PCA, K-Means

Milan Straka

December 13, 2021
Unsupervised Machine Learning
Let $A \in \mathbb{C}^{N \times N}$ be an $N \times N$ matrix.

- A vector $v \in \mathbb{C}^N$ is a (right) eigenvector, if there exists an eigenvalue $\lambda \in \mathbb{C}$, such that
  \[ Av = \lambda v. \]

- If $A \in \mathbb{R}^{N \times N}$ is a real symmetric matrix, then it has $N$ real eigenvalues and $N$ real eigenvectors, which can be chosen to be orthonormal, and we can express $A$ using the eigenvalue decomposition
  \[ A = V^T \Lambda V, \]

where:
- $V$ is a matrix, whose columns are the eigenvectors $v_1, v_2, \ldots, v_N$;
- $\Lambda$ is a diagonal matrix with the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$ on the diagonal.
Let $A \in \mathbb{R}^{N \times N}$ be a real symmetric matrix. Then if for all $x \neq 0$:

- $x^T Ax > 0$, the matrix is called **positive definite**.
  
  Note that this condition is equivalent to all eigenvalues being **positive**.

- $x^T Ax \geq 0$, the matrix is called **positive semi-definite**.
  
  This condition is equivalent to all eigenvalues being **non-negative**.

- $x^T Ax < 0$, the matrix is called **negative definite**.
  
  This condition is equivalent to all eigenvalues being **negative**.

- $x^T Ax \leq 0$, the matrix is called **negative semi-definite**.
  
  This condition is equivalent to all eigenvalues being **non-positive**.

Note that we can compute a “square root” of a positive (semi-)definite matrix, because if $A = V^T \Lambda V$, then for $\Lambda^{1/2}V$ we get

$$
(\Lambda^{1/2}V)^T \Lambda^{1/2} V = V^T \Lambda^{1/2} \Lambda^{1/2} V = V^T \Lambda V = A.
$$
Principal Component Analysis

The **principal component analysis**, **PCA**, is a linear transformation used for

- dimensionality reduction,
- feature extraction,
- whitening,
- data visualization.

To motivate the dimensionality reduction, consider a dataset consisting of a randomly translated and rotated image.

Every member of the dataset can be described just by three quantities – horizontal and vertical offsets and a rotation. We usually say that the data *lie on a manifold of dimension three*. 

*Figure 12.1 of Pattern Recognition and Machine Learning.*
We start by defining the PCA in two ways.

**Maximum Variance Formulation**

Given data \( x_1, \ldots, x_N \) with \( x_i \in \mathbb{R}^D \), the goal is to project them to a space with dimensionality \( M < D \), so that the variance of their projection is maximal.

We start by considering a projection to one-dimensional space. Such a projection is defined by a vector \( u_1 \), and because only the direction of \( u_1 \) matters, we assume that \( u_1^T u_1 = 1 \).

The projection of \( x_i \) to \( u_1 \) is given by \( (u_1^T x_i)u_1 \), because the vectors \( u_1 \) and \( x_i - (u_1^T x_i)u_1 \) are orthogonal:

\[
u_1^T (x_i - (u_1^T x_i)u_1) = u_1^T x_i - (u_1^T x_i)u_1^T u_1 = 0.
\]
Principal Component Analysis

We therefore use $\mathbf{u}_1^T \mathbf{x}_i$ as the projection of $\mathbf{x}_i$. If we define $\bar{x} = \sum_i \mathbf{x}_i / N$, the mean of the projected data is $\mathbf{u}_1^T \bar{x}$ and the variance is given by

$$ \frac{1}{N} \sum_{i=1}^{N} (\mathbf{u}_1^T \mathbf{x}_i - \mathbf{u}_1^T \bar{x})^2 = \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1, $$

where $\mathbf{S}$ is the data covariance matrix defined as

$$ \mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{x}) (\mathbf{x}_i - \bar{x})^T. $$

We can write the data covariance matrix in matrix form as

$$ \mathbf{S} = \frac{1}{N} (\mathbf{X} - \bar{x})^T (\mathbf{X} - \bar{x}). $$

If the original data is centered (it has zero mean), then $\mathbf{S} = \frac{1}{N} \mathbf{X}^T \mathbf{X}$, which we have already encountered.
Principal Component Analysis

To maximize $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$, we need to include the constraint $\mathbf{u}_1^T \mathbf{u}_1$ by introducing a Lagrange multiplier $\lambda_1$ for the constraint $\mathbf{u}_1^T \mathbf{u}_1 - 1 = 0$ and then maximizing the Lagrangian

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

By computing a derivative with respect to $\mathbf{u}_1$, we get

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

Therefore, $\mathbf{u}_1$ must be an eigenvector of $\mathbf{S}$ corresponding to eigenvalue $\lambda_1$.

