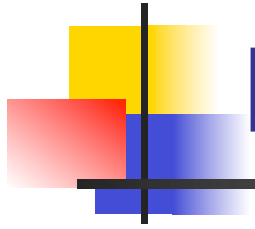




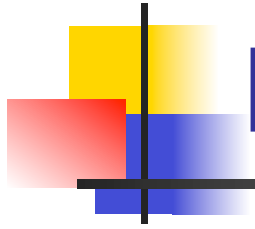
Aprendizaje de Máquina

ITAM



Menú

- Técnicas para la reducción de dimensionalidad
 - Selección de atributos
 - Selección de subconjuntos de atributos
 - Extracción de rasgos
 - Análisis de componentes principales (PCA)



Motivación

- La maldición de la dimensionalidad
 - Este fenómeno se refiere a que mientras más dimensiones (atributos) tengan los datos más ejemplos de entrenamiento se requerirán para aprender



La Maldición de la Dimensionalidad

- Si dividimos el espacio en cubos (hiper-cubos) unitarios y pensamos que debemos tener ejemplos de entrenamiento dentro de cada hiper-cubo para decir que muestreamos bien el espacio
 - ¿Cómo cambia el número de hiper-cubos unitarios para llenar el espacio?
 - Una línea de longitud 10 se “llena” con 10 hipercubos de dimensión 1
 - Un cuadrado de lado 10 se llena con 100 hipercubos de dimensión 1
 - Un cubo de lado 10 se llena con 1000 hipercubos de dimensión 1
 - Por cada dimensión extra el espacio crece en un orden de magnitud



La Maldición de la Dimensionalidad

- Otra manera de pensarlo es que cuando tenemos una medida de distancia que define, por ejemplo, una hiper-esfera, el aumento de dimensión ocasiona que la inmensa mayoría de las cosas queden lejos del centro
- Escencialmente lo que sucede es que al aumentar las dimensiones disminuye el volumen de la hiper-esfera unitaria



La Maldición de la Dimensionalidad

- La fórmula para el volumen de una hipersfera radio 1 es:

$$v_n = \left(\frac{2\pi}{n} \right) v_{n-2}, n \geq 3$$

n	Vn
1	2
2	3.141592654
3	4.188790205
4	4.934802201
5	5.263789014
6	5.16771278
7	4.72476597
8	4.058712126
9	3.298508903
10	2.55016404
11	1.884103879
12	1.335262769
13	0.910628755
14	0.599264529

La Maldición de la Dimensionalidad

x1	x2	x3	x4	x5
0.920265056	0.26452019	0.30489182	0.04169455	0.84777899
0.650880888	0.72257056	0.93792103	0.81626444	0.99629299
0.584418079	0.41394878	0.74564532	0.07076269	0.80383761
0.079327314	0.43830391	0.49845982	0.64060199	0.74885878
0.32663081	0.3137468	0.72152193	0.38348997	0.32952658
0.631366981	0.75508859	0.67049779	0.60748443	0.31974318

D(x1)	D(x1,x2)	D(x1,x2,x3)	D(x1,x2,x3,x4)	D(x1,x2,x3,x4,x5)
0.92026506	0.95752739	1.00489687	1.00576149	1.31540312
0.65088089	0.97249892	1.35109215	1.57852388	1.86663798
0.58441808	0.71616903	1.03386896	1.03628779	1.31150573
0.07932731	0.44542467	0.66847986	0.92587053	1.19080885
0.32663081	0.45290699	0.85189121	0.93422866	0.99064169
0.63136698	0.98426777	1.19094514	1.33693218	1.37463572

Distancia Promedio al origen	0.53214819	0.75479913	1.01686237	1.13626742	1.34160552
------------------------------	------------	------------	------------	------------	------------



Reducción de Dimensionalidad

- La complejidad de un modelo muchas veces depende de el número de atributos en los datos
 - Igualmente el número de ejemplos de entrenamiento necesarios crece con el número de atributos
- Adicionalmente muchas técnicas de clasificación y regresión son incapaces de identificar atributos irrelevantes
- Algunos atributos irrelevantes causan imprecisiones en los modelos



