Aprendizaje automático

Clase 7

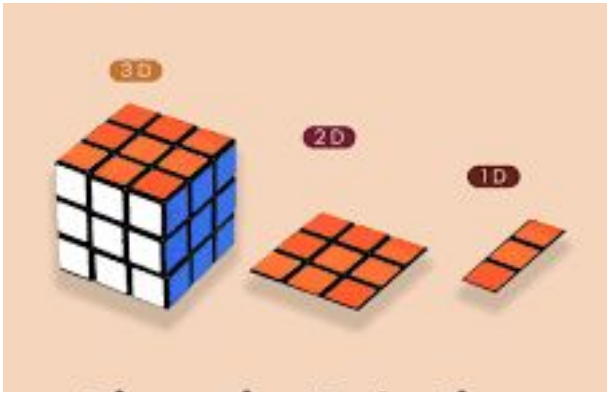
Maximiliano Beckel

UNAHUR

Reducción de la dimensionalidad

Los datos con los que se trabaja en Aprendizaje automático suelen tener una alta dimensionalidad. ¿Qué significa eso? Cada dimensión de nuestros datos está dada cada una de las variables/features que estamos registrando de nuestros datos. En Aprendizaje automático, cada observación esta relacionada con un gran número de variables (tablas con muchas columnas).

EJEMPLO: tenemos una tabla con datos de distintos pacientes a los cuales se les midió la altura, peso, presión sanguínea y concentración en sangre. Esos datos "viven" en un espacio de 4 dimensiones, una por cada variable que tenemos.



Reducción de la dimensionalidad

Los datos con los que se trabaja en Aprendizaje automático suelen tener una alta dimensionalidad. ¿Qué significa eso? Cada dimensión de nuestros datos está dada cada una de las variables/features que estamos registrando de nuestros datos. En Aprendizaje automático, cada observación esta relacionada con un gran número de variables (tablas con muchas columnas).

Técnicas de reducción de dimensionalidad

- Técnica utilizada para reducir el número de atributos (features) en un conjunto de datos, manteniendo la mayor cantidad de información posible de los datos originales.
- Proceso para transformar datos de alta dimensionalidad hacia un espacio de menor dimensionalidad que preserva la “esencia” de los datos originales.

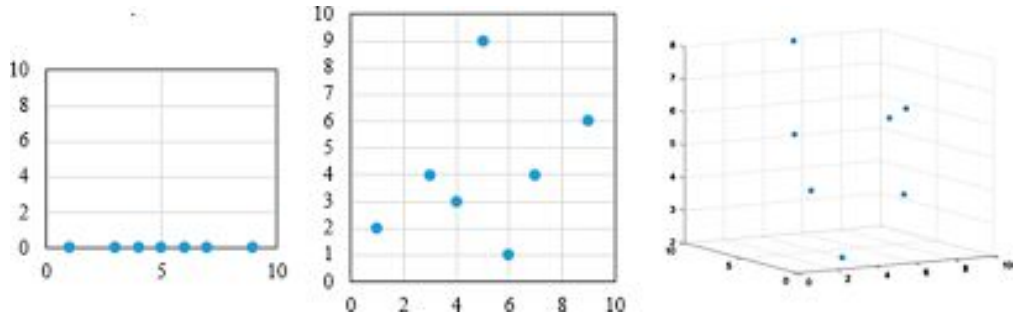
UNAHUR

Reducción de la dimensionalidad

Los datos con los que se trabaja en Aprendizaje automático suelen tener una alta dimensionalidad. ¿Qué significa eso? Cada dimensión de nuestros datos está dada cada una de las variables/features que estamos registrando de nuestros datos. En Aprendizaje automático, cada observación esta relacionada con un gran número de variables (tablas con muchas columnas).

Maldición de la dimensionalidad

- Cuando la dimensionalidad aumenta, el volumen del espacio aumenta tan rápidamente que los datos representados quedan muy dispersos (sparse).
- La cantidad de datos necesarios para analizar estos espacios crece exponencialmente con la dimensionalidad.
- En espacios de alta dimensionalidad todos los puntos tienden a ser distantes (disímiles) entre sí, lo que dificulta encontrar organización o estructura en los datos.

