Principal Component Analysis (PCA)

Professeur Samira Douzi
s.douzi@um5r.ac.ma

Qu'est-ce que l'analyse en composantes principales (PCA) ?

L'Analyse en Composantes Principales, ou PCA, est une méthode de réduction de la dimensionnalité qui est souvent utilisée pour réduire la dimensionnalité de grands ensembles de données, en transformant un grand ensemble de variables en un plus petit qui contient toujours la plupart des informations dans le grand ensemble.

Qu'est-ce que l'analyse en composantes principales (PCA) ?

Réduire le nombre de variables d'un ensemble de données se fait naturellement au détriment de la précision, mais l'astuce dans la réduction de la dimensionnalité consiste à échanger un peu de précision contre la simplicité. Parce que les ensembles de données plus petits sont plus faciles à explorer et à visualiser et rendent l'analyse des données beaucoup plus facile et plus rapide pour les algorithmes d'apprentissage automatique sans variables étrangères à traiter.

Qu'est-ce que l'analyse en composantes principales (PCA) ?

Donc, pour résumer, l'idée de PCA est simple - réduire le nombre de variables d'une Dataset, tout en préservant autant d'informations que possible.

Explication étape par étape de PCA

STEP 1: STANDARDIZATION

S'il existe de grandes différences entre les plages de variables initiales, les variables avec des plages plus grandes domineront sur celles avec de petites plages (par exemple, une variable comprise entre 0 et 100 dominera sur une variable comprise entre 0 et 1), ce qui conduira à des résultats biaisés. Ainsi, il faut transformer les données à des échelles comparables peut éviter ce problème.

Explication étape par étape de PCA

STEP 1: STANDARDIZATION

Mathématiquement, cela peut être fait en soustrayant la moyenne et en divisant par l'écart type pour chaque valeur de chaque variable.

$$z = \frac{value - mean}{standard\ deviation}$$

Une fois la standardisation effectuée, toutes les variables seront transformées à la même échelle.

Explication étape par étape de PCA

STEP 2: COVARIANCE MATRIX COMPUTATION

Le but de cette étape est de voir s'il existe une relation entre les variables du dataset. Parce que parfois, les variables sont fortement corrélées de telle sorte qu'elles contiennent des informations redondantes. Ainsi, afin d'identifier ces corrélations, nous calculons la matrice de covariance.

Explication étape par étape de PCA

STEP 2: COVARIANCE MATRIX COMPUTATION

La matrice de covariance est une matrice symétrique p × p (où p est le nombre de dimensions) qui a comme entrées les covariances associées à toutes les paires possibles des variables initiales. Par exemple, pour une data set de données tridimensionnel avec 3 variables x, y et z, la matrice de covariance est une matrice 3 × 3 de ceci à partir de :

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Covariance of Z = $Z^T Z$

Explication étape par étape de PCA

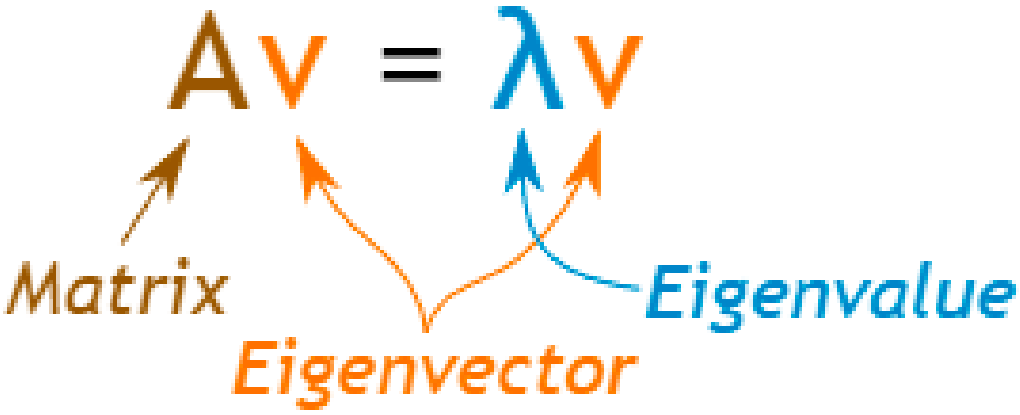
STEP 2: COVARIANCE MATRIX COMPUTATION

B13

	A	B	C	D	E	F	G
1		X	Y	Z			
2		Height	Score	Age			
3		64.0	580.0	29.0		var(X)	11.50
4		66.0	570.0	33.0		var(Y)	1250.00
5		68.0	590.0	37.0		var(Z)	110.00
6		69.0	660.0	46.0			
7		73.0	600.0	55.0		covar(XY)	50.00
8						covar(XZ)	34.75
9	m =	68.0	600.0	40.0		covar(YZ)	205.00
10							
11		n=5					
12							
13			X	Y	Z		
14		X	11.50	50.00	34.75		
15		Y	50.00	1250.00	205.00		
16		Z	34.75	205.00	110.00		
17							