Because the value to maximize $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$ is then $\mathbf{u}_1^T \lambda_1 \mathbf{u}_1 = \lambda_1 \mathbf{u}_1^T \mathbf{u}_1 = \lambda_1$, the maximum will be attained for eigenvector $\mathbf{u}_1$ corresponding to the largest eigenvalue $\lambda_1$.

The eigenvector $\mathbf{u}_1$ is known as the first principal component.

For a given $\mathbf{M}$, the principal components are eigenvectors corresponding to $\mathbf{M}$ largest eigenvalues, and maximize the variance of the projected data.
Principal Component Analysis

Minimum Error Formulation

Assume $\mathbf{u}_1, \ldots, \mathbf{u}_D$ is some orthonormal set of vectors, therefore, $\mathbf{u}_i^T \mathbf{u}_j = [i == j]$.

Every $\mathbf{x}_i$ can be then expressed using this basis as

$$\mathbf{x}_i = \sum_j (\mathbf{x}_i^T \mathbf{u}_j) \mathbf{u}_j,$$

using a similar argument as the one we used to derive the orthogonal projection.

Because we want to eventually represent the data using $M$ dimensions, we will approximate the data by the first $M$ basis vectors:

$$\tilde{\mathbf{x}}_i = \sum_{j=1}^{M} z_{i,j} \mathbf{u}_j + \sum_{j=M+1}^{D} b_j \mathbf{u}_j.$$
Principal Component Analysis

We now choose the vectors $\mathbf{u}_j$, coordinates $z_{i,j}$ and biases $b_j$ to minimize the approximation error, which we measure as a loss

$$L = \frac{1}{N} \sum_{i=1}^{N} \| \mathbf{x}_i - \tilde{\mathbf{x}}_i \|^2.$$

To minimize the error, we compute the derivative of $L$ with respect to $z_{i,j}$ and $b_j$, and utilizing the orthogonality, we obtain

$$z_{i,j} = \mathbf{x}_i^T \mathbf{u}_j, \quad b_j = \tilde{\mathbf{x}}^T \mathbf{u}_j.$$

Therefore, we can rewrite the loss as

$$L = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=M+1}^{D} (\mathbf{x}_i^T \mathbf{u}_j - \tilde{\mathbf{x}}^T \mathbf{u}_j)^2 = \sum_{j=M+1}^{D} \mathbf{u}_j^T S \mathbf{u}_j.$$

Analogously, we can minimize $L$ by choosing the eigenvectors of $D - M$ smallest eigenvalues.
We can represent the data $\mathbf{x}_i$ by the approximations $\tilde{\mathbf{x}}_i$

$$\tilde{\mathbf{x}}_i = \sum_{j=1}^{M} (\mathbf{x}_i^T \mathbf{u}_j) \mathbf{u}_j + \sum_{j=M+1}^{D} (\bar{\mathbf{x}}^T \mathbf{u}_j) \mathbf{u}_j = \bar{\mathbf{x}} + \sum_{j=1}^{M} (\mathbf{x}_i^T \mathbf{u}_j - \bar{\mathbf{x}}^T \mathbf{u}_j) \mathbf{u}_j.$$
PCA Applications – Data Compression

Figure 12.4 of Pattern Recognition and Machine Learning.

\[ J = \sum_{i=M+1}^{D} \lambda_i \]

Figure 12.3 of Pattern Recognition and Machine Learning.

Mean  \( \lambda_1 = 3.4 \times 10^5 \)  \( \lambda_2 = 2.8 \times 10^5 \)  \( \lambda_3 = 2.4 \times 10^5 \)  \( \lambda_4 = 1.6 \times 10^5 \)
PCA Applications – Whitening aka Sphering

The PCA formula allows us to perform **whitening** aka **sphering**, which is a linear transformation of the given data, so that the resulting dataset has zero mean and an identity covariance matrix.