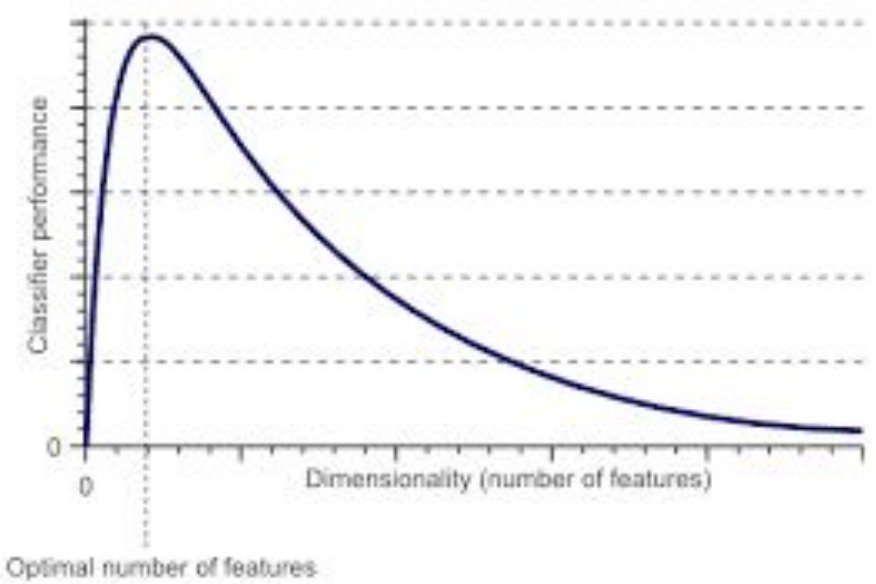


Reducción de la dimensionalidad

Los datos con los que se trabaja en Aprendizaje automático suelen tener una alta dimensionalidad. ¿Qué significa eso? Cada dimensión de nuestros datos está dada cada una de las variables/features que estamos registrando de nuestros datos. En Aprendizaje automático, cada observación esta relacionada con un gran número de variables (tablas con muchas columnas).

Maldición de la dimensionalidad

- Por otro lado, la alta dimensionalidad de los datos puede afectar la performance de algunos modelos.



