Reducción de Atributos

Jhon Jairo Padilla Aguilar, PhD.

El problema

- Hemos visto que la cantidad de atributos (features) disponibles en un conjunto de datos condiciona la construcción de buenos modelos.
- Para poder manejar esta situación, se puede recurrir a la reducción de la dimensionalidad, escogiendo un subconjunto de atributos a partir del conjunto original de atributos.
- Para lograr este objetivo, existen diferentes abordajes posibles.

Por qué reducir la dimensionalidad?

- Es fácil recolectar información y almacenarla
- En un experimento los datos no se recolectan solo para datamining, y además, se acumulan con gran velocidad.
- El preprocesamiento de datos es una parte importante para machine learning efectivo
- La reducción de dimensionalidad es un abordaje efectivo para el “downsizing” de los datos

Por qué reducir la dimensionalidad?

- La mayoría de las técnicas de machine learning no pueden ser muy efectivas si tenemos datos altamente dimensionales
- Maldición de la dimensionalidad:
  - La exactitud y la eficiencia para resolver consultas se degradan rápidamente a medida que aumenta la cantidad de dimensiones.
- La dimensión intrínseca puede ser pequeña.
- Ej: El número de genes responsables por un cierto tipo de enfermedad puede ser pequeño

Por qué reducir la dimensionalidad?

- Visualización: proyección de datos altamente dimensionales en 2D o 3D.
- Compresión de datos: lograr almacenamiento y recuperación eficiente.
- Remoción de ruido: se logra así un efecto positivo en resolver las consultas.

Principal Components Analyisis (PCA)

- PCA : usado para reducir dimensiones de los datos sin mucha pérdida de información (suposición fuerte: correlación lineal).
- PCA es “una transformación lineal ortogonal que transfiere los datos a un nuevo sistema de coordenadas, de forma tal que la mayor varianza de cualquier proyección de los datos está asociada a la primer coordenada (first principal component), la segunda mayor varianza está asociada a la segunda coordenada (second principal component), y así sucesivamente.”

Principios básicos

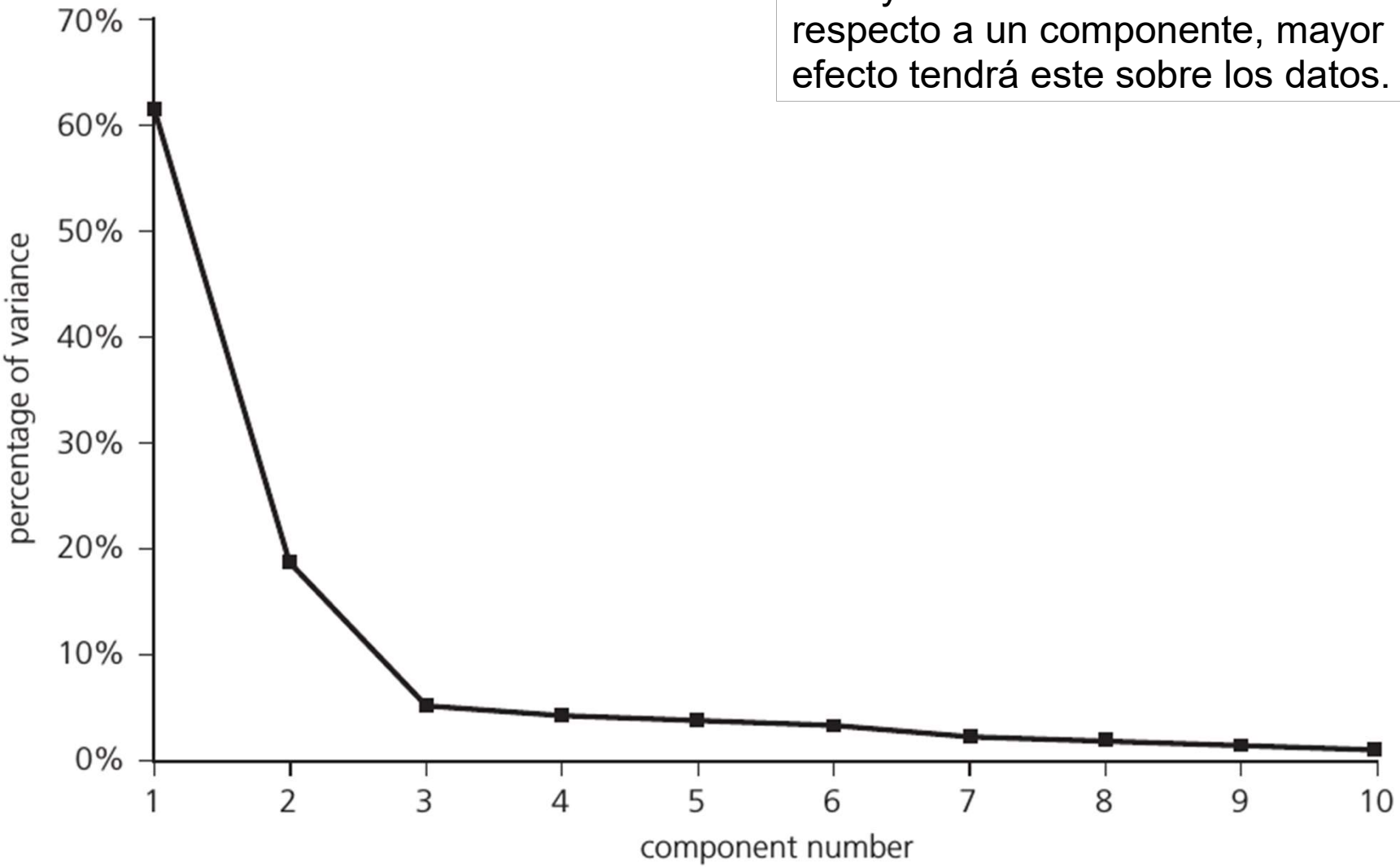
- Supongamos que los atributos son A1 y A2, y tenemos n ejemplos de entrenamiento. Los x's denotan valores de A1 y los y's denotan valores de A2 asociados a los ejemplos de entrenamiento.
- La varianza de un atributo mide la dispersión de sus valores respecto a la media.
- La covarianza contrasta dos atributos: si la covarianza es positiva, ambas dimensiones/atributos son directamente proporcionales. Si es negativa, son inversamente proporcionales.
- Covarianza Cero = los atributos son independientes entre sí.