Notably, if $\mathbf{U}$ are the eigenvectors of $\mathbf{S}$ and $\mathbf{\Lambda}$ is the diagonal matrix of the corresponding eigenvalues (i.e., $\mathbf{SU} = \mathbf{U} \Lambda$), we can define the transformed data as

$$
\mathbf{z}_i \overset{\text{def}}{=} \mathbf{\Lambda}^{-1/2} \mathbf{U}^T (\mathbf{x}_i - \bar{\mathbf{x}}).
$$

Then, the mean of $\mathbf{z}_i$ is zero and the covariance is given by

$$
\frac{1}{N} \sum_i \mathbf{z}_i \mathbf{z}_i^T = \frac{1}{N} \sum_i \mathbf{\Lambda}^{-1/2} \mathbf{U}^T (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{U} \mathbf{\Lambda}^{-1/2}
$$

$$
= \mathbf{\Lambda}^{-1/2} \mathbf{U}^T \mathbf{SU} \mathbf{\Lambda}^{-1/2} = \mathbf{\Lambda}^{-1/2} \mathbf{\Lambda} \mathbf{\Lambda}^{-1/2} = \mathbf{I}.
$$
PCA Applications – Whitening or Sphering

Figure 12.6 of Pattern Recognition and Machine Learning.
Note that PCA does not have access to supervised labels, so it may not give a solution favorable for further classification.

Figure 12.7 of Pattern Recognition and Machine Learning.
It can be proven that if we construct a MLP autoencoder, which is a model trying to reconstruct input as close as possible, then even if the hidden layer uses non-linear activation, the solution to a MSE loss is a projection onto the $M$-dimensional subspace defined by the first $M$ principal components (but is not necessary orthonormal or orthogonal).
However, non-linear PCA can be achieved, if both the encoder and the decoder are non-linear.
There are two frequently used algorithms for performing PCA.

If we want to compute all (or many) principal components, we can compute directly the eigenvectors and eigenvalues of the covariance matrix.

We can even avoid computing the covariance matrix. If we instead compute the singular value decomposition of \((X - \bar{x}) = UDV^T\), it holds that

\[
(X - \bar{x})^T (X - \bar{x}) = VDU^T UDV^T = VD^2V^T.
\]

Therefore,

\[
(X - \bar{x})^T (X - \bar{x}) V = VD^2,
\]

which means that \(V\) are the eigenvectors of \((X - \bar{x})^T (X - \bar{x})\) and therefore of the data covariance matrix \(S\). The eigenvalues of \(S\) are the squares of the singular values of \((X - \bar{x})\) divided by \(N\).
If we want only the first (or several first) principal components, we might use the **power iteration algorithm**.

The power iteration algorithm can be used to find a **dominant** eigenvalue (an eigenvalue with absolute value strictly larger than absolute value of all other eigenvalues) and the corresponding eigenvector (it is used for example to compute PageRank). It works as follows:

**Input**: Real symmetric matrix $A$ with a dominant eigenvalue.

**Output**: The dominant eigenvalue $\lambda$ and the corresponding eigenvector $v$, with probability close to 1.

- Initialize $v$ randomly (for example each component from $U[-1, 1]$).
- Repeat until convergence (or for a fixed number of iterations):
  - $v \leftarrow Av$
  - $\lambda \leftarrow \|v\|$
  - $v \leftarrow v/\lambda$

If the algorithm converges, then $v = Av/\lambda$, so $v$ is an eigenvector with eigenvalue $\lambda$. 
In order to analyze the convergence, let \((\lambda_1, \lambda_2, \lambda_3, \ldots)\) be the eigenvalues of \(A\), in the descending order of absolute values, so \(|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots\), where the strict equality is the consequence of the dominant eigenvalue assumption.

If we express the vector \(v\) in the basis of the eigenvectors as \((a_1, a_2, a_3, \ldots)\), then \(Av/\lambda_1\) is in the basis of the eigenvectors:

\[
\frac{Av}{\lambda_1} = \left( \frac{\lambda_1}{\lambda_1} a_1, \frac{\lambda_2}{\lambda_1} a_2, \frac{\lambda_3}{\lambda_1} a_3, \ldots \right) = \left( a_1, \frac{\lambda_2}{\lambda_1} a_2, \frac{\lambda_3}{\lambda_1} a_3, \ldots \right).
\]

Therefore, all but the first coordinates decreased by at least a factor of \(|\lambda_2/\lambda_1|\).

If the initial \(v\) had a non-zero first coordinate \(a_1\) (which has probability very close to 1), then repeated multiplication with \(A\) will converge to the eigenvector corresponding to \(\lambda_1\).
After we get the largest eigenvalue $\lambda_1$ and its eigenvector $v_1$, we can modify the matrix $A$ to “remove the eigenvalue $\lambda_1$”. Considering $A - \lambda_1 v_1 v_1^T$:

- multiplying it by $v_1$ returns zero:

$$
(A - \lambda_1 v_1 v_1^T) v_1 = \lambda_1 v_1 - \lambda_1 v_1 v_1^T v_1 = 0,
$$

- multiplying it by other eigenvectors $v_i$ gives the same result as multiplying $A$:

$$
(A - \lambda_1 v_1 v_1^T) v_i = A v_i - \lambda_1 v_1 v_1^T v_i = A v_i.
$$

Therefore, $A - \lambda_1 v_1 v_1^T$ has the same set of eigenvectors and eigenvalues, except for $v_1, \lambda_1$. 