Reducción de Dimensionalidad

- El reducir el número de atributos reduce la complejidad del modelo
 - Modelos más simples suelen ser más consistentes: generalizan mejor
- En ocasiones la recolección de atributos es costosa. Si el atributo no es relevante, podemos ahorrar el costo de recolección



Reducción de Dimensionalidad

- Existen técnicas supervisadas y no supervisadas
 - Que necesitan o no el valor de la función objetivo
- Además estos métodos se dividen en:
 - Selección de atributos
 - Del conjunto de p atributos seleccionar k
 - Extracción de rasgos
 - Encontrar k rasgos formados a partir de los p atributos
- Vamos a ver una técnica supervisada de selección de atributos y una técnica no supervisada de extracción de rasgos
 - Selección de subconjuntos
 - Análisis de componentes principales



Selección de Rasgos

Selección de Subconjuntos

- El objetivo es encontrar el subconjunto de atributos que contribuye más a la precisión del modelo
 - ¿Cuántos posibles subconjuntos hay?
 - Demasiados
- Es impráctico explorar todas las posibles combinaciones de los p atributos
- Existen dos técnicas que, utilizando un heurístico, encuentran un “buen” subconjunto en tiempo razonable
- Es una técnica supervisada pues necesita el valor de la función objetivo para los datos de entrenamiento



Selección de Subconjuntos

Selección Incremental

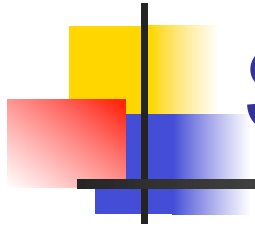
- Sea $A=\{x_1, \dots, x_p\}$ el conjunto de atributos y ERROR un error máximo aceptable
- $F=\{\}$
- Do{
 - Para cada atributo x_j de A
 - Calcular el error ε de clasificación o regresión del modelo usando $F \cup \{x_j\}$
 - Guardar el atributo x_m (y el error ε) que causa la mayor disminución
 - $F \leftarrow F \cup \{x_m\}$
 - $A \leftarrow A - \{x_m\}$
- Until ($\varepsilon < \text{ERROR}$) or ($A=\{\}$)



Selección de Subconjuntos

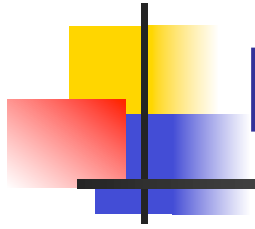
Selección Decremental

- Sea $F = \{x_1, \dots, x_p\}$ el conjunto de todos los atributos y ERROR un error máximo aceptable
- $F \leftarrow A$
- Do{
 - Para cada atributo x_j de F
 - Calcular el error ε de clasificación o regresión del modelo usando $F - \{x_j\}$
 - Desechar el atributo x_m (y el error ε) que causa la menor reducción en el error
 - $F \leftarrow F - \{x_m\}$
- Until ($\varepsilon < \text{ERROR}$) or ($A = \{\}$)



Selección de Subconjuntos

- Es muy importante entrenar el modelo con un grupo de datos y calcular el error con otro grupo independiente
- La selección decremental es un poco más costosa pues entrenar un modelo con más atributos suele ser más costoso.
 - Pero puede encontrar algunos subconjuntos de variables que contribuyen a reducir el error sólo cuando están juntos
 - E.g. Saldo, deuda y edad. Puede ser que individualmente (o en parejas) ni el saldo de un cliente, ni su deuda, ni su edad ayuden a un clasificador, pero que juntos si lo hagan
- Sirven para cualquier técnica de aprendizaje supervisado



Extracción de Rasgos

- En lugar de seleccionar un subconjunto de atributos estas técnicas forman nuevos rasgos a partir de combinaciones de atributos
 - E.g. En lugar de tener saldo y deuda, creamos saldo/deuda



Componentes Principales (Principal Components Analysis)

- Esta técnica encuentra relaciones entre los atributos para formar nuevos rasgos
- Es un técnica no-supervisada
 - No se necesita el valor de la función objetivo para los datos de entrenamiento
- Sólo encuentra relaciones lineales
- Asume que la co-varianza es un buen indicador de lo interesante de una relación entre atributos



