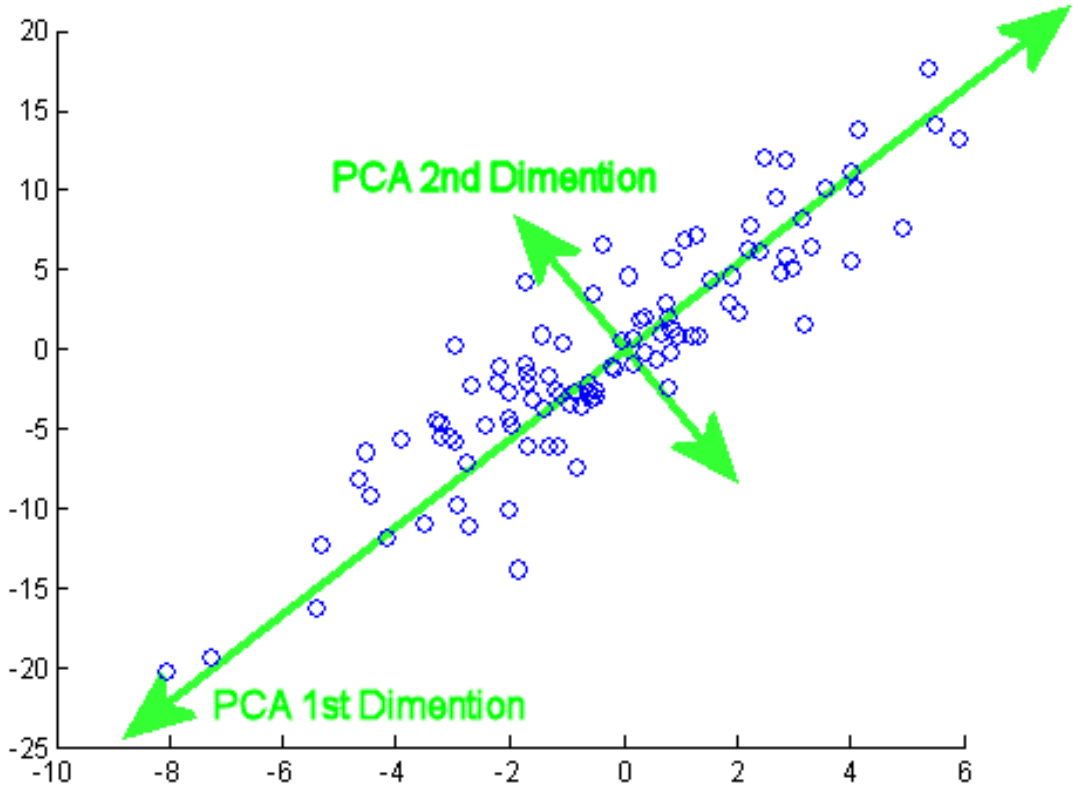


Análisis de componentes principales - PCA

Jonnatan Arias Garcia

Análisis de Componentes Principales PCA

Es una técnica estadística usada para simplificar la complejidad de los datos de alta dimensión y, a su vez, conservar las tendencias y los patrones.

Para ello, transforma los datos en menos dimensiones, que actúan como resúmenes de características. Esta transformación se realiza conservando la mayor cantidad posible de variación de los datos.

Analogía:

Imagina que estás en un mercado de frutas y tratas de decidir qué frutas comprar. Puedes considerar una variedad de características, como el color, el tamaño, el dulzor y el precio.

Pero ¿qué pasaría si solo tuvieras unos segundos para elegir? Probablemente priorizarías solo una o dos características clave para tomar una decisión rápidamente.

Análisis de Componentes Principales PCA

La varianza:

Indica cuánta variabilidad total de los datos captura cada componente en un Análisis de Componentes Principales (PCA). Los componentes principales se ordenan de mayor a menor según la **varianza** que explican.

Por ejemplo, en un conjunto de datos sobre vehículos, con características como eficiencia de combustible, velocidad máxima, potencia del motor, precio y prestigio de la marca, el PCA podría revelar que el primer componente principal (quizás una combinación de eficiencia de combustible y precio) explica el 40% de la varianza, mientras que el segundo (quizás velocidad y potencia) explique el 30%. Esto significa que estos primeros componentes capturan la mayor parte de la información relevante del conjunto de datos.

