Principal Component Analysis

Iván Andrés Trujilllo Abella

ivantrujillo1229@gmail.com

References

The main course is:

- Mathematics for machine learning: PCA (Coursera)

Books:

- Introduction to Linear Algebra 2ed (Gilbert Strang)
- Linear Algebra Done Right (Undergraduate Texts in Mathematics)

Covariance $Cov(X, Y)$

...

$$cov(X, Y) = E[(X - E[X])(Y - E[Y])] \tag{1}$$

...

$$cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N - 1} \tag{2}$$

Covariance Matrix S

...

for k features

$$S = \begin{pmatrix} cov(X_1, X_1) & cov(X_1, X_2) & \ldots & cov(X_1, X_k) \\ cov(X_2, X_1) & cov(X_2, X_2) & \ldots & cov(X_2, X_k) \\ \vdots & \vdots & \ldots & \vdots \\ cov(X_k, X_1) & cov(X_k, X_2) & \ldots & Cov(X_k, X_k) \end{pmatrix} \tag{3}$$

Important points

- Is symmetric given that $cov(X, Y) = cov(Y, X)$
- Also named matrix of variance covariance given that $cov(X_j, X_j) = var(X_j)$

Covariance Matrix

for the $\boldsymbol{X}$ matrix of features with dimensions $n \times k$ and μ as the matrix of means can be expressed as

...

$$\boldsymbol{S} = \frac{1}{N-1}(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T \tag{4}$$

Now you need diagonalize this matrix.

Vector representation

You have a set of vectors, such that $x \in \mathbb{R}^{D\times 1}$

...

$$\tilde{x}_n = \sum_{i=1}^{M} B_{in} b_i + \sum_{i=M+1}^{D} B_{in} b_i \tag{5}$$

We are seeking M basis vectors.

in PCA

- We ignore $\sum_{i=M+1}^{D} B_{in} b_i$
- interested in subspace spanned by the basis vectors $\sum_{i=1}^{M} B_{in} b_i$

Lost function

...

$$J = \frac{1}{N}\sum_{i=1}^{N} \|x_n - \tilde{x}_n\|^2 \tag{6}$$

...

Now the problem is find B_{in} and b_i such that the average squared error is minimized.

Minimize lost function

...

$$\frac{\partial J}{\partial b_i} = \frac{\partial J}{\partial \tilde{x}_n} \frac{\partial \tilde{x}_n}{\partial b_i} \tag{7}$$

$$\frac{\partial J}{\partial B_{in}} = \frac{\partial J}{\partial \tilde{x}} \frac{\partial \tilde{x}}{\partial B_{in}} \tag{8}$$

The gradient

Definition

for the function $z = f(x_1, ..., x_n)$ The gradient is a vector that points in the direction of the greatest rate of incresase.

$$\frac{\partial z}{\partial x} = \nabla f \tag{9}$$

$$\nabla f = \begin{pmatrix} \frac{\partial z}{\partial x_1} \\ \frac{\partial z}{\partial x_2} \\ \vdots \\ \frac{\partial z}{\partial x_n} \end{pmatrix} \tag{10}$$

rules

...

$$x^T y = \sum x_i y_i \tag{11}$$

$$\frac{\partial x^T y}{\partial y_i} = x_i \tag{12}$$

...

$$\frac{\partial x^T y}{\partial y} = x^T \tag{13}$$

Rules

...

$$z = \|x - y\|^2 \tag{14}$$

$$\begin{aligned} z &= (x-y)^T(x-y) \\ &= (x^T - y^T)(x-y) \\ &= x^Tx - x^Ty - y^Tx + y^Ty \\ &= x^Tx - 2x^Ty + y^Ty \end{aligned} \tag{15}$$

...

$$\|x - y\|^2 = x^Tx - 2x^Ty + y^Ty \tag{16}$$

...

$$J = \frac{1}{N}\sum_{i=1}^{N}\|x_n - \tilde{x}_n\|^2 = \frac{1}{N}\sum_{i=1}^{N} = \left(x_n^T x_n - 2x_n^T\tilde{x}_n + \tilde{x}_n^T\tilde{x}\right) \tag{17}$$

Computing:

...

$$\frac{\partial\left(-2x_n^T\tilde{x}_n\right)}{\partial\tilde{x}_n} = -2x_n^T \tag{18}$$

...

$$\frac{\partial\left(\tilde{x}_n^T\tilde{x}_n\right)}{\partial\tilde{x}_n} = 2\tilde{x}_n^T \tag{19}$$

...

$$\frac{\partial J}{\partial \tilde{x}_n} = \frac{2}{n}\left(\tilde{x}_n - x_n\right)^T \tag{20}$$

...

$$\frac{\partial J}{\partial b_i} = \frac{2}{N}(\tilde{x}_n - x_n)^T B_{in} \tag{21}$$

...

$$\frac{\partial J}{\partial B_{in}} = \frac{2}{N}\left(\tilde{x}_n - x_n\right)^T b_i \tag{22}$$

Rewrite derivative

...

Given that are orthonormal basis:

$$B_{in} = x_n^T b_i \;\; \textbf{(why?)} \tag{23}$$

...

$$\frac{\partial J}{\partial B_{in}} = \frac{2}{N} \left(\left(\sum_{j=1}^{M} B_j n b_j \right)^T - x_n^T \right) b_i \tag{24}$$

Rewrite derivative

Take in mind that

...

$$\left(\sum_{j=1}^{M} B_j n b_j\right)^T = \sum_{j=1}^{M} B_j n b_j^T \textbf{ (why)?} \tag{25}$$

...

$$\frac{\partial J}{\partial B_{in}} = \sum B_{jn} b_j^T b_i - x_n^T b_i \tag{26}$$

definition (Kronecker delta)

for two orthonormal vectors its inner product

$$b_i.b_j = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

...

$$\frac{\partial J}{\partial B_{in}} = \sum_{j=1}^{M} B_{jn}\delta_{ij} - x_n^T b_i = B_{in} - x_n^T b_i \tag{27}$$

...

$$\frac{\partial J}{\partial B_{in}} = \frac{2}{N}(B_{in} - x^T b_i) \tag{28}$$

optimizing coordinates

...

$$\frac{\partial J}{\partial B_{in}} = 0 \tag{29}$$

...

$$B_{in} = x_n^T b_i \tag{30}$$

rewrite the vector

...

remember that $B_{jn} = x_n^T b_j$ is a scalar.

...

$$\tilde{x} = \sum_{j=1}^{M} (x_n^T b_j) b_j = \sum_{j=1}^{M} (b_j b_j^T) x_n \tag{31}$$

...

$$x_n = \sum_{j=1}^{M} \left(b_j b_j^T \right) x_n + \sum_{j=M+1}^{D} \left(b_j b_j^T \right) x_n \tag{32}$$

rewrite loss function

$$x_n - \tilde{x}_n = \sum_{j=M+1}^{D} \left(b_j b_j^T \right) x_n \quad (33)$$

now think in transformation $b_j b_j^T x_n = b_j x_n b_j^T$ or remember that $b_j^T x_n$ is an scalar.

...

$$\frac{1}{N} \sum_{j=M+1}^{D} \|(b_j^T x_n) b_j\|^2 \quad (34)$$

norm as combination of basis vectors

Remember that:

$$\begin{aligned} \|v\|^2 &= v.v \\ &= \left(\sum_{j=1}^{n} \alpha_j b_j\right) . \left(\sum_{k=1}^{n} \alpha_k b_k\right) \end{aligned} \tag{35}$$

also we could expressed as an review the properties:

$$\langle \sum_{j=1}^{n} \alpha_j b_j, \sum_{k=1}^{n} \alpha_k b_k \rangle \tag{36}$$

Try solve it.

Rewrite lost function

See apendix, given that are ortonormal vectors:

$$\|v\|^2 = \sum \alpha_i^2 \|b_i\|^2 \tag{37}$$

$v = \left(b_j^T x_n\right) b_j$ using the above equation and normal vectors (orthonormal given that orthogonality was assumed previously)

$$\|v\|^2 = \sum (b_j^T x_n)^2 \tag{38}$$

Rember that $b_j^T x_n$ is an scalar the

...

$$\|v\|^2 = \sum b_j^T x_n x_n^T b_j \tag{39}$$

Rewrite loss function

...

$$J = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=M+1}^{D} b_j^T x_n x_n^T b_j \tag{40}$$

Rewriting as:

...

$$J = \sum_{j=M+1}^{D} b_j^T \left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n^T\right) b_j \tag{41}$$

Note that $\frac{1}{N}\sum_{i=1} x_n x_n^T = S$ (Covariance matrix).