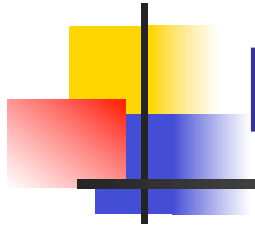
Componentes Principales (PCA)

- Tenemos N datos de entrenamiento con p atributos cada uno (p dimensiones)
 - El objetivo es encontrar un nuevo espacio de $k \leq p$ dimensiones que describa las características importantes (para el aprendizaje)



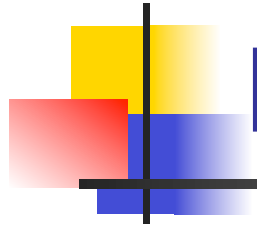
Componentes Principales (PCA)

- La idea general es encontrar combinaciones lineales de los atributos (rasgos) y conservar aquellos que maximizan la varianza de los datos
 - Los rasgos que “explican” la mayor variabilidad de los datos son los que mejor pueden servir para clasificar nuevas instancias (también sirve para regresión)
 - Estos son los rasgos que mejor caracterizan las diferencias entre grupos de datos y por ende los que mejor describen a datos similares



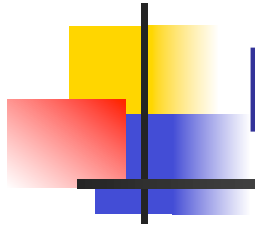
Proceso (idea)

- Calculamos la matriz de covarianza
- Encontramos en que dirección hay mayor variabilidad y repetimos para cada una de las direcciones perpendiculares a ésta
- Conservamos sólo las direcciones “más significativas”
- Transformamos los datos usando los rasgos encontrados



Pequeño Repaso

- Covarianza
- Vectores y valores característicos



Matriz de Covarianza

- Indica la variabilidad de un atributo con respecto a otro
- $\text{Cov}(A_1, A_2) =$

$$\left(\sum (A_{1,i} - \text{Media}(A_1)) (A_{2,i} - \text{Media}(A_2)) \right) / (n-1)$$

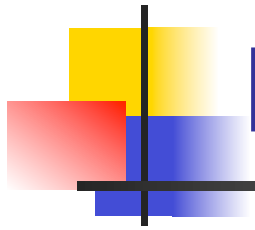
- La matriz de covarianza contiene todas las combinaciones: A_1 con A_2 , A_2 con A_3, \dots, A_p con A_1, \dots, A_p con A_p .
- Esto resulta en una matriz con p^2 entradas
- Note que la diagonal de la matriz tiene la varianza de un atributo