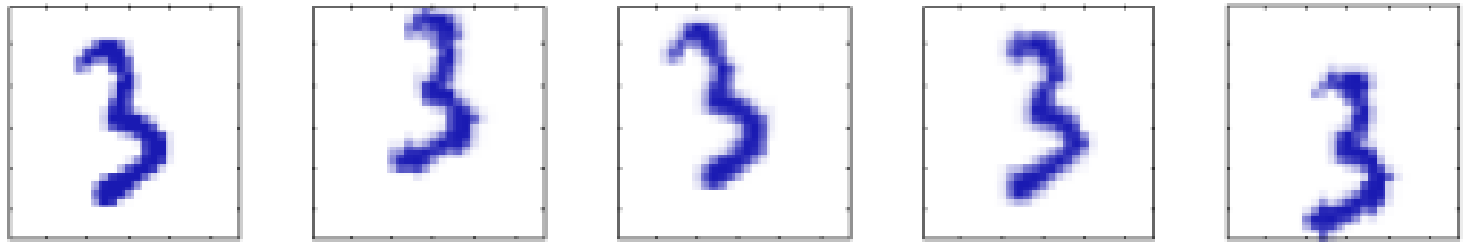
Selección de característica / Extracción I

- La solución a varios problemas en el reconocimiento de patrones se puede lograr eligiendo un espacio de características mejor.

- **Maldición de la dimensionalidad:** el número de ejemplos necesarios para entrenar una función clasificadora crece exponencialmente con el número de dimensiones.

- **¿Qué características caracterizan mejor a la clase?**: qué palabras caracterizan mejor a una clase de documentos.

Selección de característica / Extracción II



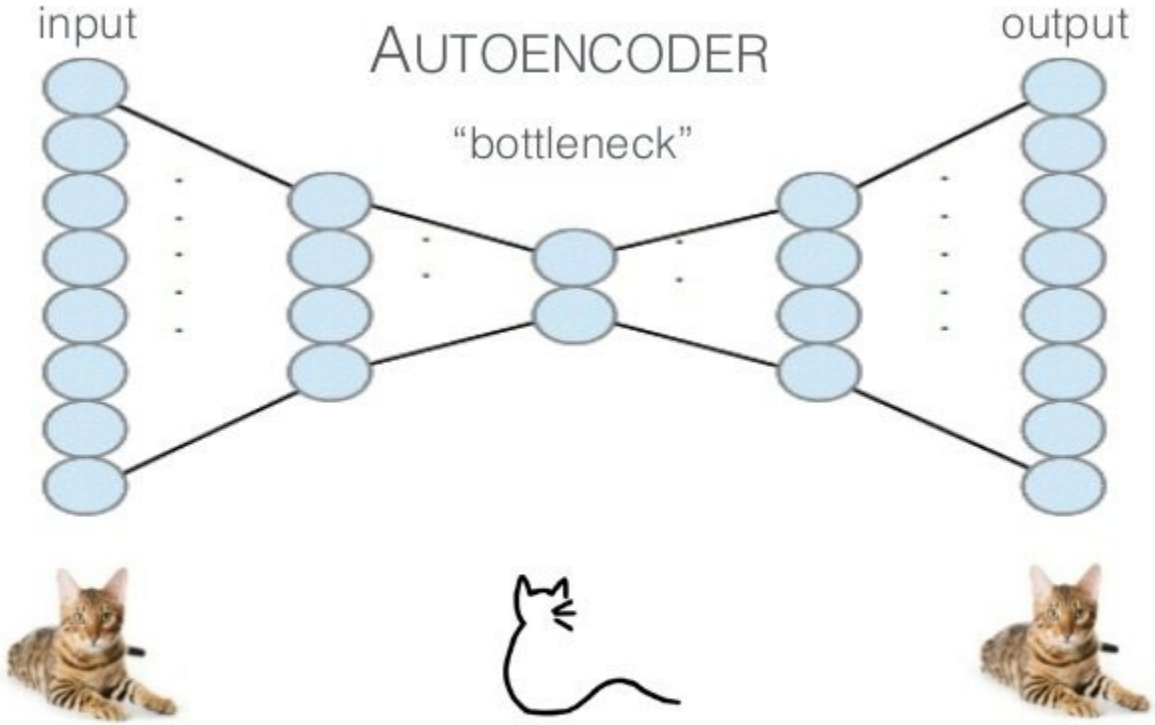
Un conjunto de datos sintéticos obtenido tomando una de las imágenes de dígitos fuera de línea y creando múltiples copias en cada una de las cuales el dígito ha sufrido un desplazamiento y rotación aleatorios dentro de un campo de imagen más grande. Las imágenes resultantes tienen cada una 100 × 100 = 10000 píxeles.