Reducción de la dimensionalidad

Ventajas

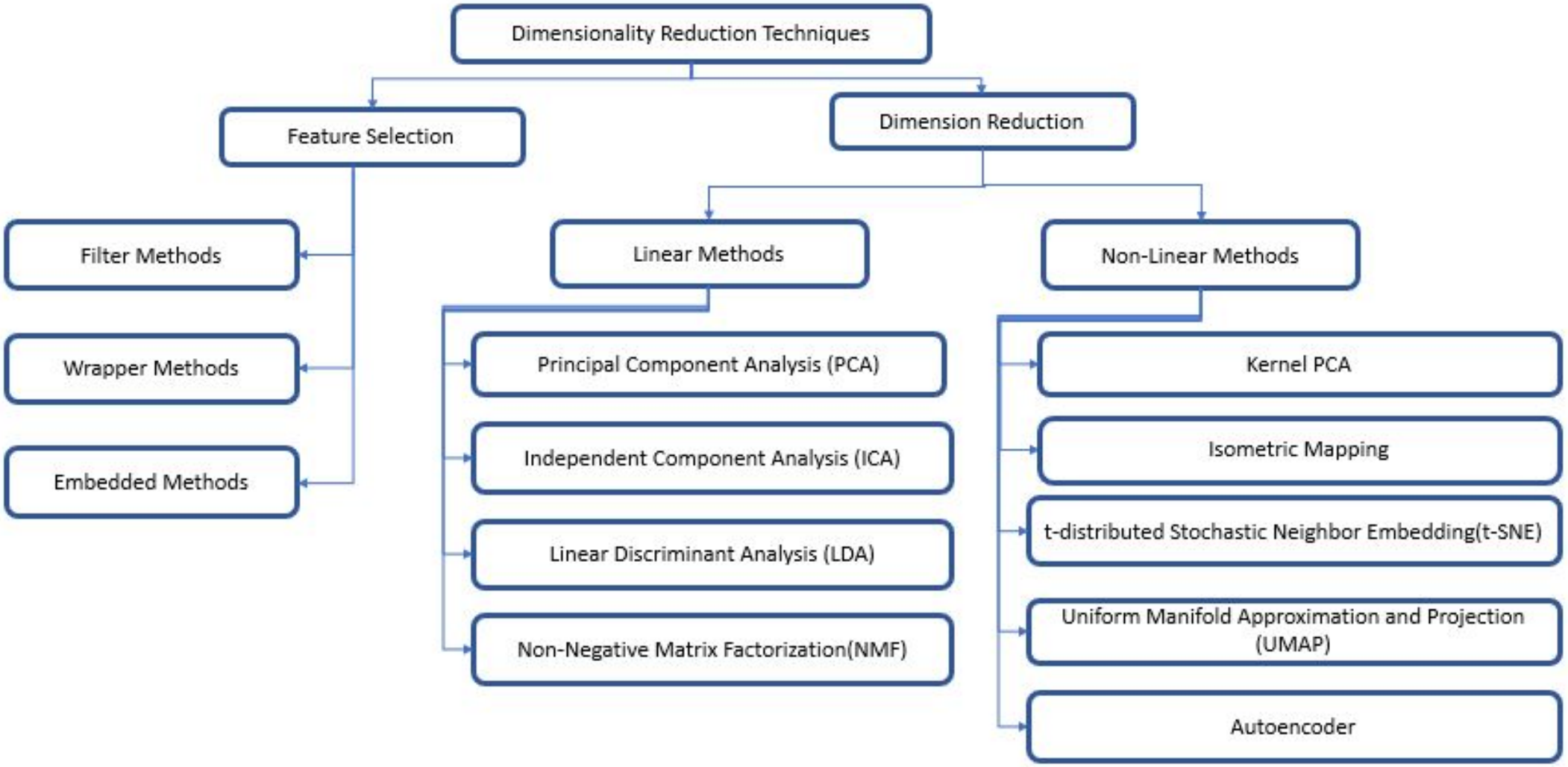
- Menos atributos se traduce en menor complejidad
- Reducción de “ruido” o redundancias, lo que reduce el overfitting
- Reducción de espacio de almacenamiento (data compression)
- Menores tiempos de cómputo
- La precisión de los modelos puede aumentar al eliminar datos irrelevantes/engañosos
- Mejora la visualización de los datos

Desventajas

- Podemos perder información, potencialmente afectando el análisis
- Puede requerir mucho poder de cómputo, dependiendo de la técnica
- La interpretación de atributos transformados puede ser compleja o imposible