PCA

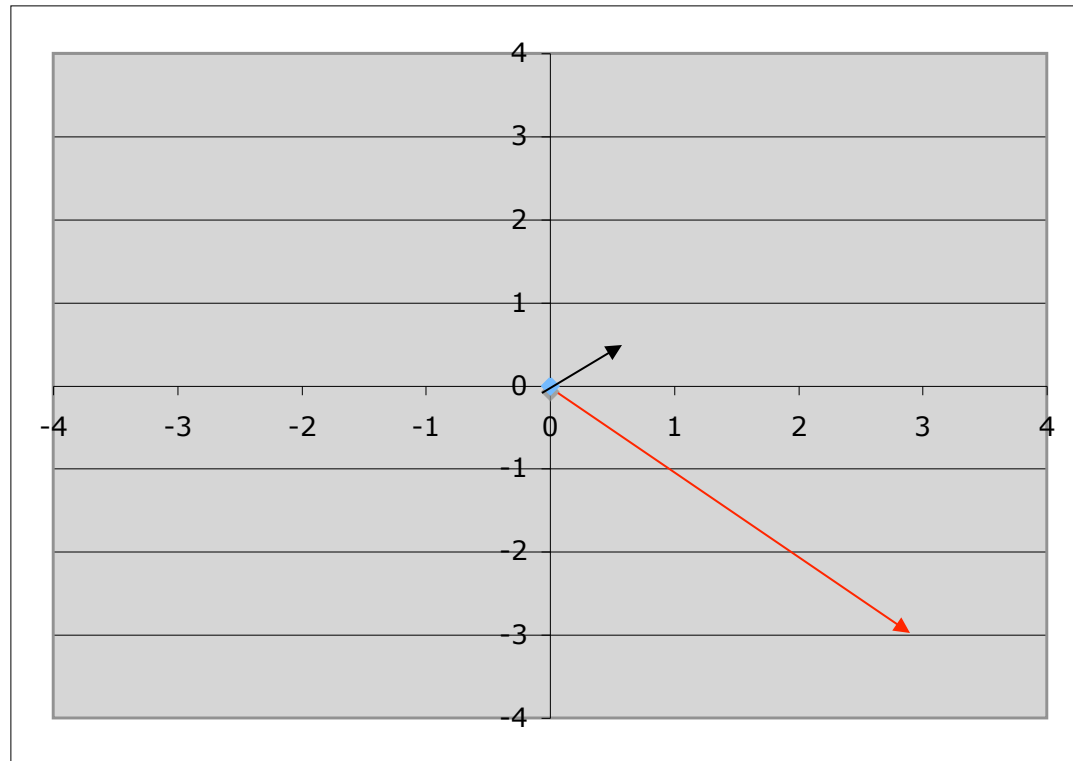
Vectores Característicos

- Un vector característico es aquel que no cambia más que de magnitud por una transformación dada
- Pensamos en una matriz como una transformación



PCA Ejemplo

Transformación $\begin{bmatrix} 4 & 2 \\ -4 & -2 \end{bmatrix}$ vector $\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$

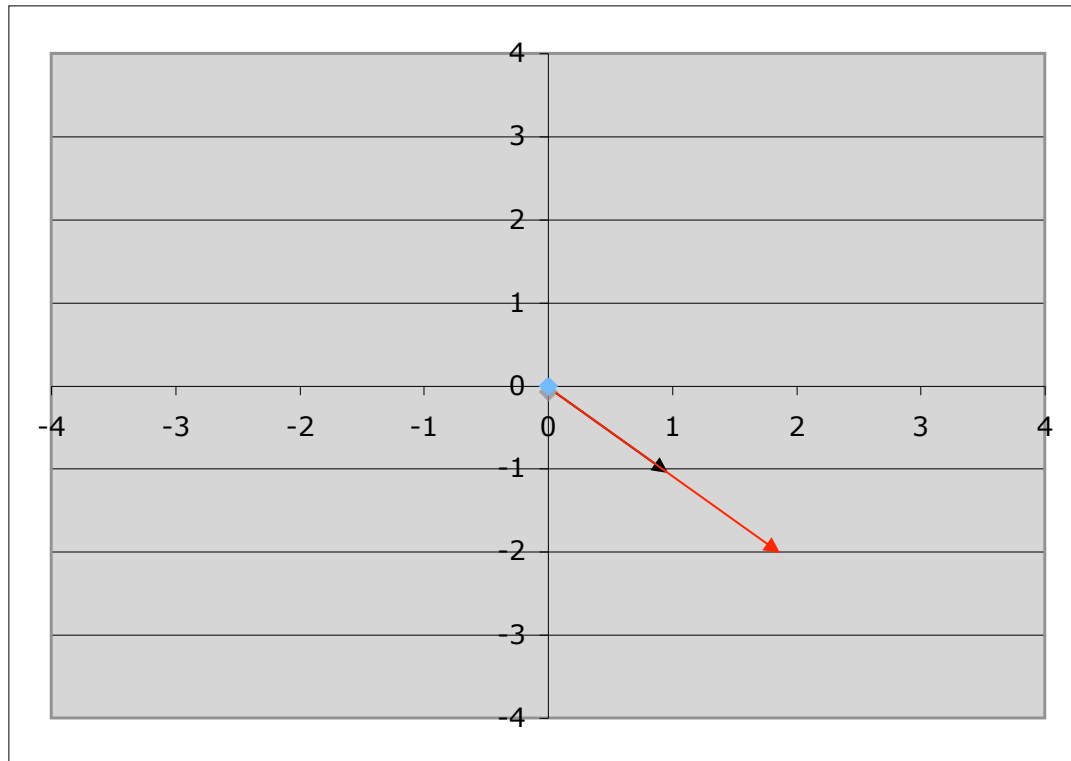


Se
transforma
en

$$\begin{bmatrix} 3 \\ -3 \end{bmatrix}$$

PCA Ejemplo

Transformación $\begin{bmatrix} 4 & 2 \\ -4 & -2 \end{bmatrix}$ vector $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$