$$\mathrm{var}(A_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}$$

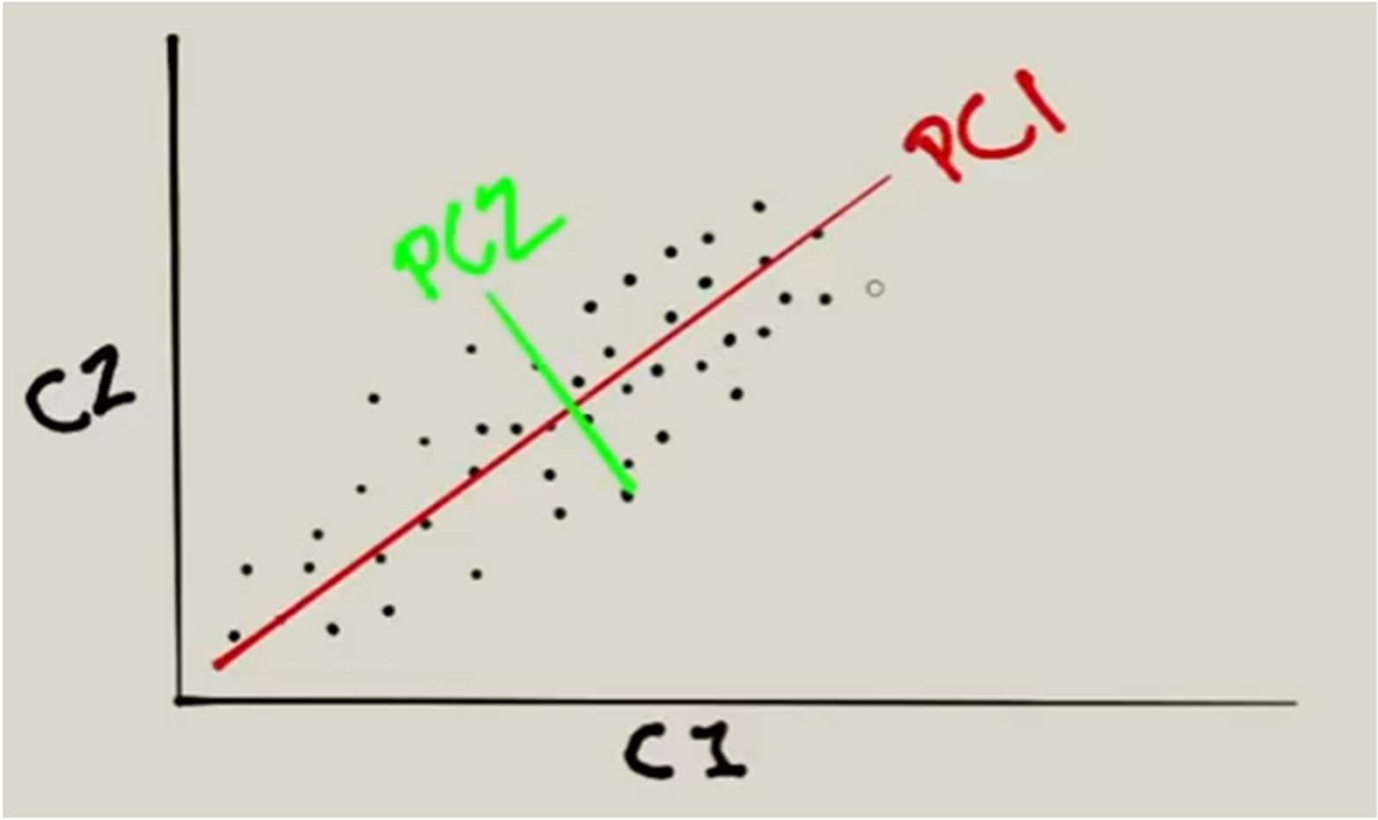
$$\mathrm{cov}(A_1, A_2) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