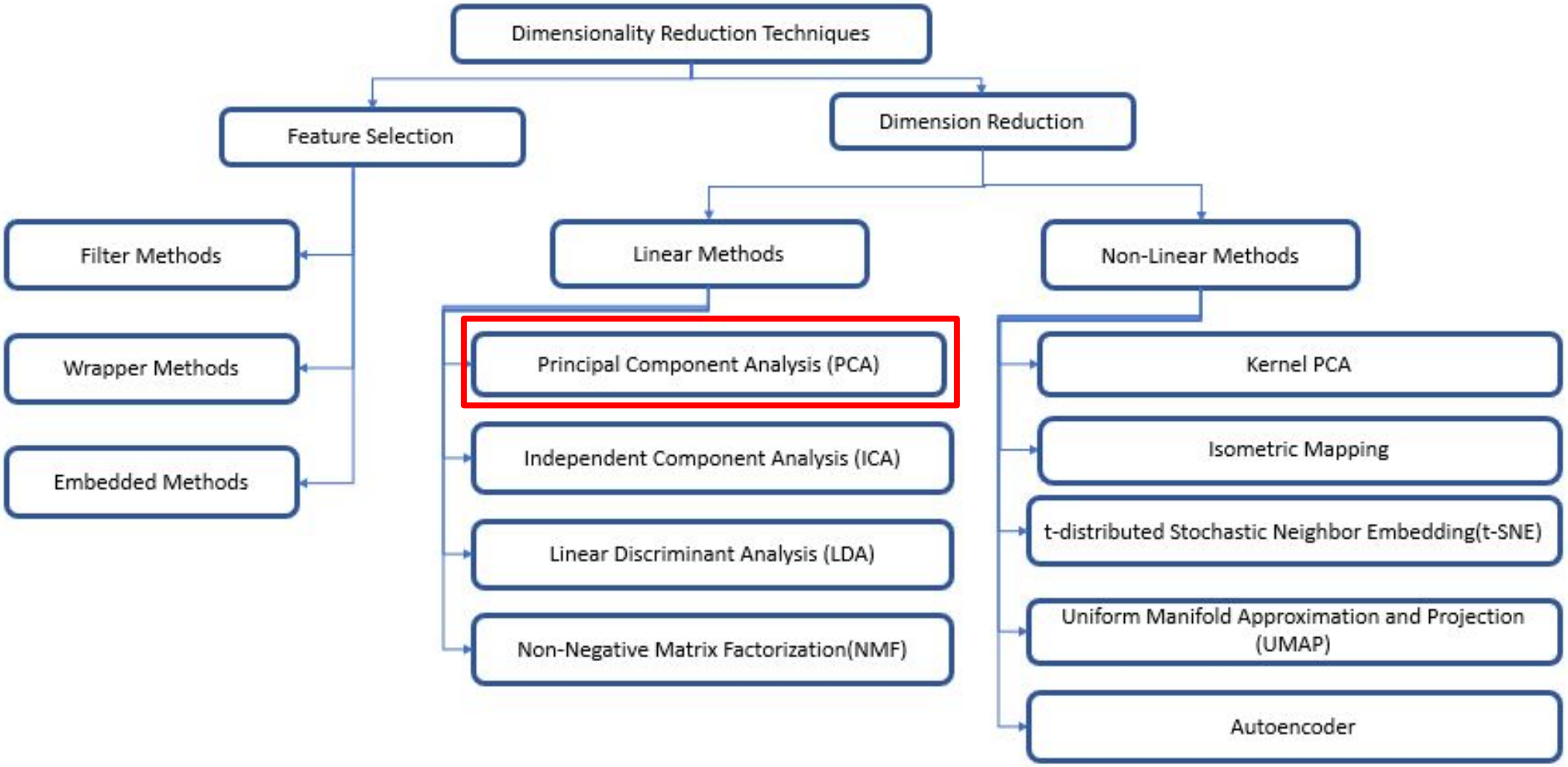
UNAHUR

Técnicas de reducción de la dimensionalidad



UNAHUR

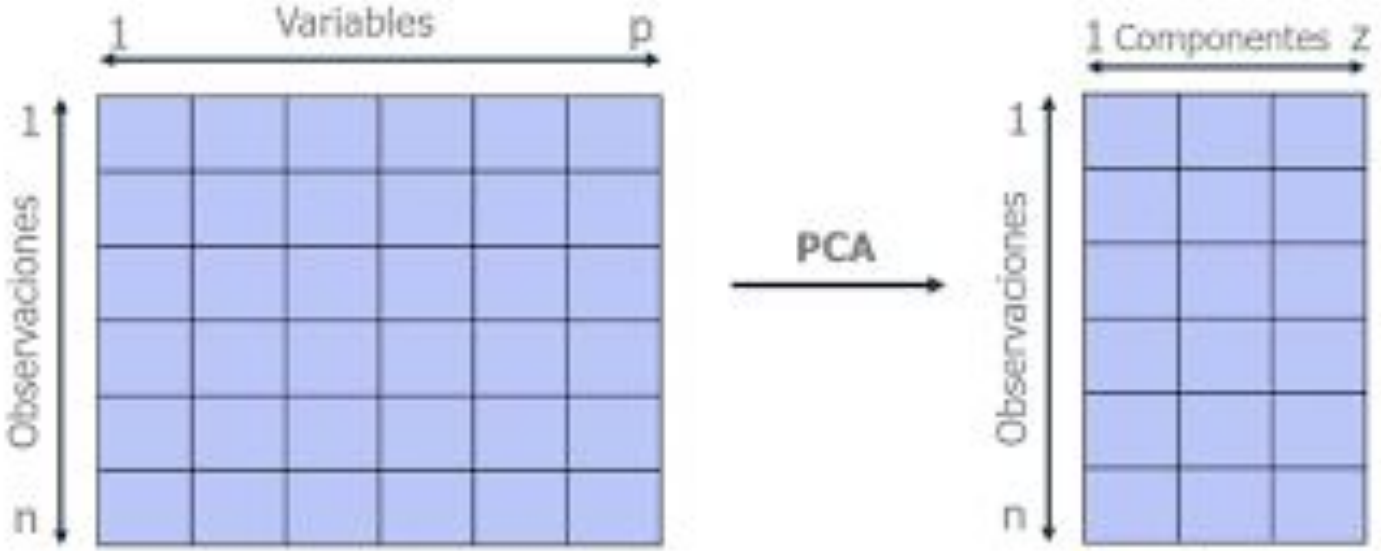
Técnicas de reducción de la dimensionalidad



