the following dataset, can you find the "best" basis $\mathbf{v}_1, \mathbf{v}_2$?
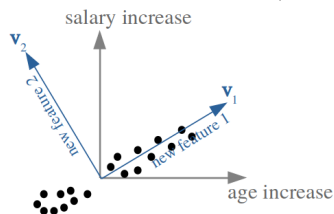
Best new feature space

Intuitive algorithm:

Find $\mathbf{v}_1$ such that the projection captures most of the variance:

$$\mathbf{v}_1 = \arg\max_{\mathbf{v}} \mathbf{v}^T \cdot \mathbf{X}^T\mathbf{X} \cdot \mathbf{v}$$

$$\text{s.t. } \mathbf{v}^T \cdot \mathbf{v} = 1$$

With the following dataset, can you find the "best" basis $\mathbf{v}_1, \mathbf{v}_2$?

Best new feature space

Intuitive algorithm:

Find $\mathbf{v}_1$ such that the projection captures most of the variance:

$$\mathbf{v}_1 = \arg\max_{\mathbf{v}} \mathbf{v}^T \cdot \mathbf{X}^T\mathbf{X} \cdot \mathbf{v}$$
$$\text{s.t. } \mathbf{v}^T \cdot \mathbf{v} = 1$$

With the following dataset, can you find the "best" basis $\mathbf{v}_1, \mathbf{v}_2$?

Then, find $\mathbf{v}_2$:

$$\mathbf{v}_2 = \arg\max_{\mathbf{v}} \mathbf{v}^T \cdot \mathbf{X}^T\mathbf{X} \cdot \mathbf{v}$$
$$\text{s.t. } \mathbf{v}^T \cdot \mathbf{v} = 1$$
$$\mathbf{v}^T \cdot \mathbf{v}_1 = 0$$

Best new feature space

Intuitive algorithm:

Find $\mathbf{v}_1$ such that the projection captures most of the variance:

$$\mathbf{v}_1 = \arg\max_{\mathbf{v}} \mathbf{v}^T \cdot \mathbf{X}^T\mathbf{X} \cdot \mathbf{v}$$
$$\text{s.t. } \mathbf{v}^T \cdot \mathbf{v} = 1$$

With the following dataset, can you find the "best" basis $\mathbf{v}_1, \mathbf{v}_2$?
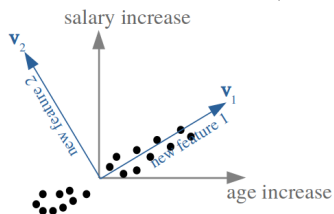
Then, find $\mathbf{v}_2$:

$$\mathbf{v}_2 = \arg\max_{\mathbf{v}} \mathbf{v}^T \cdot \mathbf{X}^T\mathbf{X} \cdot \mathbf{v}$$
$$\text{s.t. } \mathbf{v}^T \cdot \mathbf{v} = 1$$
$$\mathbf{v}^T \cdot \mathbf{v}_1 = 0$$

Then, find $\mathbf{v}_3$:

$$\mathbf{v}_3 = \arg\max_{\mathbf{v}} \mathbf{v}^T \cdot \mathbf{X}^T\mathbf{X} \cdot \mathbf{v}$$
$$\text{s.t. } \mathbf{v}^T \cdot \mathbf{v} = 1$$
$$\mathbf{v}^T \cdot \mathbf{v}_1 = 0$$
$$\mathbf{v}^T \cdot \mathbf{v}_2 = 0$$

Theorem ([4])

Any matrix $\mathbf{X}$ *can be decomposed as follows*

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

where $\mathbf{\Sigma}$ *is a diagonal matrix of dimension* $N \times N$ *and* $N = \min(N, M)$.
The elements on the diagonal of $\mathbf{\Sigma}$ *are the singular values of* $\mathbf{X}$ *and are ordered:* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N \geq 0$.
Denoting with $\mathbf{I}_N$ *the identity matrix of dimension* N, *we also have*

$$\mathbf{V}^T \cdot \mathbf{V} = \mathbf{V} \cdot \mathbf{V}^T = \mathbf{I}_N \text{ and } \mathbf{U}^T \cdot \mathbf{U} = \mathbf{I}_N.$$

The columns $\mathbf{v}_1, \ldots, \mathbf{v}_N$ *are thus an orthonormal basis of* $\mathbb{R}^N$.

Variance of the new features
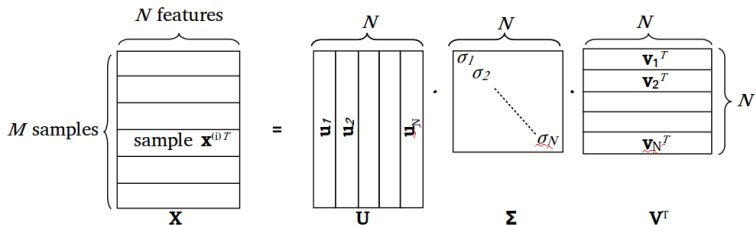
Corollary

σ_j^2 *is the variance of the j-th new feature.*
Therefore, the new features are ordered from the one with the most variance to the least.
Moreover, the total variance of the dataset is $\sum_{j=1}^{N} \sigma_j^2$ (independent of the basis).

