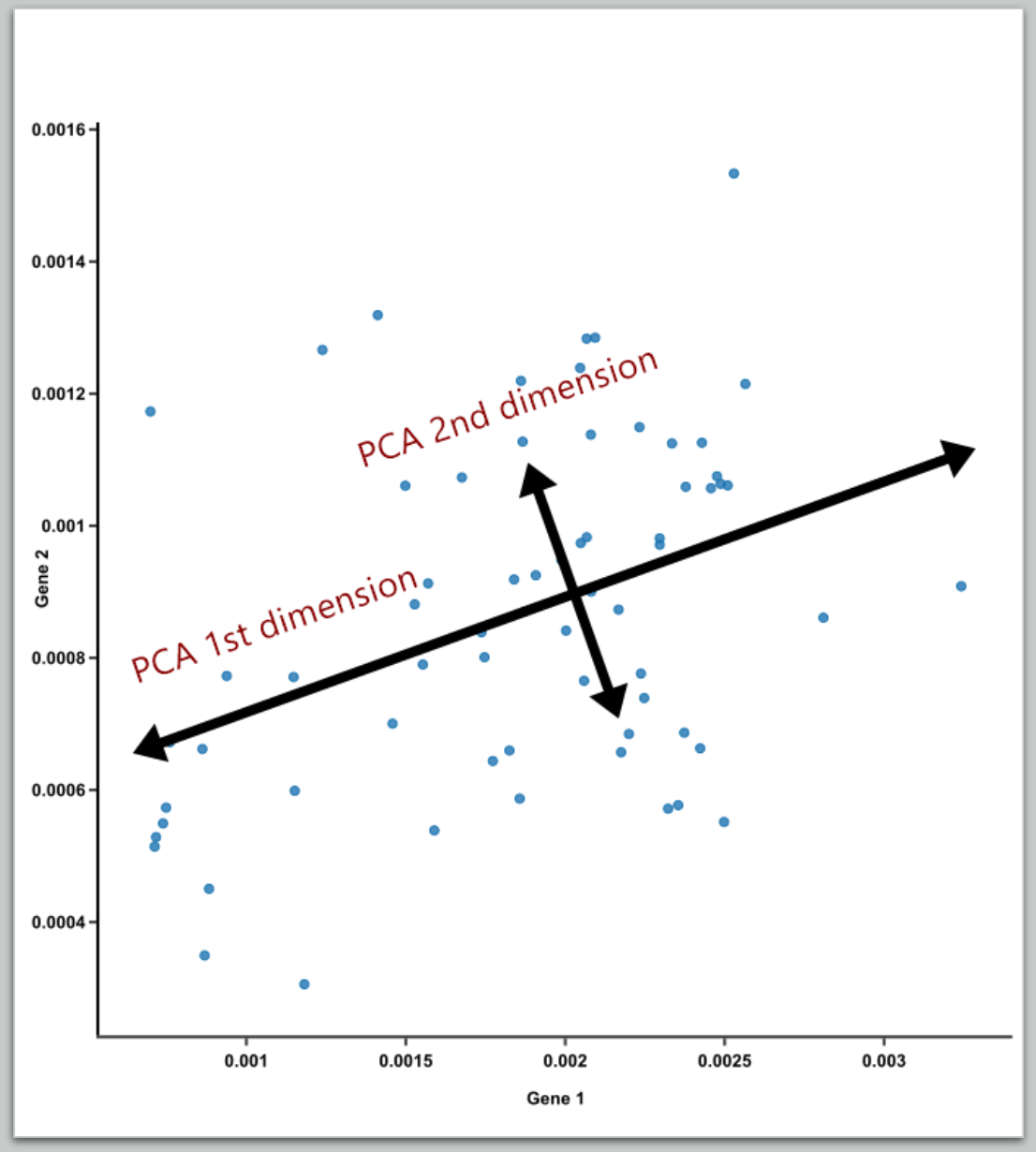
Machine Learning

PCA

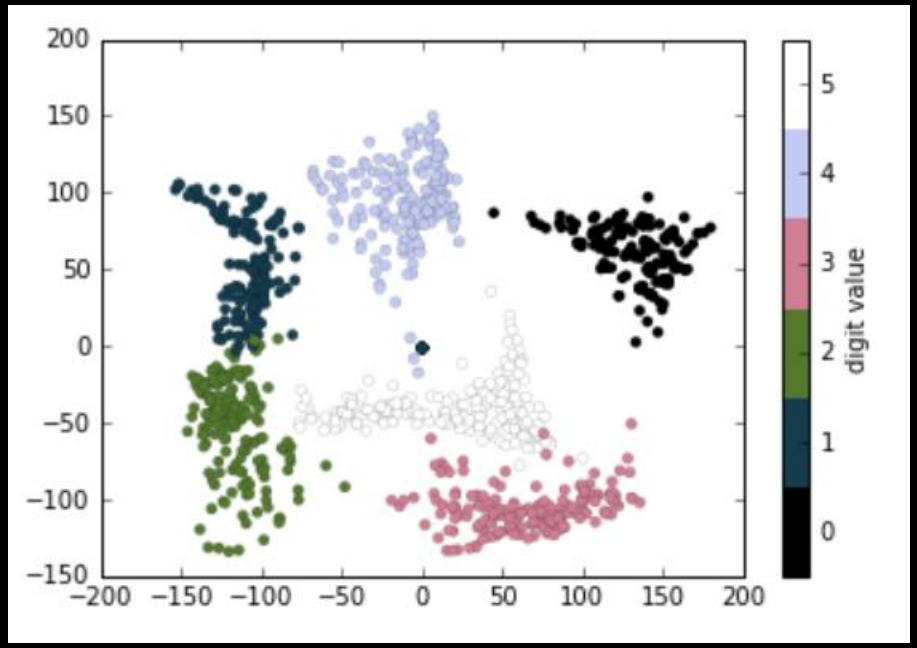
¿Qué es el PCA?

- Principal Component Analysis (PCA)
- Método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información.
- Supóngase que existe una muestra con n individuos cada uno con p variables (X1, X2, …, Xp), es decir, el espacio muestral tiene p dimensiones. PCA permite encontrar un número de factores subyacentes (z<p) que explican aproximadamente lo mismo que las p variables originales.