UNAHUR

Análisis de Componentes Principales (PCA)

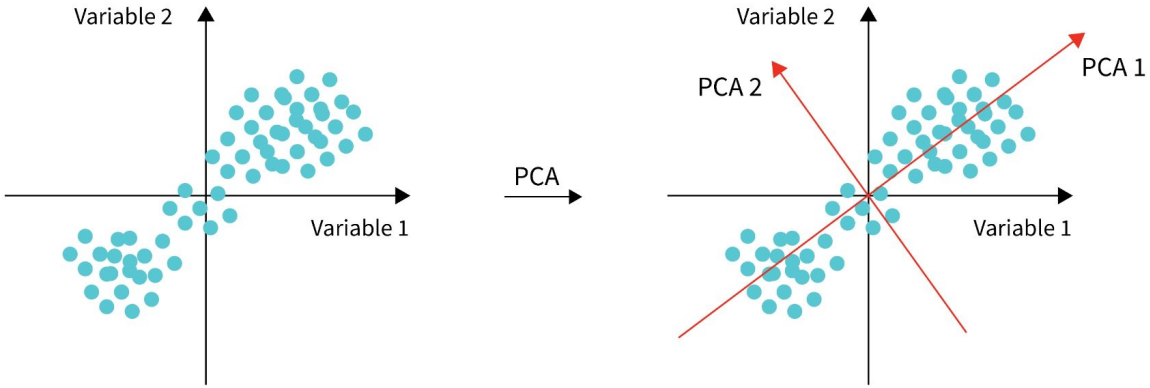
Principal Component Analysis (PCA) es un método estadístico que permite simplificar la complejidad de los datos con muchas dimensiones a la vez que conserva su información. Supóngase que existe una muestra con n individuos cada uno con p variables. PCA permite encontrar un número de factores subyacentes (z < p) que explican aproximadamente lo mismo que las p variables originales. Donde antes se necesitaban p valores para caracterizar a cada individuo, ahora bastan z valores. Cada una de estas z nuevas variables recibe el nombre de componente principal.

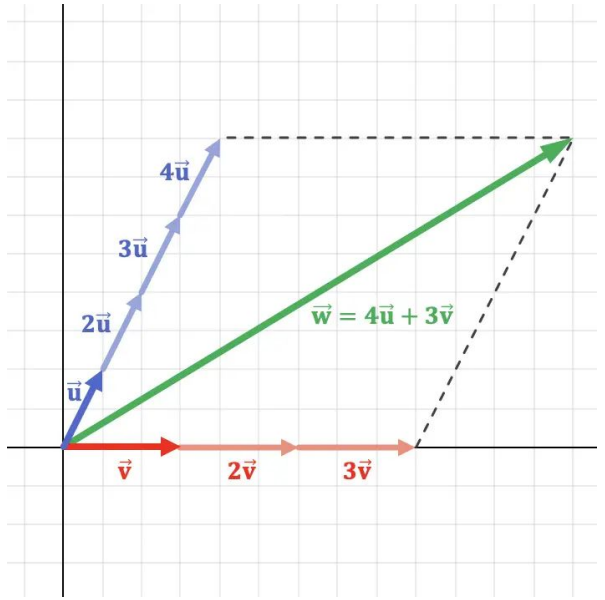


https://cienciadedatos.net/documentos/py19-pca-python

Análisis de Componentes Principales (PCA)

