Método de la Potencia y PCA

Métodos Numéricos

7 de mayo de 2022

Autovalores y Autovectores

Definición

Sea $A \in \mathbb{R}^{n \times n}$. Un *autovector* de A es un vector no nulo tal que $Ax = \lambda x$, para algun escalar λ. Un escalar λ es denominado *autovalor* de A si existe una solución no trivial x del sistema $Ax = \lambda x$. En este caso, x es llamado *autovector asociado a* λ.

Consideramos:

$$A = \begin{bmatrix} 3 & -2 \\ 1 & 0 \end{bmatrix}, u = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, v = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$Au = \begin{bmatrix} -5 \\ -1 \end{bmatrix}, Av = \begin{bmatrix} 4 \\ 2 \end{bmatrix} = 2 \begin{bmatrix} 2 \\ 1 \end{bmatrix} = 2v$$

Gráficamente....A sólo estira (o encoge) el vector v.

Autovalores y autovectores

Dada una matriz $A \in \mathbb{R}^{n \times m}$, un vector $\mathbf{x} \in \mathbb{R}^m$ no nulo puede ser paralelo a $A\mathbf{x}$, lo que significa que existe una constante λ tal que $A\mathbf{x} = \lambda\mathbf{x}$.

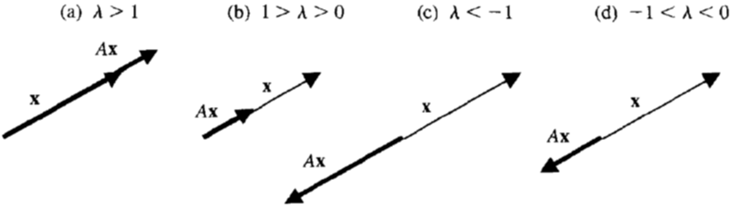
A todos los $\mathbf{x}$ que cumplen esta propiedad se los denomina autovectores de A y λ su autovalor asociado.

Autovalores y autovectores

Dada una matriz $A \in \mathbb{R}^{n\times m}$, un vector $\mathbf{x} \in \mathbb{R}^m$ no nulo puede ser paralelo a $A\mathbf{x}$, lo que significa que existe una constante λ tal que $A\mathbf{x} = \lambda\mathbf{x}$.

A todos los $\mathbf{x}$ que cumplen esta propiedad se los denomina autovectores de A y λ su autovalor asociado.

Representación gráfica:



En muchos casos, la presencia de autovectores-autovalores puede ser utilizada para encontrar una factorización $A = PDP^{-1}$, donde D es una matriz diagonal.

Intuición

Podemos encontrar una base donde la transformación lineal A se comporta como si fuese diagonal.

Observación

No toda matriz $A \in \mathbb{R}^{n \times n}$ es diagonalizable.

Teorema

Una matriz $A \in \mathbb{R}^{n \times n}$ es diagonalizable sí y solo sí A tiene n autovectores linealmente independientes (las columnas de P).

Más teoremas

Teorema Espectral

Si $A \in \mathbb{R}^{n \times n}$ es simétrica, entonces existe una base ortonormal de autovectores $\{v_1, \ldots, v_n\}$ asociados a $\lambda_1, \ldots, \lambda_n$.

Consecuencia: Existe P, y $P^{-1} = P^t$. Luego, $A = PDP^t$.

Teorema

Sea $A \in \mathbb{R}^{m \times n}$. Luego, si λ es autovalor de $A^T A$, con autovector v entonces λ es también autovalor de AA^T con autovector Av.

Cálculo de autovalores/autovectores

- La matriz de covarianza $M_x = \frac{1}{n-1}X^tX$ es simétrica y semidefinida positiva.
- Vamos a querer diagonalizar M_X para obtener la transformación que queremos. Para eso vamos a calcular sus autovectores.

Cálculo de autovalores/autovectores

- La matriz de covarianza $M_x = \frac{1}{n-1} X^t X$ es simétrica y semidefinida positiva.
- Vamos a querer diagonalizar M_X para obtener la transformación que queremos. Para eso vamos a calcular sus autovectores.
- Podemos considerar el Método de la Potencia para calcular λ_1 y v_1.
  1. MetodoPotencia(B,x_0,niter)
  2. $v \leftarrow x_0$.
  3. Para $i = 1, \ldots, niter$
  4. $v \leftarrow \frac{Bv}{||Bv||}$
  5. Fin Para
  6. $\lambda \leftarrow \frac{v^t Bv}{v^t v}$
  7. Devolver λ, v.