Se
transforma
en

$$\begin{bmatrix} 2 \\ -2 \end{bmatrix}$$



Valor Característico

- Un valor característico es la cantidad por la que cambia la magnitud de un eigen vector al ser multiplicado por la transformación

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4 \times \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 6 \\ 4 \end{bmatrix} = \begin{bmatrix} 24 \\ 16 \end{bmatrix} = 4 \times \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

- El valor característico es 4



Propiedades de los Vectores Característicos y relación a PCA

- Sólo existen para matrices cuadradas
 - La matriz de covarianza es cuadrada
- Una matriz de $n \times n$ tiene a lo más n vectores característicos diferentes de magnitud unitaria
 - Estos usamos para PCA
- Los vectores característicos son perpendiculares (ortogonales) entre si
 - Lo que ocasiona que la matriz de covarianza de los datos transformados tenga ceros menos en la diagonal
- El valor característico asociado al vector indica la fuerza (importancia) de dicha relación
 - Un valor entre la suma de todos los valores característicos puede interpretarse como la cantidad de la varianza de la que da cuenta



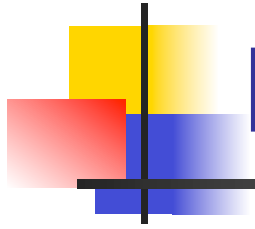
Componentes Principales (PCA-Algoritmo)

- Sea \mathbf{X} la matriz de datos con p dimensiones
 1. Para cada atributo i , calcular la media m_i . Para los N datos restar las correspondientes m_i de sus atributos. La matriz resultante es \mathbf{X}'
 2. Calcular la matriz de covarianza Σ de \mathbf{X}' . La matriz será de $p \times p$
 - Promedio de la multiplicación de las diferencias a la media
 - Qué tanto varía cada pareja de atributos
 - Qué tanto cambia la dimensión uno con respecto a la dos,



Componentes Principales (PCA-Algoritmo)

3. Calcular los vectores y valores característicos de Σ (eigenvectores y eigenvalores)
 - Los eigen vectores las direcciones que mejor caracterizan la covarianza de los atributos
4. Escoger los k componentes principales (los vectores cuyos valores característicos sean los mayores) y formar el vector de rasgos **R**
 - **R** tiene los componentes principales en cada columna. Es una matriz de $p \times k$ (contiene los k rasgos que encontramos)
5. Finalmente aplicamos esta transformación a nuestros datos **X'**)
 - Datos transformados = $\mathbf{R}^T \mathbf{X}'^T$



Proceso (idea)

- Sirve imaginarse PCA como un ajuste de mínimos cuadrados en las distintas direcciones de variabilidad de los datos
 - Calculamos la matriz de covarianza
 - Encontramos en que dirección hay mayor variabilidad y repetimos para cada una de las direcciones perpendiculares a ésta
 - Conservamos sólo las direcciones “más significativas”
 - Transformamos los datos usando los rasgos encontrados

Ejemplo PCA

(Datos de Lindsay I Smith)

Paso 1

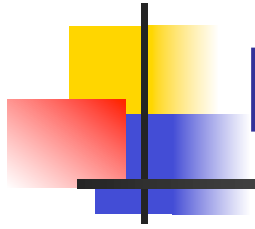
Datos

x1	x2	x1-m1	x2-m2
2.5	2.4	0.69	0.49
0.5	0.7	-1.31	-1.21
2.2	2.9	0.39	0.99
1.9	2.2	0.09	0.29
3.1	3	1.29	1.09
2.3	2.7	0.49	0.79
2	1.6	0.19	-0.31
1	1.1	-0.81	-0.81
1.5	1.6	-0.31	-0.31
1.1	0.9	-0.71	-1.01