- PCA es una técnica muy útil si existen atributos que se encuentran correlacionados. ¿Por qué? Porque en esos casos se podrá encontrar una única variable que logre resumir la información que contenían las variables originales que estaban correlacionadas.
- Los componentes principales (PC) son una combinación lineal de los atributos originales.
- Cada PC se caracteriza por el porcentaje total de la varianza de los datos que logra capturar. Según este valor, se ordenan de mayor a menor. Por lo que el PC1 es el componente principal que mayor varianza explica, el PC2 el segundo y así…



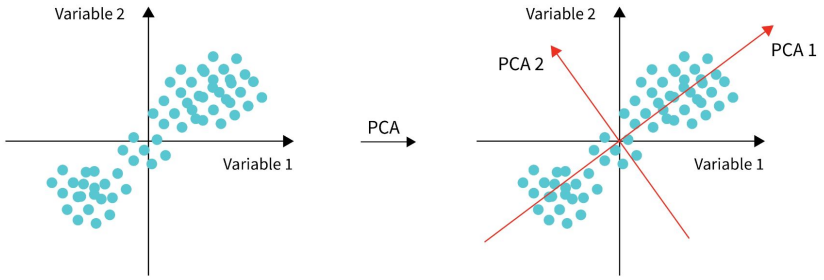


https://cienciadedatos.net/documentos/py19-pca-python

Análisis de Componentes Principales (PCA)

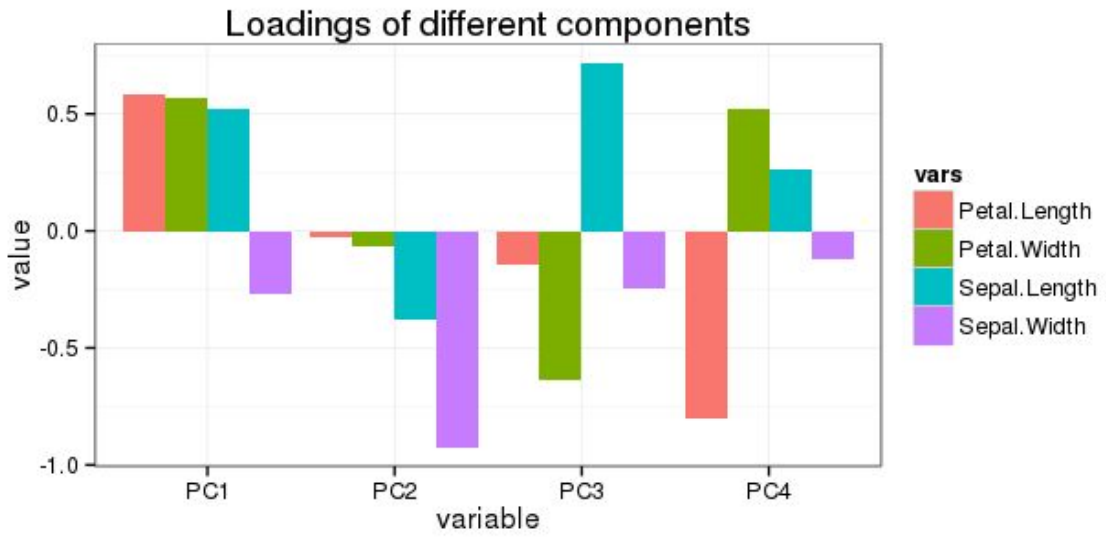
Receta para realizar un PCA

- Estandarización de datos: Antes de aplicar PCA, es necesario estandarizar los datos para que todas las variables tengan una media de 0 y una varianza de 1.
- Cálculo de la matriz de covarianza: La matriz de covarianza es una matriz cuadrada que contiene las covarianzas entre todas las posibles parejas de variables del conjunto de datos. Esta matriz ayuda a identificar las relaciones lineales entre las variables.
- Cálculo de los autovalores y autovectores: Se determinan los autovalores y autovectores de la matriz de covarianza. Los autovectores representan las direcciones de los componentes principales, mientras que los autovalores indican la cantidad de varianza que cada componente principal puede explicar.
- Selección de los componentes principales: Se ordenan los autovalores de mayor a menor y se seleccionan los primeros k componentes principales, donde k es el número de componentes que se desea mantener.
- Proyección de los datos: Finalmente, los datos originales se proyectan sobre los k componentes principales seleccionados, creando un nuevo conjunto de datos de menor dimensión.