Cálculo de autovalores/autovectores

Una vez que tenemos λ_1 y v_1, como seguimos?

Cálculo de autovalores/autovectores

Una vez que tenemos λ_1 y v_1, como seguimos?

Deflación

Sea $B \in \mathbb{R}^{n \times n}$ una matriz con autovalores distintos $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$ y una base ortonormal de autovectores. Entonces, la matriz $B - \lambda_1 v_1 v_1^t$ tiene autovalores $0, \lambda_2, \ldots, \lambda_n$ con autovectores asociados $v_1, \ldots, v_n$.

Cálculo de autovalores/autovectores

Una vez que tenemos λ_1 y v_1, como seguimos?

Deflación

Sea $B \in \mathbb{R}^{n \times n}$ una matriz con autovalores distintos $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$ y una base ortonormal de autovectores. Entonces, la matriz $B - \lambda_1 v_1 v_1^t$ tiene autovalores $0, \lambda_2, \ldots, \lambda_n$ con autovectores asociados $v_1, \ldots, v_n$.

- $(B - \lambda_1 v_1 v_1^t) v_1 = B v_1 - \lambda_1 v_1 (v_1^t v_1) = \lambda_1 v_1 - \lambda_1 v_1 = 0 v_1$.
- $(B - \lambda_1 v_1 v_1^t) v_i = B v_i - \lambda_1 v_1 (v_1^t v_i) = \lambda_i v_i$.

Observación

En nuestro caso, no hace falta que todos los autovalores tengan magnitudes distintas.

Análisis de Componentes Principales
Motivación

En el contexto de nuestro problema tenemos que:

- Las instancias están en un espacio de dimensión alta. Problemas?

En el contexto de nuestro problema tenemos que:

- Las instancias están en un espacio de dimensión alta. Problemas?
- Tiene sentido pensar que el hecho de que las instancias estén en ese espacio es circunstancial. Probablemente existan otros, de menor dimensión, donde las instancias se puedan "expresar mejor".

Análisis de Componentes Principales

Intuición

- Para ir al nuevo espacio (S) vamos a aplicar una transformación de cambio de base ortogonal.

Análisis de Componentes Principales

Intuición

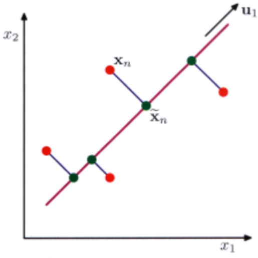
- Para ir al nuevo espacio (S) vamos a aplicar una transformación de cambio de base ortogonal.
- Vamos a aproximar cada instancia tomando un subespacio de menor dimensión de S.

Análisis de Componentes Principales
Intuición

- Para ir al nuevo espacio (S) vamos a aplicar una transformación de cambio de base ortogonal.
- Vamos a aproximar cada instancia tomando un subespacio de menor dimensión de S.
- Vamos a querer que elegir los nuevos vectores de manera que maximicen la varianza de las instancias trasnformadas y a su vez disminuir la covarianza (redundancia entre las direcciones).

Análisis de Componentes Principales

Intuición



En concreto, ¿Que buscamos en el nuevo subespacio de menor dimensión?

Dualidad

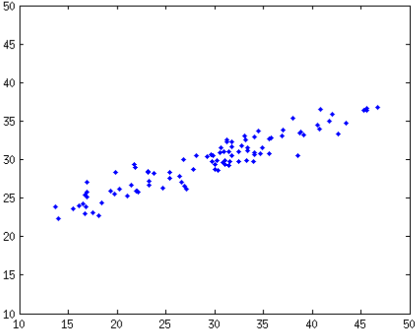
- encontrar direcciones tal que se maximice la varianza de los datos proyectados
- encontrar direcciones tal que se minimice la suma de los cuadrados de los errores de las proyecciones

Análisis de Componentes Principales

Ejemplo datos en $\mathbb{R}^2$

Sean $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ una secuencia de n datos, con $x^{(i)} \in \mathbb{R}^2$.