Varianza de cada componente

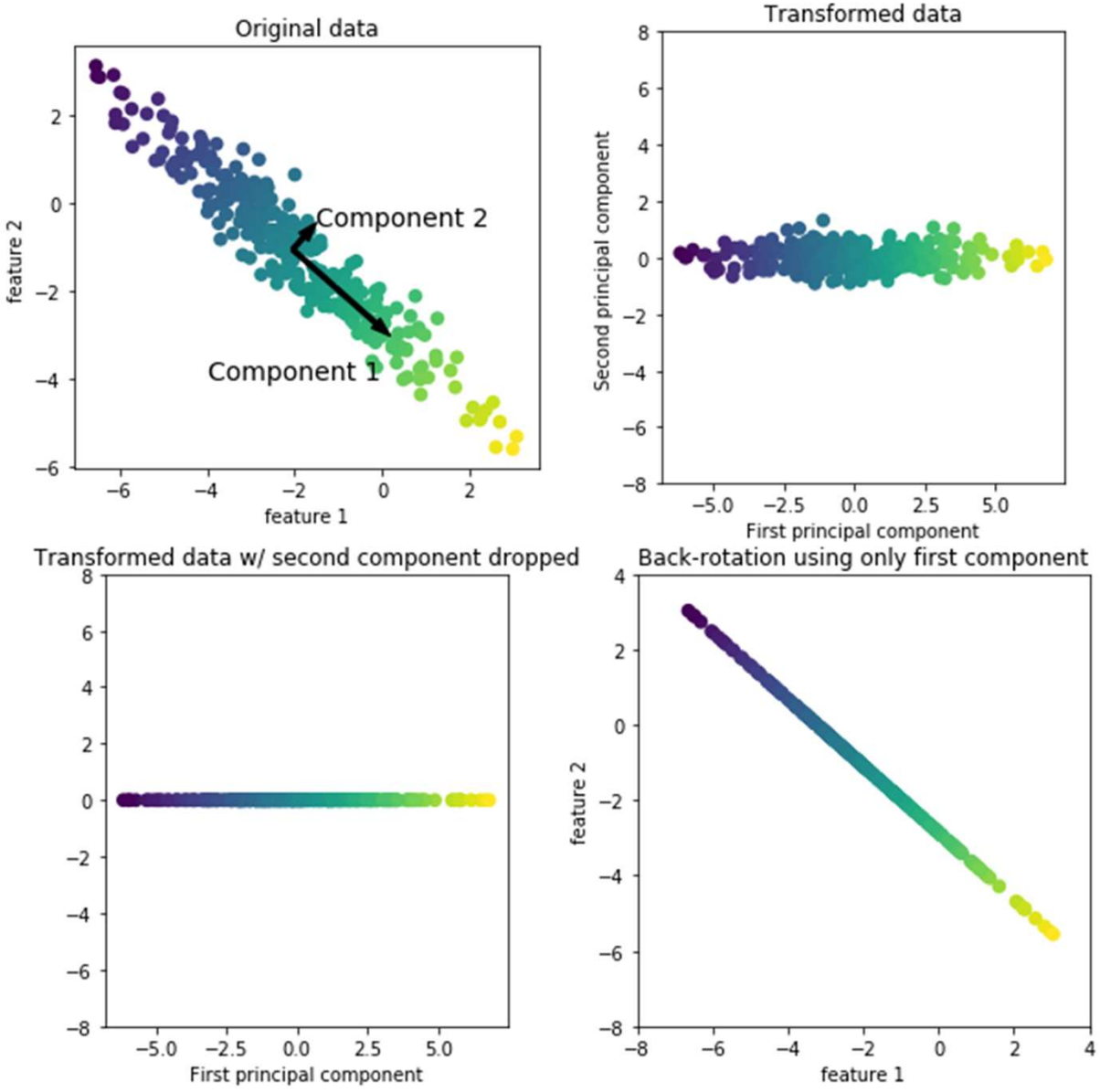
A mayor varianza de los datos con respecto a un componente, mayor efecto tendrá este sobre los datos.