- Cada una de estas z nuevas variables recibe el nombre de componente principal.



¿Para qué se usa PCA?

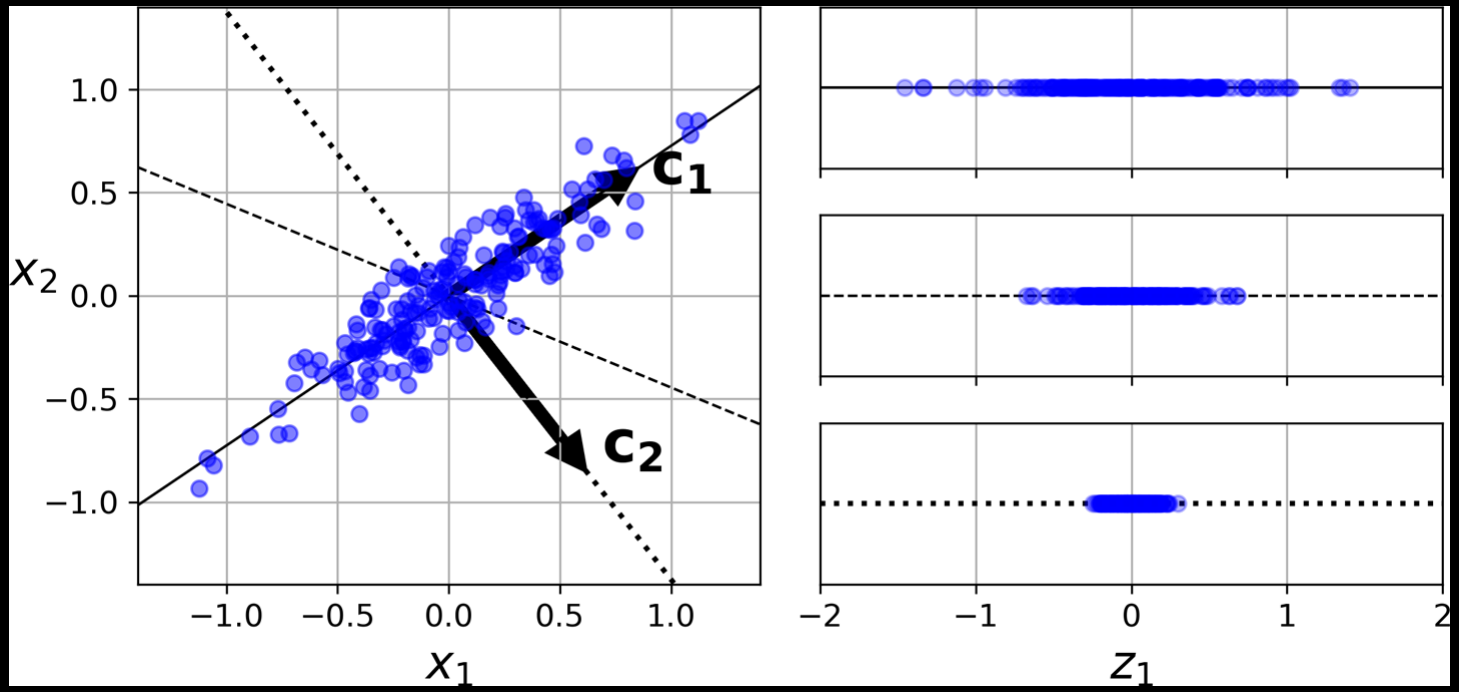
1. **Extracción de patrones en los datos**: convierte el dataset para ver similitudes y diferencias en los datos.
2. **Feature reduction**: Comprimir la información de un dataset en menos variables. Imprescindible con datasets de cientos o miles de features.
3. **Visualización para clasificación**: Datasets de más de tres variables son imposibles de representar en una gráfica. Con PCA podemos.