Explication étape par étape de PCA

STEP 3: COMPUTE THE EIGENVECTORS AND EIGENVALUES



Explication étape par étape de PCA

STEP 3: COMPUTE THE EIGENVECTORS AND EIGENVALUES

Eigenvalues : Polynôme Caractéristique

Soit A une matrice

Son polynôme caractéristique est :

$$P_{\chi}(\lambda)=\det(A-\lambda I_i)$$

Polynôme Caractéristique : Exemple

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$$

$$|\mathbf{A} - \lambda \cdot \mathbf{I}| = \left| \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = 0$$

$$\left| \begin{bmatrix} -\lambda & 1 \\ -2 & -3-\lambda \end{bmatrix} \right| = \lambda^2 + 3\lambda + 2 = 0$$

λ_1=-1, λ_2=-2

Explication étape par étape de PCA

STEP 3: COMPUTE THE EIGENVECTORS AND EIGENVALUES

Soit X une matrice et λ une valeur propre de A
On dira que V un vecteur propre associé à la valeur propre

$$(A-\lambda I_i).V=0 \quad Ou \quad A.V=V.\lambda$$

EigenVectors : Exemple

$$\mathbf{A} \cdot \mathbf{v}_1 = \lambda_1 \cdot \mathbf{v}_1$$

$$(\mathbf{A} - \lambda_1) \cdot \mathbf{v}_1 = 0$$

$$\begin{bmatrix} -\lambda_1 & 1 \\ -2 & -3-\lambda_1 \end{bmatrix} \cdot \mathbf{v}_1 = 0$$

$$\begin{bmatrix} 1 & 1 \\ -2 & -2 \end{bmatrix} \cdot \mathbf{v}_1 = \begin{bmatrix} 1 & 1 \\ -2 & -2 \end{bmatrix} \cdot \begin{bmatrix} v_{1,1} \\ v_{1,2} \end{bmatrix} = 0$$

$$v_{1,1} + v_{1,2} = 0, \quad \text{so}$$

$$v_{1,1} = -v_{1,2}$$

$$-2 \cdot v_{1,1} + -2 \cdot v_{1,2} = 0, \quad \text{so again}$$

$$v_{1,1} = -v_{1,2}$$

$$\mathbf{v}_1 = k_1 \begin{bmatrix} +1 \\ -1 \end{bmatrix}$$

EigenVectors : Exemple

$$\mathbf{A} \cdot \mathbf{v}_2 = \lambda_2 \cdot \mathbf{v}_2$$

$$(\mathbf{A} - \lambda_2) \cdot \mathbf{v}_2 = \begin{bmatrix} -\lambda_2 & 1 \\ -2 & -3-\lambda_2 \end{bmatrix} \cdot \mathbf{v}_2 = \begin{bmatrix} 2 & 1 \\ -2 & -1 \end{bmatrix} \cdot \begin{bmatrix} v_{2,1} \\ v_{2,2} \end{bmatrix} = 0 \quad \text{so}$$

$$2 \cdot v_{2,1} + 1 \cdot v_{2,2} = 0 \quad (\text{or from bottom line: } -2 \cdot v_{2,1} - 1 \cdot v_{2,2} = 0)$$

$$2 \cdot v_{2,1} = -v_{2,2}$$

$$\mathbf{v}_2 = k_2 \begin{bmatrix} +1 \\ -2 \end{bmatrix}$$

Explication étape par étape de PCA

STEP 4 : SORT THE EIGEN VECTORS

prendre les valeurs propres λ_1, λ_2, …, λp et les trier de la plus grande à la plus petite. Ce faisant, triez les vecteurs propres dans le même ordre. (Par exemple, si $\lambda 3$ est la plus grande valeur propre, alors prenez le troisième vecteur et placez-le dans la première position de colonne de la matrice des composants principaux P.)

Explication étape par étape de PCA

Exemple

$$v1 = \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix} \qquad \lambda_1 = 1.284028$$

$$v2 = \begin{bmatrix} -0.7351785 \\ 0.6778736 \end{bmatrix} \qquad \lambda_2 = 0.04908323$$

pour calculer le pourcentage de variance (information) représenté par chaque composant, nous divisons Eigenvalue de chaque composant par la somme des Eigenvalues. Si on applique cela sur l'exemple ci-dessus, on constate que PC1 et PC2 portent respectivement 96% et 4% de la variance des données.

Explication étape par étape de PCA

STEP 5: FEATURE VECTOR

Dans cette étape, nous choisissons soit de conserver toutes ces composantes ou de rejeter celles de moindre importance (de faibles Eigenvalues), et de former avec les autres une matrice de vecteurs que nous appelons Feature vector.

Cela en fait la première étape vers la réduction de la dimensionnalité, car si nous choisissons de ne conserver que p EigenVectors (composants) sur n , l'ensemble de données final n'aura que p dimensions.

Explication étape par étape de PCA

DERNIÈRE ÉTAPE : REDEFINIR LES DONNÉES SELON LES PRINCIPAUX AXES DES COMPOSANTES

La dernière étape consiste à utiliser Feature vector formé à l'aide des EigenVectors, pour réorienter les données des axes d'origine vers ceux représentés par les composantes principales. Cela peut être fait en multipliant la transposée de dataset d'origine par la transposée du Feature vector .

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

TP python : PCA