Variable Latente continua

- Espacio latente:

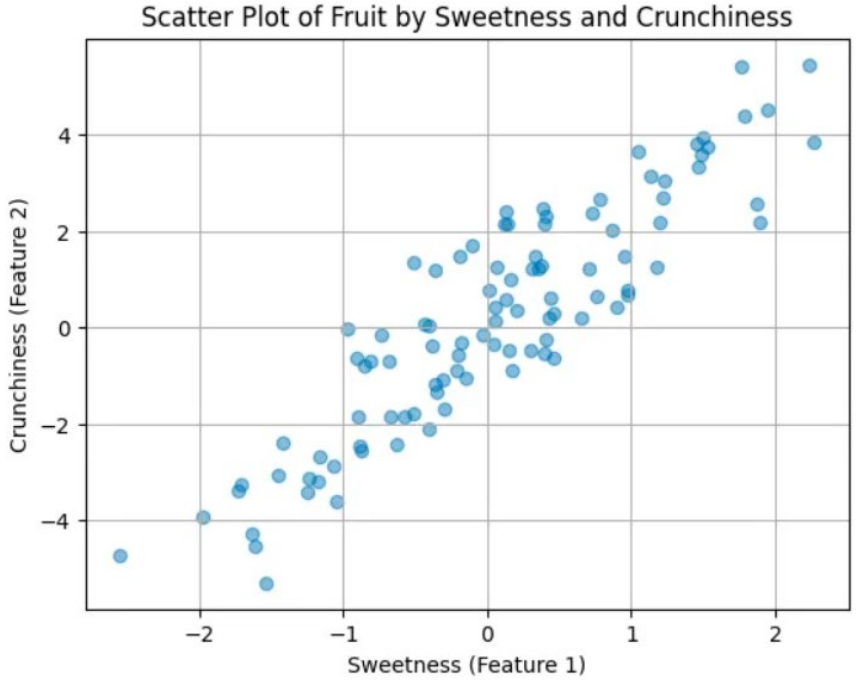
El espacio latente es el espacio en el que se encuentran los datos en la capa de cuello de botella.

El espacio latente contiene una representación comprimida de la imagen, que es la única información que el decodificador puede usar para tratar de reconstruir la entrada con la mayor fidelidad posible



Explicación Base I

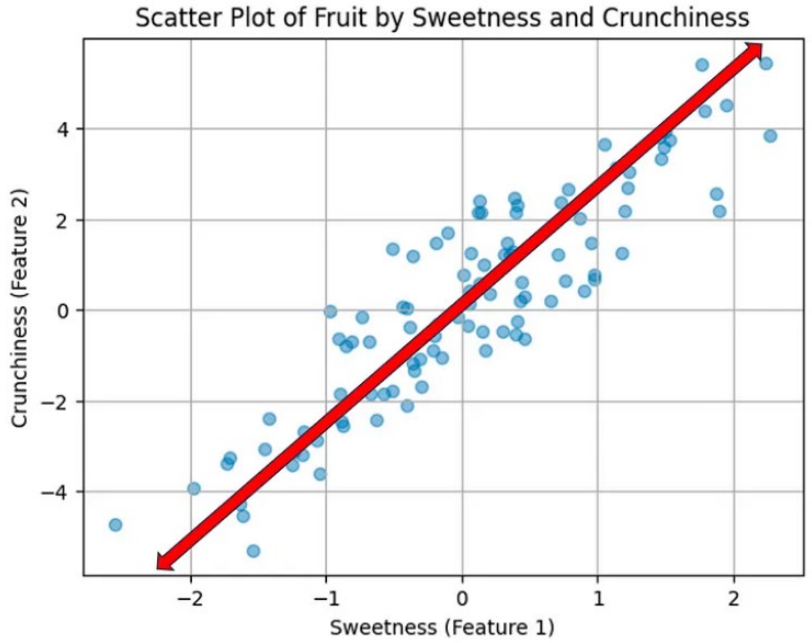
El siguiente grafico muestra una correlación positiva entre características, indicando que la frutas mas dulces tienden a ser las mas crujientes.