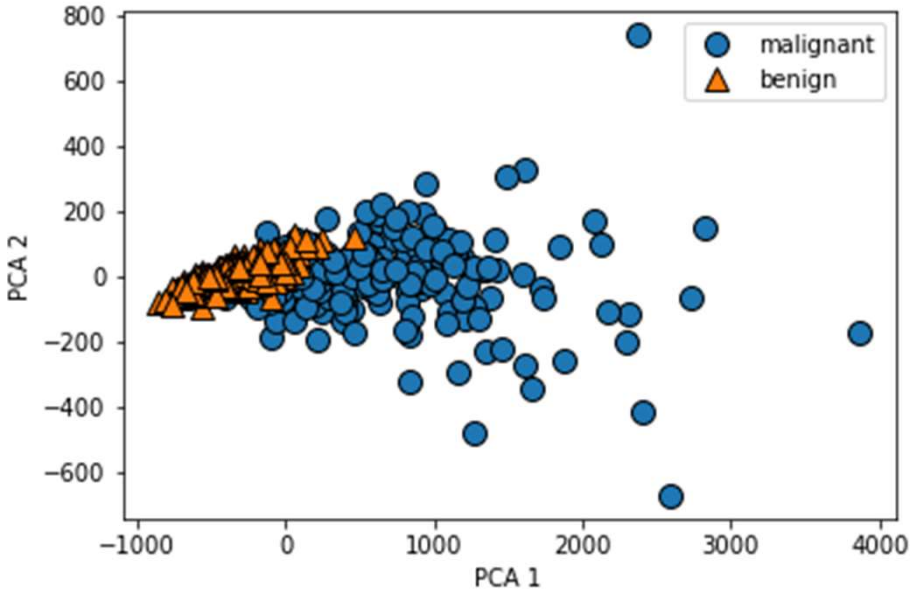
Principio básico para elección de Componentes principales



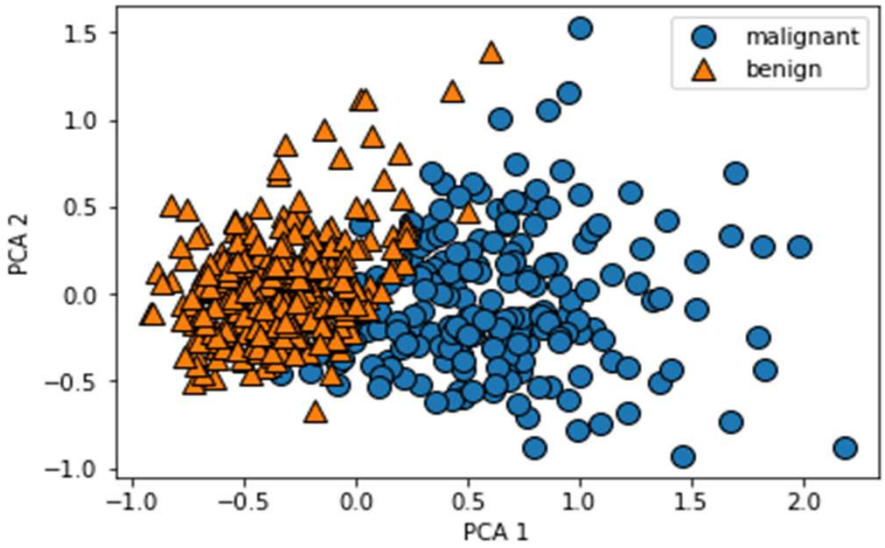
Cómo opera PCA?



Escalado de los datos antes de aplicar PCA



Datos sin Escalar



Datos Escalados previamente