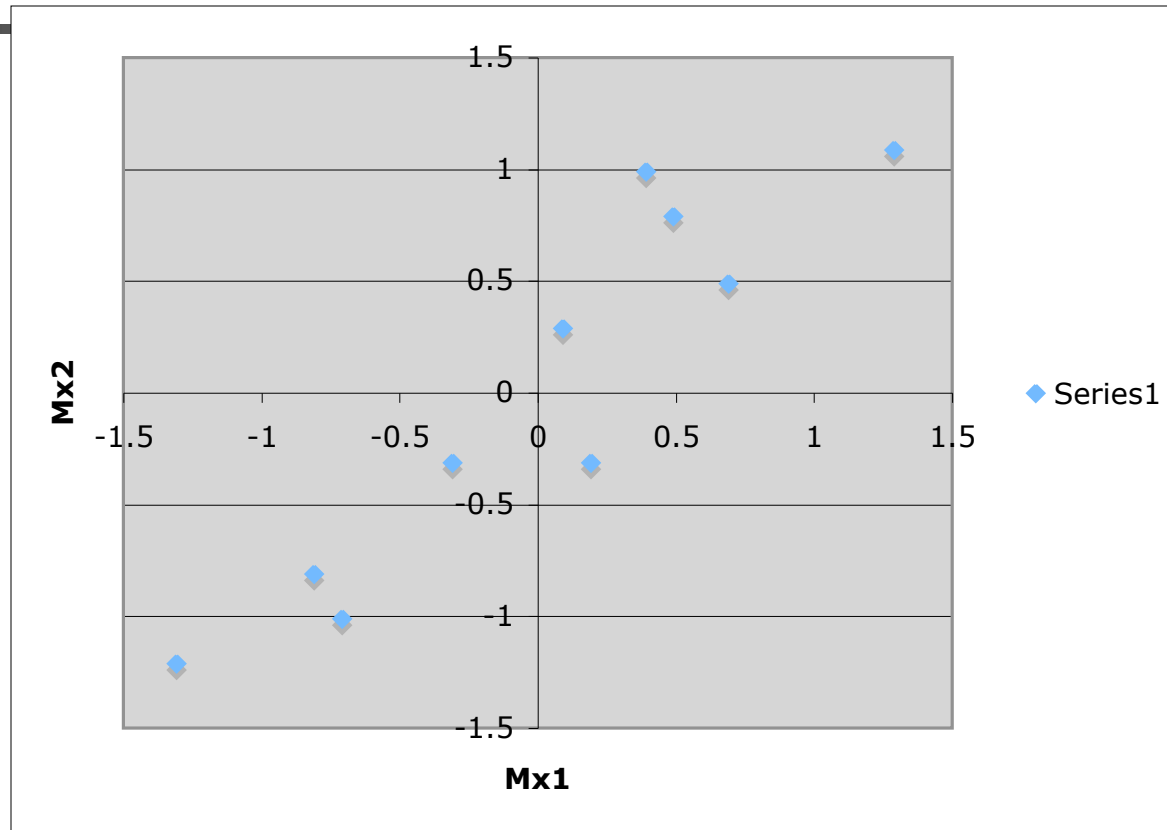
m_1, m_2

1.81

1.91



Ejemplo PCA



- Gráfica de los datos-media

Ejemplo PCA

Paso 2

- Covarianza

$E((x_1 - m_1)(x_1 - m_1))$	$E((x_1 - m_1)(x_2 - m_2))$
$E((x_2 - m_2)(x_1 - m_1))$	$E((x_2 - m_2)(x_2 - m_2))$