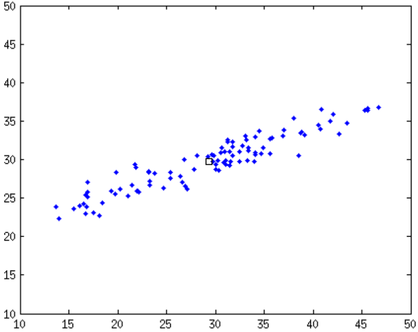
$$X = \begin{bmatrix} {x^{(1)}}^t \\ {x^{(2)}}^t \\ {x^{(3)}}^t \\ {x^{(4)}}^t \\ {x^{(5)}}^t \\ {x^{(6)}}^t \\ \vdots \\ {x^{(n)}}^t \end{bmatrix} = \begin{bmatrix} 26{,}4320 & 27{,}7740 \\ 26{,}8846 & 26{,}5631 \\ 23{,}3309 & 26{,}6983 \\ 30{,}6387 & 31{,}5619 \\ 30{,}5171 & 30{,}8993 \\ 45{,}6364 & 36{,}6035 \\ \vdots & \vdots \\ 16{,}0650 & 24{,}0210 \end{bmatrix}$$



Análisis de Componentes Principales

Ejemplo datos en $\mathbb{R}^2$

$$X = \begin{bmatrix} 26{,}4320 & 27{,}7740 \\ 26{,}8846 & 26{,}5631 \\ 23{,}3309 & 26{,}6983 \\ 30{,}6387 & 31{,}5619 \\ 30{,}5171 & 30{,}8993 \\ 45{,}6364 & 36{,}6035 \\ \vdots & \vdots \\ 16{,}0650 & 24{,}0210 \end{bmatrix}$$



Media:

$\mu = \frac{1}{n}(x^{(1)} + \cdots + x^{(n)})$

$\mu = (29{,}3623, 29{,}7148)$

Varianza de una variable x_k: Medida para la dispersión de los datos.

$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_k^{(i)} - \mu_k)^2$

$\sigma_{x_1}^2 = 66{,}2134, \;\; \sigma_{x_2}^2 = 12{,}5491$

Covarianza: Medida de cuánto dos variables varían de forma similar. Variables con mayor covarianza inducen la presencia de cierta dependencia o relación.

$$\sigma_{x_j x_k} = \frac{1}{n-1}\sum_{i=1}^{n}(x_j^{(i)} - \mu_j)(x_k^{(i)} - \mu_k)$$

Análisis de Componentes Principales

Ejemplo datos en $\mathbb{R}^2$ - Covarianza

Dadas n observaciones de dos variables x_1, x_2, y $v = (1, \ldots, 1)^t$:

$$\sigma_{x_1x_2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_1^{(i)} - \mu_1)(x_2^{(i)} - \mu_2) = \frac{1}{n-1}(x_2 - \mu_2 v)^t(x_1 - \mu_1 v)$$

Matriz de Covarianza:

$$X = \begin{bmatrix} 26{,}4320 - \mu_1 & 27{,}7740 - \mu_2 \\ 26{,}8846 - \mu_1 & 26{,}5631 - \mu_2 \\ 23{,}3309 - \mu_1 & 26{,}6983 - \mu_2 \\ 30{,}6387 - \mu_1 & 31{,}5619 - \mu_2 \\ 30{,}5171 - \mu_1 & 30{,}8993 - \mu_2 \\ 45{,}6364 - \mu_1 & 36{,}6035 - \mu_2 \\ \vdots & \vdots \\ 16{,}0650 - \mu_1 & 24{,}0210 - \mu_2 \end{bmatrix}$$

$$M_X = \frac{1}{n-1}X^tX = \begin{bmatrix} \sigma_{x_1x_1} & \sigma_{x_1x_2} \\ \sigma_{x_1x_2} & \sigma_{x_2x_2} \end{bmatrix} = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1x_2} \\ \sigma_{x_1x_2} & \sigma_{x_2}^2 \end{bmatrix}$$

$$M_X = \begin{bmatrix} 66{,}2134 & 27{,}1263 \\ 27{,}1263 & 12{,}5491 \end{bmatrix}$$

¿Cómo expresar mejor nuestros datos?

Objetivo

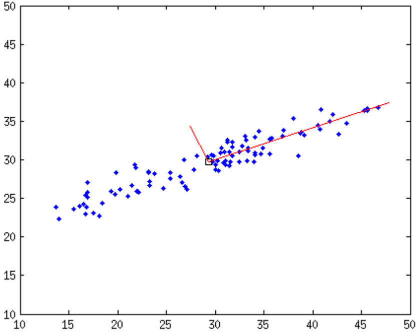
Buscamos una transformación de los datos que disminuya la redundancia (es decir, disminuir la covarianza).
Alternativa y equivalentemente: encontrar direcciones ortogonales que maximicen la varianza de nuestros datos

- Cambio de base: $\hat{X}^t = PX^t$.
- Cómo podemos hacerlo? Diagonalizar la matriz de covarianza. Esta matriz tiene la varianza de cada variable en la diagonal, y la covarianza en las restantes posiciones. Luego, al diagonalizar buscamos variables que tengan covarianza cero entre sí y la mayor varianza posible.
- Si quieren ver una demo, *Pattern Recognition and Machine Learning* de *Christopher Bishop*

¿Cómo expresar mejor nuestros datos?