¿Qué es el PCA?

Normalmente se usa para reducir la dimensionalidad. Identifica los hiperplanos que maximizan la varianza y proyecta los datos en esos hiperplanos, de tal manera que minimicemos la pérdida de información.

Las líneas discontinuas de la siguiente imagen serían una proyección con muy poca varianza de la variable.



Algunos conceptos matemáticos

Matriz de Covarianza

La covarianza es el valor que refleja en qué cuantía dos variables aleatorias varían de forma conjunta respecto a sus medias.

Nos permite saber cómo se comporta una variable en función de lo que hace otra variable. Es decir, cuando X sube ¿Cómo se comporta Y? Así pues, la covarianza puede tomar los siguiente valores:

Covarianza (X,Y) es menor que cero cuando “X” sube e “Y” baja. Hay una relación negativa.

Covarianza (X,Y) es mayor que cero cuando “X” sube e “Y” sube. Hay una relación positiva.

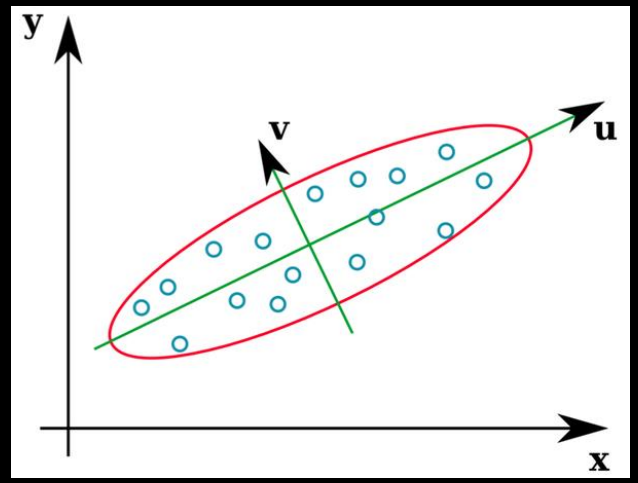
$$Cov(X,Y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Autovalores y autovectores

En álgebra lineal, los vectores propios, eigenvectores o autovectores de un operador lineal son los vectores no nulos que, cuando son transformados por el operador, dan lugar a un múltiplo escalar de sí mismos, con lo que no cambian su dirección. Este escalar lambda recibe el nombre valor propio, autovalor o valor característico

1. La matriz sobre la que se calculan los autovalores y autovectores tiene que ser cuadrada.
2. Hay tantos autovalores como dimensiones tenga la matriz. Se pueden repetir.
3. Todos los autovectores son perpendiculares entre sí
4. La longitud del autovector es 1 y su autovalor representa el poder de cada autovector.

$$A.v = \lambda.v \quad , \quad v \neq 0_V$$



[Explicación detallada de cálculo de autovalores y autovetores]

PCA

Cálculo de los Principal Components

Para calcular el PCA necesitamos:

$$\mathbf{V(b)} = \mathbf{V}\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{bmatrix} var(b_0) & cov(b_0, b_1) \\ cov(b_0, b_1) & V(b_1) \end{bmatrix}$$

1. Obtener la matriz de covarianza de nuestros datos.
2. Descomponer la matriz de en sus autovalores y autovectores.
3. Obtenemos un ranking de vectores, ordenando los autovalores de mayor a menor.
4. Conseguimos los Principal Components a partir de las features originales y los autovectores.

$$A.v = \lambda.v \quad , \quad v \neq 0_V$$

$$newDataset^T = featureVector^T.dataset^T$$

Aquellos valores más altos, son los que representan la mayor varianza de nuestros datos.