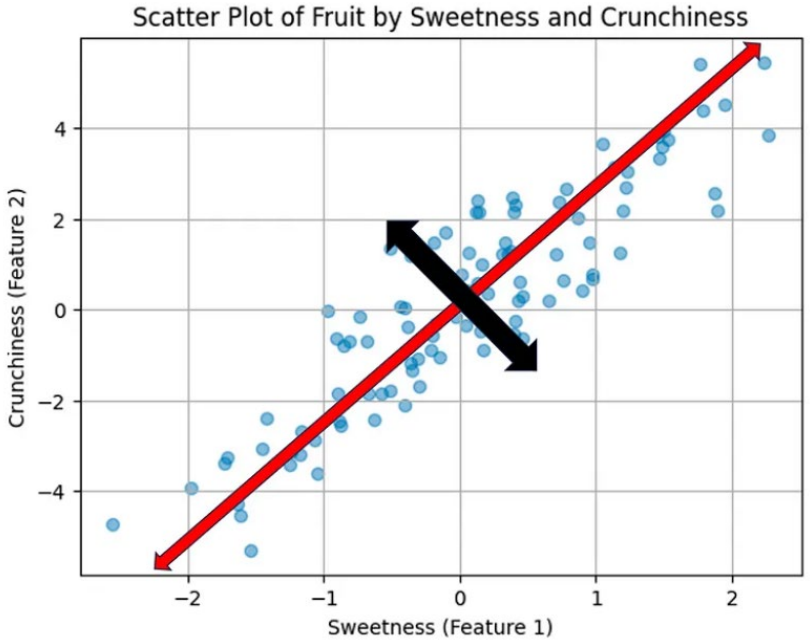


Explicación Base II

Necesitamos encontrar una forma de visualizar y reducir todo esto a una sola características PC1.

Esta línea explica la mayor dispersión del conjunto de datos. Siendo por ejemplo mayor dispersión frente a la línea ortogonal (negra)