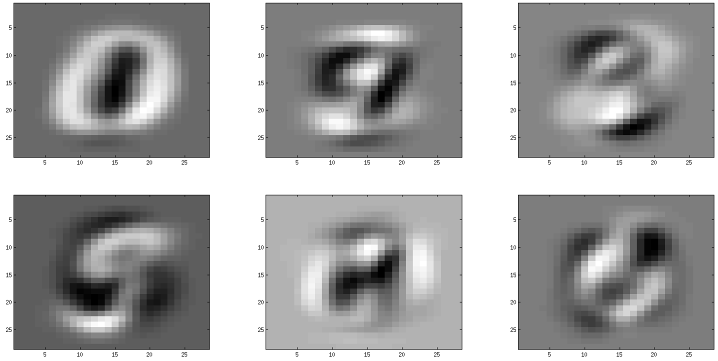
Volvemos al ejemplo

$$
\begin{aligned}
M_X &= \begin{bmatrix} 66{,}2134 & 27{,}1263 \\ 27{,}1263 & 12{,}5491 \end{bmatrix} \\
&= \underbrace{\begin{bmatrix} 0{,}9228 & -0{,}3852 \\ 0{,}3852 & 0{,}9228 \end{bmatrix}}_{V} \underbrace{\begin{bmatrix} 77{,}5362 & 0 \\ 0 & 1{,}2263 \end{bmatrix}}_{D=M_{\hat{X}}} \underbrace{\begin{bmatrix} 0{,}9228 & 0{,}3852 \\ -0{,}3852 & 0{,}9228 \end{bmatrix}}_{V^t}
\end{aligned}
$$



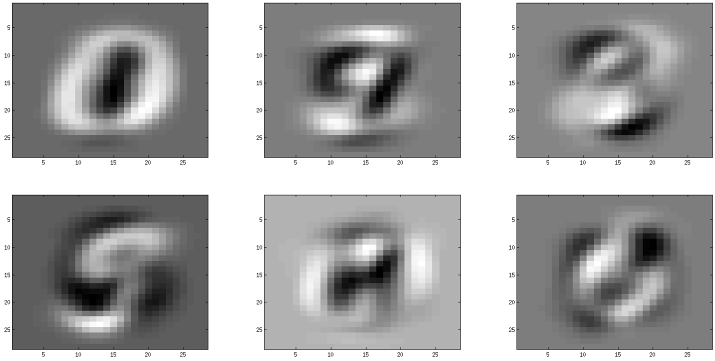
Reconocimiento de dígitos

Cualquier digito del dataset puede aproximarse como la combinación lineal de los primeros 6 eigendigits de la nueva base. Notar que ahora las instancias son de $\mathbb{R}^6$ (las coordenadas que salen de la cl).



Reconocimiento de dígitos

Cualquier digito del dataset puede aproximarse como la combinación lineal de los primeros 6 eigendigits de la nueva base. Notar que ahora las instancias son de $\mathbb{R}^6$ (las coordenadas que salen de la cl).



También es útil utilizar este tipo de técnicas para ayudar a la**visualización de datos**

Análisis de Componentes Principales

Resumen hasta acá

- Tenemos n muestras de m variables.
- Calculamos el vector μ que contiene la media de cada de una las variables.
- Construimos la matriz $X \in \mathbb{R}^{n\times m}$ donde cada muestra corresponde a una fila de X y tienen media cero (i.e., $X_i := (x^{(i)} - \mu)$).
- Diagonalizamos la matriz de covarianzas $M_X = \frac{X^t X}{n-1}$. La matriz V (ortogonal) contiene los autovectores de M_X.

Propiedades del cambio de base

- Disminuye redundancias.
- El cambio de base $\hat{X}^t = PX^t = V^t X^t$ asigna a cada muestra un nuevo *nombre* mediante un cambio de coordenadas.
- Las columnas de V (autovectores de M_X) son las componentes principales de los datos.
- En caso de m grande, es posible tomar sólo un subconjunto de las componentes principales para estudiar (i.e., aquellas que capturen mayor proporción de la varianza de los datos).