Computing PCA — The Power Iteration Algorithm

We are now ready to formulate the complete algorithm for computing the PCA.

**Input**: Matrix $X$, desired number of dimensions $M$.

- Compute the mean $\mu$ of the examples (the rows of $X$).
- Compute the covariance matrix $S \leftarrow \frac{1}{N} (X - \mu)^T (X - \mu)$.

for $i$ in $\{1, 2, \ldots, M\}$:

  - Initialize $v_i$ randomly.
  - Repeat until convergence (or for a fixed number of iterations):
    - $v_i \leftarrow Sv_i$
    - $\lambda_i \leftarrow \|v_i\|$
    - $v_i \leftarrow v_i / \lambda_i$

  - $S \leftarrow S - \lambda_i v_i v_i^T$

- Return $XV$, where the columns of $V$ are $v_1, v_2, \ldots, v_M$. 
Clustering is an unsupervised machine learning technique, which given input data tries to divide them into some number of groups, or clusters.

The number of clusters might be given in advance, or we should infer it.

When clustering documents, we usually normalize TF-IDF normalized so that each feature vector has length 1 (i.e., L2 normalization), because then

$$1 - \text{cosine similarity}(x, y) = \frac{1}{2} \|x - y\|^2.$$
K-Means Clustering

Let \( \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \) be a collection of \( N \) input examples, each being a \( D \)-dimensional vector \( \mathbf{x}_i \in \mathbb{R}^D \). Let \( K \), the number of target clusters, be given.

Let \( z_{i,k} \in \{0, 1\} \) be binary indicator variables describing whether an input example \( \mathbf{x}_i \) is assigned to cluster \( k \), and let each cluster be specified by a point \( \mathbf{\mu}_1, \ldots, \mathbf{\mu}_K \), usually called the cluster center.

Our objective function \( J \), which we aim to minimize, is

\[
J = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{i,k} \| \mathbf{x}_i - \mathbf{\mu}_k \|^2.
\]
K-Means Clustering

**Input:** Input points $x_1, \ldots, x_N$, number of clusters $K$.

- Initialize $\mu_1, \ldots, \mu_K$ as $K$ random input points.
- Repeat until convergence (or until patience runs out):
  - Compute the best possible $z_{i,k}$. It is easy to see that the smallest $J$ is achieved by
    $$z_{i,k} = \begin{cases} 
    1 & \text{if } k = \arg\min_j \|x_i - \mu_j\|^2, \\
    0 & \text{otherwise}.
    \end{cases}$$
  - Compute the best possible $\mu_k = \arg\min_{\mu} \sum_i z_{i,k}\|x_i - \mu\|^2$. By computing a derivative with respect to $\mu$, we get
    $$\mu_k = \frac{\sum_i z_{i,k}x_i}{\sum_i z_{i,k}}.$$
K-Means Clustering

Figure 9.1 of Pattern Recognition and Machine Learning.
It is easy to see that:

- updating the cluster assignment $z_{i,k}$ decreases the loss $J$ or keeps it the same;
- updating the cluster centers again decreases the loss $J$ or keeps it the same.

K-Means clustering therefore converges to a local optimum. However, it is quite sensitive to the starting initialization:

- It is common practice to run K-Means algorithm multiple times with different initialization and use the result with the lowest $J$ (scikit-learn uses `n_init=10` by default).
- Instead of using random initialization, k-means++ initialization scheme might be used, where the first cluster center is chosen randomly and others are chosen proportionally to the square of their distance to the nearest cluster center. It can be proven that with such initialization, the found solution has $O(\log K)$ approximation ratio in expectation.
K-Means Clustering

$K = 2$  
$K = 3$  
$K = 10$  
Original image

Figure 9.3 of Pattern Recognition and Machine Learning.
Gaussian Mixture vs K-Means

It could be useful to consider that different clusters might have different radii or even be ellipsoidal.

Different cluster analysis results on "mouse" data set:

Original Data             k-Means Clustering             EM Clustering

https://commons.wikimedia.org/wiki/File:ClusterAnalysis_Mouse.svg