Proof.
The variance of the *j*-th new feature is:

$$\begin{aligned}
(\mathbf{X} \cdot \mathbf{v}_j)^T \cdot (\mathbf{X} \cdot \mathbf{v}_j) &= \mathbf{v}_j^T \cdot \mathbf{X}^T \mathbf{X} \cdot \mathbf{v}_j \\
&= \mathbf{v}_j^T \cdot \mathbf{V} \mathbf{\Sigma} \mathbf{U}^T \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T \cdot \mathbf{v}_j \\
&= (\mathbf{v}_j^T \cdot \mathbf{V}) \mathbf{\Sigma}^2 (\mathbf{V}^T \cdot \mathbf{v}_j) \\
\text{(thanks to orthonormality of the columns of } \mathbf{V}) \quad &= \mathbf{e}_j^T \cdot \mathbf{\Sigma}^2 \cdot \mathbf{e}_j \\
&= \sigma_j^2,
\end{aligned}$$

where $\mathbf{e}_j$ is a vector having a 1 in the *j*-th position and 0 everywhere else.

Section 3

Dimensionality reduction

Low-rank approximation

After applying SVD, we can just consider the first r new features (or components):

$\mathbf{X}_{(r)}$ is a *low-rank* approximation of $\mathbf{X}$.

Each component j "captures" (has) a variance $\frac{\sigma_j^2}{\sum_{j'=1}^r \sigma_{j'}^2}$.

By taking only the first r features, we capture a fraction $\frac{\sum_{j=1}^r \sigma_j^2}{\sum_{j'=1}^N \sigma_{j'}^2}$

If you do not scale:

- The features with higher magnitude will concentrate most of the variance
- They will automatically be more present in the first principal components
  - Not beacause they are the most important, but just because of their magnitude
- All the theory only holds if features have zero-mean (see definition of variance).

Advantages of dimensionality reduction

- Scalability: work with $r < N$ features.
- Stability: we only preserve the "important" information (by keeping the components that capture most of the variance) and remove the "details".
- Supervised learning and dimensionality reduction
  - Given a dataset (already scaled) $\mathbf{X}$ with true lables $\mathbf{y}$:
  - Partition the dataset: $(\mathbf{X}^{\text{train}}, \mathbf{y}^{\text{train}})$ and $(\mathbf{X}^{\text{test}}, \mathbf{y}^{\text{text}})$
  - Apply SVD: $\mathbf{X}^{\text{train}} = \mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V}^T$
  - Keep the r most important components: $\tilde{\mathbf{X}}^{\text{train}} = \mathbf{X} \cdot \mathbf{V}_{(r)}$
  - Train your model $h(\cdot)$ on $(\tilde{\mathbf{X}}^{\text{train}}, \mathbf{y}^{\text{train}})$
  - For a new test sample $\mathbf{x}$:
    - Project it on the first r components: $\tilde{\mathbf{x}}^T = \mathbf{x} \cdot \mathbf{V}_{(r)}$.
    - The prediction is $\hat{y} = h(\tilde{\mathbf{x}})$.
  - By doing so, your predictions are usually more stable (they do not overfit on details)

Go to notebook 07.dimensionality-reductuion/b.svd-and-regression.ipynb

Directly in the code

Let's code ...

Go to notebook
07.dimensionality-reductuion/a.singular-value-decomposition.ipynb

- Change of basis
- Singular Value Decomposition
- Dimensionality Reduction
- Application to supervised learning
- Application to anomaly detection

Section 4

Backup slides (you can ignore)

Component matrices

Theorem ([1])

The sample matrix can be written as $\mathbf{X} = \sum_{z=0}^{N} \sigma_z \cdot \mathbf{u}_z \cdot \mathbf{v}_z^T$.
The sample matrix is thus a linear combination of component matrices
$\mathbf{u}_i \cdot \mathbf{v}_i^T$ *with weights σ_i.*

Proof:

Matrix $\mathbf{u}_i \cdot \mathbf{v}_i^T$ contributes to the sample matrix for a fraction $\frac{\sigma_i}{\sum_{j=1}^{r} \sigma_j}$

The first component matrices are the most important.

[1] Singular Value Decomposition.
In *STAT 555 - Penn State University - Lecture Notes*, chapter 16.1.

[2] Samuel Chowning.
The singular value decomposition, 2020.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman.
The Elements of Statistical Learning, volume 1.
Springer, 2nd edition, 2009.
URL: http://www.springerlink.com/index/10.1007/b94608,
arXiv:1010.3003, doi:10.1007/b94608.

[4] Kevin P. Murphy.
Machine learning: A Probabilistic Perspective, chapter 12.2.3.
MIT Press, 2012.

[5] Fereshteh Sadeghi.
Dimensionality Reduction.
Lecture notes - CSEP 546.
URL: https://courses.cs.washington.edu/courses/csep546/16sp/slides/PCA_csep546.pdf.