Análisis de Componentes Principales (PCA)

- Se puede medir la importancia de cada una de las variables originales en cada una de las componentes principales. A esto se llama **loadings.**
- ¿Con cuántos componentes principales me quedo? La pregunta de siempre… Como en otros casos, hay varios métodos para determinar el número óptimo de PCs, uno de los más usados es el método del codo.
- Si el objetivo del PCA es poder visualizar los datos no nos va a quedar otra opción que quedarnos con las primeras 2 o 3 componentes principales.

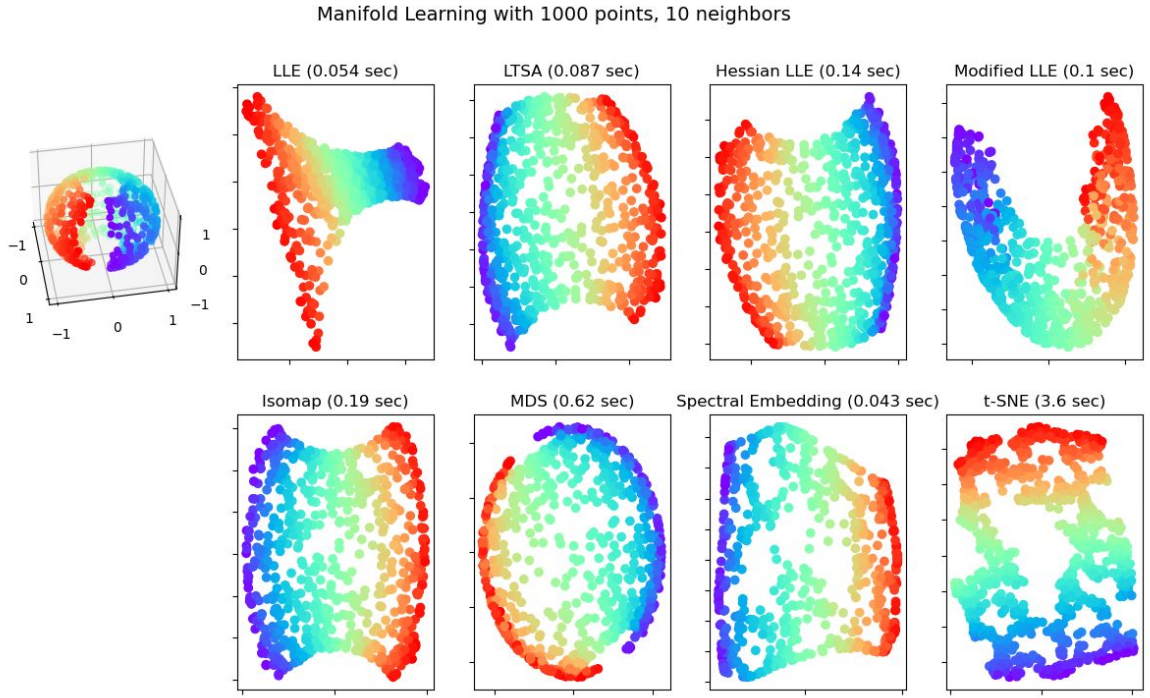


Manifold Learning

No todo en la vida es lineal

PCA es una buena técnica de reducción de dimensionalidad si hay una relación lineal entre las variables, si éste no es el caso vamos a tener que buscar alternativas…

Manifold Learning es un conjunto de técnicas de reducción de dimensionalidad que no requieren de linealidad en la relación que mantienen los datos. En el ejemplo de abajo, los datos originales están distribuidos en forma esférica.



https://scikit-learn.org/stable/auto_examples/manifold/plot_manifold_sphere.html