Machine Learning: Dimensionality Reduction: Principal Components Analysis

PCA: Principal Components Analysis

Informal Motivation. A common situation in data analysis is this.

- A dataset has **a large number of features**: sometimes exceeding the number of available data points.
- Simple exploratory analysis of data suggests that a lot of features are not independent of each other (i.e., correlated to one degree or another).
- Analyst wants to obtain a representation of data that keeps the data variability intact (or almost intact), but uses fewer dimensions.

PCA in a nutshell. Principal Components Analysis (PCA for short) is an orthogonal transformation of a dataset into a new system of coordinates where

- each coordinate is orthogonal to others, and
- the coordinates are enumerated *in the order of decreased variance*.

PCA has the following properties:

- **Independence of dimensions.** Because each dimension in the new represenation is orthogonal to others, the ”features” that the new dimensions represent are all *independent of each other*.
- **Variability of data.** The new dimensions combined capture the same variability of the data as the original dataset.
- **Dimensionality reduction.** The number of dimensions can be reduced by selecting only the top k dimensions. The resulting representation will be an *approximation* of the original dataset, but this approximation will use *significantly fewer dimensions* than the original dataset.

Why maximize variability? Given a collection of data points, we want to be able to tell them apart as best as we can.

Finding a dimension along which these data point vary the most (have the highest variability) allows us to observe the actual differences between these data points.

PCA: The Math

Let $V = \{V_1, \ldots, V_n\}$ be a set of observed variables, $dom(V_i) = \mathbb{R}$.

Let $D = \{\mathbf{d_1}, \mathbf{d_2}, \ldots, \mathbf{d_m}\}$ be a dataset:

$$D = \begin{pmatrix} d_{11} & d_{12} & \ldots & d_{1n} \\ d_{21} & d_{22} & \ldots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \ldots & d_{mn} \end{pmatrix}$$

Step 1. Centralization. Let μ_i be the *sample mean* of V_i on dataset D. We centralize the dataset D as follows:

$$X = D - \begin{pmatrix} \mu_1 & \mu_2 & \ldots & \mu_n \\ \mu_1 & \mu_2 & \ldots & \mu_n \\ \vdots & \vdots & \ddots & \vdots \\ \mu_1 & \mu_2 & \ldots & \mu_n \end{pmatrix} = \begin{pmatrix} d_{11}-\mu_1 & d_{12}-\mu_2 & \ldots & d_{1n}-\mu_n \\ d_{21}-\mu_1 & d_{22}-\mu_2 & \ldots & d_{2n}-\mu_n \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1}-\mu_1 & d_{m2}-\mu_2 & \ldots & d_{mn}-\mu_n \end{pmatrix} =$$

$$= \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \ldots & x_{mn} \end{pmatrix} = \begin{pmatrix} - & \mathbf{x_1} & - \\ - & \mathbf{x_2} & - \\ \vdots & \vdots & \vdots \\ - & \mathbf{x_m} & - \end{pmatrix}$$

In dataset X, the means of all variables V_i are set to 0.

Step 2. Maximization of Variability. We want to find direction $\mathbf{v} = (v_1, \ldots, v_n)$ of the maximal variability of X. This means that we want to consider the following values:

$$s_i = \mathbf{x_i} \cdot \mathbf{v},$$

and find v such that the variance of the set $\{s_1, s_2, \ldots, s_m\}$ is the largest.

That is, we want to maximize the function:

$$Var(\mathbf{s}) = \sum_{i=1}^{m} s_i^2 = \sum_{i=1}^{m} (\mathbf{x_i} \cdot \mathbf{v})^2 = \mathbf{v}^T X^T X \mathbf{v}$$

Note: We can have $Var(\mathbf{s})$ be arbitrarily high if we pick $\mathbf{v}$ with arbitrarily high values.

We need to limit the *scale* of $\mathbf{v}$.

Step 3. Constraints on Solution. To limit the scale of $\mathbf{v}$ we introduce a constraint on the vectors $\mathbf{v}$:

$$||\mathbf{v}|| = 1.$$

This can be rewritten as

$$||\mathbf{v}|| = \mathbf{v} \cdot \mathbf{v} = \mathbf{v}^T\mathbf{v} = 1$$

We thus arrive to the following optimization problem.

Maximize

$$Var(\mathbf{v}) = \mathbf{v}^T X^T X \mathbf{v}$$

subject to

$$\mathbf{v}^T\mathbf{v} = 1$$

Step 4. Solution. We want to switch to an unconstrained optimization problem. To do this, *we introduce Lagrangian penalty* into our function:

$$L(\mathbf{v}, \lambda) = \mathbf{v}^T X^T X \mathbf{v} + \lambda(1 - \mathbf{v}^T\mathbf{v})$$

This function can now be optimized. We take the derivative of L w.r.t. $\mathbf{v}$:

$$\frac{\partial L}{\partial \mathbf{v}} = 2X^T X\mathbf{v} - 2\lambda\mathbf{v},$$

and set it to 0:

$$2X^T X\mathbf{v} - 2\lambda\mathbf{v} = 0,$$

i.e.

$$X^T X v = \lambda\mathbf{v}$$

What does this mean?

The solution is an eigenvector of the matrix $X^T X$. Which vector is it?

$$\mathbf{v}^T X^T X\mathbf{v} = \mathbf{v}^T(X^T X\mathbf{v}) = \mathbf{v}^T(\lambda\mathbf{v}) = \lambda(\mathbf{v}^T\mathbf{v}) = \lambda.$$

Because we want to maximize $\mathbf{v}^T X^T X\mathbf{v}$, this means that we are looking for $\mathbf{v}$ to be an eigenvector of the **largest eigenvalue** of matrix $X^T X$.

Spectral Theorem. If A is a symmetric matrix than A has an orthonormal basis of eigenvectors with real eigenvalues.

References

[1] Mohammed J. Zaki, Wagner Meira Jr., *Data Mining and Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press, 2014.