Rewrite loss function

...

$$J = \sum_{j=M+1}^{D} b_j^T S b_j \tag{42}$$

We can uses Lagrange multiplier to solve this problem.

...

$$\mathcal{L} = \sum_{j=M+1}^{D} b_j^T S b_j - \sum_{j=M+1}^{D} \lambda_j \left(1 - b_j^T b_j\right) \tag{43}$$

...

$$\frac{\partial \mathcal{L}}{\partial b_j} = 2b_j^T S - 2\lambda_j b_j^T = 0 \quad \forall j, j = 1, ..., D \tag{44}$$

....

$$Sb_j = \lambda_j b_j \quad \forall j, j = 1, ..., D \tag{45}$$

...

$$\frac{\partial \mathcal{L}}{\partial \lambda_j} = 1 - b_j^T b_J = 0 \quad \forall j, j = 1, ..., D \tag{46}$$

...

$$b_j^T b_j = 1 \quad \forall j, j = 1, ..., D \tag{47}$$

rewrite loss function

...

$$J = \sum_{j=M+1}^{D} b_j^T S b_j = \sum b_j^T b_j \lambda_j \tag{48}$$

...

$$J = \sum_{j=M+1}^{D} \lambda_j \tag{49}$$

To minimize J we minimize the eigenvalues associated with the covariance matrix.

Appendix

The norm of v expressed as linear combination of its basis vectors:

....

$$\|v\|^2 = \sum_{i=1} \sum_{j=1, j \neq i} \alpha_i \alpha_j \langle b_i, b_j \rangle + \sum_{i=1}^{n} \alpha_i^2 \|b_i\|^2 \tag{50}$$

...

$$v = \sum_{i=1}^{n} \alpha_i b_i \tag{51}$$

$$\langle v, v \rangle = \langle \sum_{i=1}^{n} \alpha_i b_i, \sum_{i=1}^{n} \alpha_i b_i \rangle \tag{52}$$

Using linearity of the left term and shorting the notation

$$\langle \sum_{i=1}^{n} \alpha_i b_i, v \rangle \tag{53}$$

$$\sum_{i=1}^{n} \alpha_i \langle b_i, v \rangle \tag{54}$$

$$\sum_{i=1}^{n} \alpha_i \langle b_i, \sum_{k=1}^{n} \alpha_k b_k \rangle \tag{55}$$

using linearity of the right term, (we add the k index given we need sum to all pair of terms).

$$\alpha_1 \langle b_1, \sum_{k=1}^{n} \alpha_k b_k \rangle + ... + \alpha_n \langle b_n, \sum_{k=1}^{n} \alpha_k b_k \rangle \tag{56}$$

$$\alpha_1 \left(\alpha_1 \langle b_1, b_1\rangle + .. + \alpha_n \langle b_n, b_n\rangle\right) + ... + \alpha_n \left(\alpha_1 \langle b_1, b_1\rangle + .. + \alpha_n \langle b_n, b_n\rangle\right)$$

$$\alpha_1\alpha_1 \langle b_1, b_1\rangle + \alpha_1\alpha_2 \langle b_1, b_2\rangle + ... + \alpha_n\alpha_1 \langle b_n, b_1\rangle + ... + \alpha_n\alpha_n \langle b_n, b_n\rangle$$

...

$$\sum_{i=1} \sum_{k=1} \alpha_i \alpha_k \langle b_i, b_k\rangle = \sum_{i=1} \sum_{k=1, k\neq i} \alpha_i \alpha_k \langle b_i, b_k\rangle + \sum \alpha_i^2 \|b_i\|^2 \qquad (57)$$

Apendix

...

$$\frac{\partial \boldsymbol{Ax}}{\partial \boldsymbol{x}} = \boldsymbol{A} \tag{58}$$