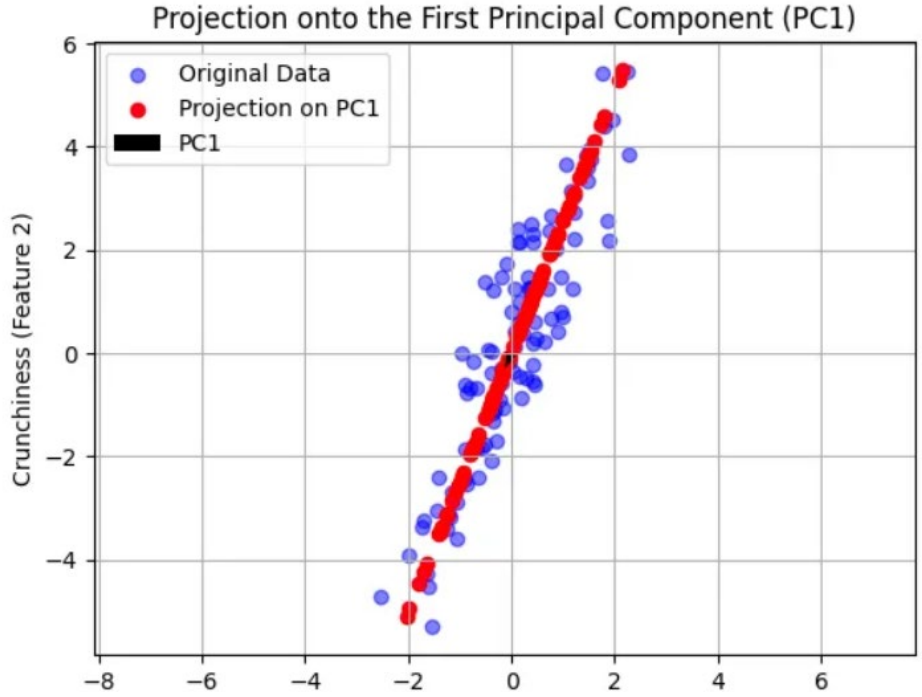


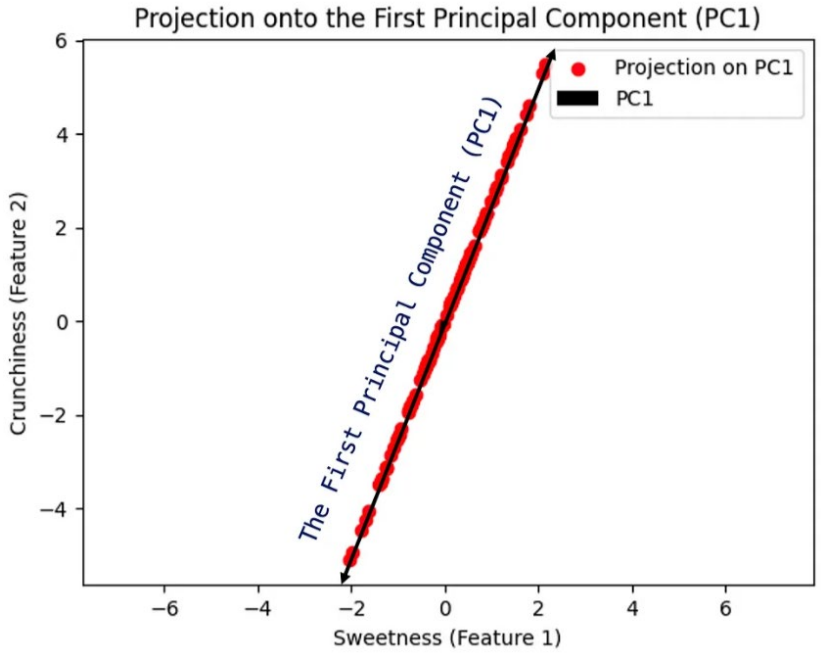


Explicación Base III

Podemos colapsar los datos en una nueva característica, llamada Componente Principal 1 (PC1 es el vector propio)

Si eliminamos los datos originales se vería así:





Explicación Base IV

Los nuevos componentes son combinaciones lineales de las características originales, lo que significa que se construyen como sumas ponderadas de las características originales.

El primer componente principal está alineado con **la mayor varianza** , lo que significa **que captura la mayor dispersión de los datos** . El segundo componente principal captura la mayor varianza posible mientras es ortogonal al primero, y así sucesivamente para los componentes subsiguientes.

Explicación Base V

A menudo no es posible retener toda la información (varianza del 100 %), se pueden extraer conocimientos importantes de los datos si se conservan algunas características importantes que contribuyen en mayor medida a la varianza de los datos.

Esta es la esencia del PCA: encontrar las características más informativas y usarlas para representar el conjunto de datos de manera eficiente.

El modelo I

- PCA se utiliza ampliamente para aplicaciones como la reducción de dimensionalidad, compresión de datos con pérdida, extracción de características y visualización de datos (Jolliffe, 2002). También se conoce como la transformada de Karhunen-Loève.

- Consideremos un conjunto de observaciones $\{\boldsymbol{x}_n\}$ donde n = 1, . . . , N, y $\boldsymbol{x}_n$ es una variable euclidiana con dimensionalidad D.

- Nuestro objetivo es proyectar los datos en un espacio de dimensionalidad M < D maximizando la varianza de los datos proyectados.

El modelo II

- La media de los datos proyectados es

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Y la varianza de los datos proyectados esta dada por:

$$\frac{1}{N}\sum_{n=1}^{N}\left\{u_1^T x_n - u_1^T \bar{x}\right\} = u_1^T S u_1,$$

Donde **S** es la matriz de covarianza de los datos

$$S = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})\text{^}T$$

El modelo III

Introducimos a multiplicador de Lagrange que denotaremos λ_1 y luego hacer una maximización ilimitada de:

$$\boldsymbol{u}_1^T \boldsymbol{S} \boldsymbol{u}_1 + \lambda_1 (1 - \boldsymbol{u}_1^T \boldsymbol{u}_1)$$

Al igualar a cero la derivada con respecto a $\boldsymbol{u}_1$, vemos que esta cantidad tendrá un punto estacionario cuando

$$\boldsymbol{S} \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1$$

Lo que dice que $\boldsymbol{u}_1$ debe ser un vector propio de **S**

El modelo IV

Tip

En resumen, el análisis de componentes principales implica evaluar la media $\overline{\boldsymbol{x}}$ y la matriz de covarianza **S** del conjunto de datos y luego encontrar la **M** vectores propios de **S** correspondientes a los **M** valores propios más grandes λ_i

Algoritmo I

Algoritmo PCA (**X**, k): top k valores propios/vectores propios

DATOS: **X** Matriz de datos m × N

Cada punto $\boldsymbol{x}_n$ es un vector columna de **x**

1. Reste la media de cada columna de **x** (centre los datos)
2. Calcular la matriz de covarianza de **x**
3. Realice la descomposición SVD de **S** de modo que $\{\lambda_i, \boldsymbol{u}_i\}$ son los valores propios y vectores propios de **S**
4. Devuelva los k componentes principales de modo que $k \leq m$(conservando un valor dado del porcentaje de la varianza de los datos)