0.5549	0.5539
0.5539	0.6449



Ejemplo PCA

Paso 3

■ Vectores y valores característicos

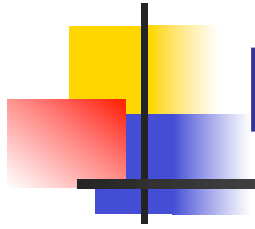
Vectores característicos

eigen1	eigen2
-0.735179	0.677873
0.677873	0.735179

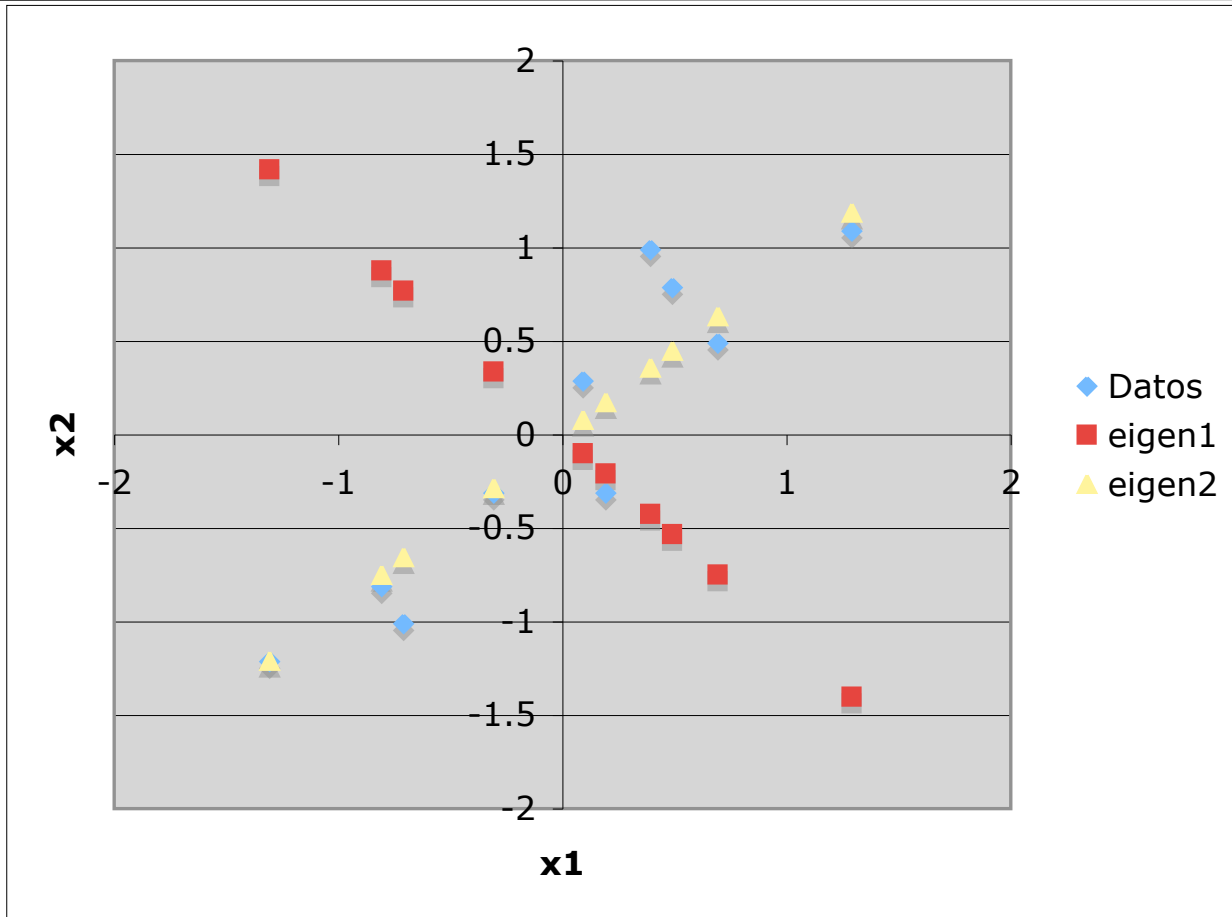
Valores característicos

eigen1	0.044175
eigen2	1.155625

El segundo vector da cuenta del 72% de la varianza
 $1.155/(1.155+0.044)=0.72$



Ejemplo PCA



- Gráfica de datos y vectores característicos



Ejemplo PCA

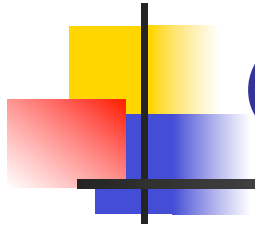
Paso 4 y 5

- Escogemos el vector con mayor valor característico (el segundo) y multiplicamos por los datos.