Think the following:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1,m} \\ a_{21} & a_{21} & \dots & a_{2,m} \\ \vdots & \vdots & \dots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,m} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} a_{1i} x_i \\ \sum_{i=1}^{m} a_{2i} x_i \\ \vdots \\ \sum_{i=1}^{m} a_{ni} x_i \end{bmatrix}$$

If apply we derivate with respect to $\boldsymbol{x}$

$$\begin{bmatrix} \frac{\partial f_1()}{\partial x_1} & \frac{\partial f_1()}{\partial x_2} & \dots & \frac{\partial f_1()}{\partial x_m} \\ \frac{\partial f_2()}{\partial x_1} & \frac{\partial f_2()}{\partial x_2} & \dots & \frac{\partial f_2()}{\partial x_m} \\ \vdots & \vdots & \dots & \vdots \\ \frac{\partial f_n()}{\partial x_1} & \frac{\partial f_n()}{\partial x_2} & \dots & \frac{\partial f_n()}{\partial x_m} \end{bmatrix}$$

Notice that

$$\frac{\partial \left(\sum_{i=1}^{n} a_{ki} x_i \right)}{\partial x_k} = a_{ki}$$

Therefore for k row and i column:

$$\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A} \tag{59}$$

.

...

$$\frac{\partial \left(\boldsymbol{x}^T \boldsymbol{A} x\right)}{\partial \boldsymbol{A}} = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{x} \tag{60}$$

$$\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} a_{11}x_1 + a_{12}x_3 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{bmatrix}$$

that could be expressed as:

$$\begin{gathered} x_1(a_{11}x_1 + a_{12}x_1 + a_{13}x_3)+ \\ x_2(a_{21}x_1 + a_{22}x_2 + a_{23}x_3)+ \\ x_3(a_{31}x_1 + a_{32}x_2 + a_{33}x_3) \end{gathered} \tag{61}$$

Note the following:
$\frac{\partial(f(x_1,\ldots,x_n))}{x_1} = 2a_{11}x_1 + \sum_{i=1,i\neq 1}^{3} a_{1i}x_i + \sum_{i=1,i\neq 1}^{3} a_{i1}x_i$
Therefore is easily extended:

$$\frac{\partial f}{\partial x_j} = 2a_{jj}x_j + \sum_{i=1,i\neq j}^{n} a_{ji}x_i + \sum_{i=1,i\neq j}^{n} a_{ij}x_i \tag{62}$$

The before equation we have that:

$$\left(\sum_{i=1,i\neq j}^{n} a_{ji}x_i + a_{jj}x_j\right) + \left(\sum_{i=1,i\neq j}^{n} a_{ij}x_i + a_{jj}x_j\right) \tag{63}$$

This can be rewritten as:

$$\left(\sum_{i=1}^{n} a_{ji} + \sum_{i=1}^{n} a_{ij}\right) x_i = \sum_{i=1}^{n}(a_{ji} + a_{ij})x_i \tag{64}$$

if $\vec{a}_j$ is the $j - row$ row of $\boldsymbol{A}$ matrix then $\vec{a}_j\vec{x} = \sum_{i=1}^{n} a_{ji}x_i$ therefore for the $j - th$ row of $\boldsymbol{A}^T$ $\vec{a}_j\vec{x} = \sum_{i=1}^{n} a_{ij}x_i$. for m variables therefore we have that:

$$(A + A^T)_{ji} = [a_{ji} + a_{ij}] \tag{65}$$

Therefore:

$$\frac{\partial\left(\boldsymbol{x}^T\boldsymbol{A}x\right)}{\partial\boldsymbol{A}} = (\boldsymbol{A}+\boldsymbol{A}^T)\boldsymbol{x} \tag{66}$$

Notice when the matrix $\boldsymbol{A}$ is symmetric then:

$$\frac{\partial\left(\boldsymbol{x}^T\boldsymbol{A}x\right)}{\partial\boldsymbol{A}} = 2\boldsymbol{A}\boldsymbol{x} \tag{67}$$

give that A is symmetric also can be written as: $2\boldsymbol{x}^T\boldsymbol{A}$.