Como estamos buscando variables que reduzcan la dimensionalidad, los autovalores de la matriz de covarianza se calculan para obtener patrones (autovalores) con su significado (autovectores). Los autovalores de la matriz de coviarianza serán nuevas variables cuyo poder discriminante serán sus autovectores.

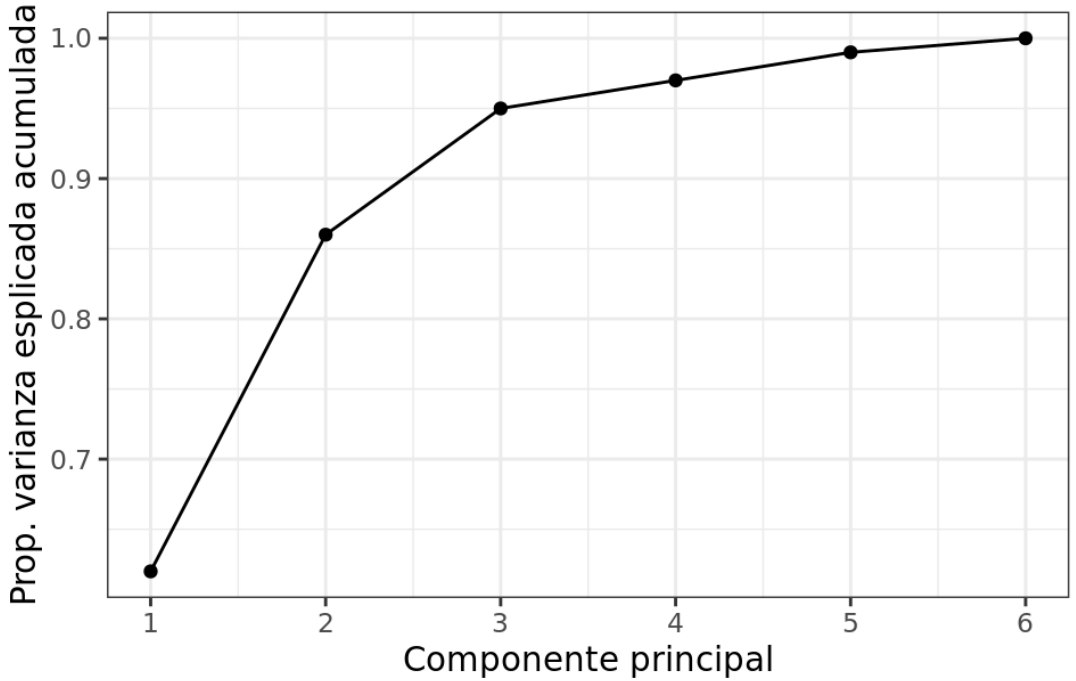
¿Cómo funciona?

Veamos en la siguiente demo cómo trabaja el PCA, de manera gráfica

https://setosa.io/ev/principal-component-analysis/

A TENER EN CUENTA

- Escalado de variables: PCA identifica direcciones cuya varianza es mayor. Por ello deberemos tener los datos en la misma escala. **StandardScaler**.
- Influencia de outliers: al trabajar con varianzas, PCA es altamente sensible a outliers. Es muy recomendable estudiar si los hay.
- ¿Cuánta información presente en el set de datos original se pierde al proyectar las observaciones en un espacio de menor dimensión? (Varianza explicada de cada componente principal).
- Es de interés utilizar el número mínimo de componentes que resultan suficientes para explicar los datos.



Bibliografía

https://learning.oreilly.com/library/view/hands-on-machine-learning/9781492032632/ch08.html#idm45022161212088

https://medium.com/@kyasar.mail/pca-principal-component-analysis-729068e28ec8

https://setosa.io/ev/principal-component-analysis/