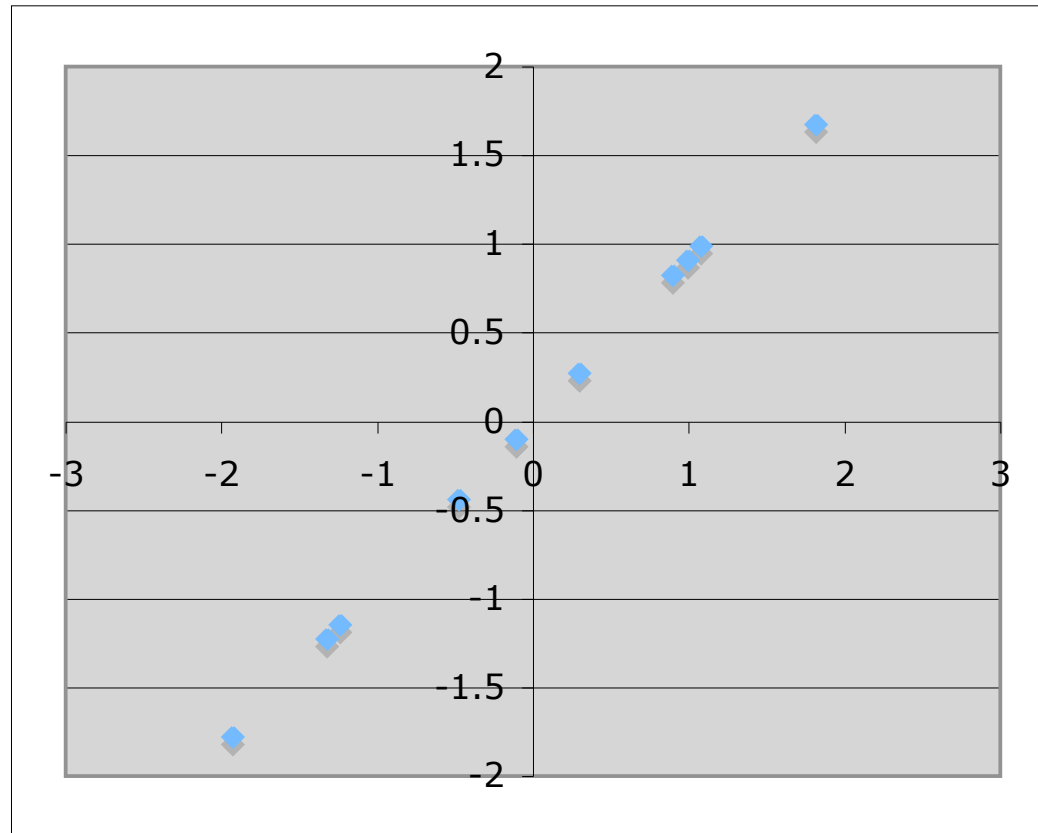
Nuevos datos

0.82797008
-1.77758022
0.99219768
0.27421048
1.67580128
0.91294918
-0.09910962
-1.14457212
-0.43804612
-1.22382062

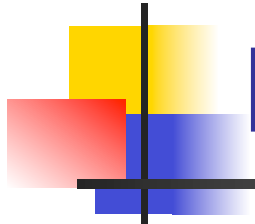
$$\text{Nuevos} = \text{eigen2}^T \mathbf{X}'^T$$



Gráfica de los Nuevos Datos



Cada vector propio define la proyección de un dato sobre el nuevo eje. Los ejes capturan las combinaciones lineales de los atributos originales donde hay más variación, que mejor separan los datos



PCA

- En general conservamos k vectores característicos y de esta manera expresamos los datos en términos de los patrones que mejor describen los datos (covarianzas)
- Si k es menor a p entonces también hemos reducido la dimensionalidad del problema



Ejercicio

- Repita el ejercicio en python