Exploración y Curación de Datos

EyCD 2024

¿Qué hemos visto en esta materia?

Herramientas para el pre-procesamiento de datos

- Herramientas de estadística descriptiva e inferencial
  - Análisis univariado y multivariado
- Transformaciones de datos: indexado, agrupación y agregación
- Selección de características
- Combinación de conjuntos de datos
- Imputación de valores faltantes
- Detección y corrección de sesgos

Hoy agregamos

- **Codificación de variables categóricas**

Es el proceso de convertir variables categóricas o cualitativas en una forma numérica que pueda ser utilizada en análisis estadísticos o algoritmos de aprendizaje automático.

- **Reducción de dimensionalidad con PCA**

La reducción de dimensionalidad con Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica utilizada para reducir la cantidad de variables o características en un conjunto de datos, conservando al mismo tiempo la mayor parte de la información relevante.

Encodings

Los algoritmos de aprendizaje automático requieren **exclusivamente** datos numéricos

Es necesario transformar nuestras variables categóricas a algún formato numérico

One-hot encoding

Codificación one-hot (binaria): Crea variables binarias (0 o 1) para cada categoría. Se crea una nueva columna para cada categoría posible, y se asigna 1 si la observación pertenece a esa categoría y 0 si no. Por ejemplo, si hay tres categorías (A, B, C), se crearán tres columnas binarias, y una observación tendrá un 1 en la columna correspondiente a su categoría y 0 en las otras dos.

One-hot encoding

Id	Barrio
1	San Vicente
2	Cerro de las Rosas
3	Maipú
4	San Vicente
5	Ituzaingó

Id	Barrio=San Vicente	Barrio=Cerro de las Rosas	Barrio=Maipú	Barrio=Ituzaingó
1	1	0	0	0
2				
3				
4				
5				

One-hot encoding

Id	Barrio
1	San Vicente
2	Cerro de las Rosas
3	Maipú
4	San Vicente
5	Ituzaingó

Id	Barrio=San Vicente	Barrio=Cerro de las Rosas	Barrio=Maipú	Barrio=Ituzaingó
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	1	0	0	0
5	0	0	0	1

The curse of dimensionality

Al codificar los datos de esta manera, generamos vectores esparsos de alta dimensionalidad

- Ocupa mucho espacio en memoria
- Los vectores resultantes son ortogonales.
  - Todos los vectores están a la misma distancia entre ellos (si tienen norma 1)
  - No podemos calcular operaciones como el producto punto.

Es importante tener en cuenta estas desventajas al utilizar la codificación one-hot y evaluar si es la mejor opción en función del contexto y el objetivo del análisis. En algunos casos, pueden ser preferibles otras técnicas de codificación que aborden estas limitaciones, como la codificación ordinal o la codificación de frecuencia.

Encoding ordinal

La codificación ordinal es una técnica utilizada para transformar variables categóricas ordenadas en valores numéricos que preservan el orden y la jerarquía de las categorías. A diferencia de la codificación one-hot, que crea variables binarias independientes para cada categoría, la codificación ordinal asigna un valor numérico a cada categoría en función de su posición en la escala de orden.

Notebook

04_Encodings_TiposDeVariables.ipynb

Reducción de dimensionalidad

Objetivo

Reducir el número de columnas o variables de nuestro conjunto de datos

Conservar la mayor cantidad de información posible

¿Qué técnicas conocemos hasta ahora?

Formalización matemática

Vamos a expresar el conjunto de datos como una matriz X con n filas y m columnas. Cada filas es un vector x_i que habita un espacio matemático con m dimensiones. Cada dimensión corresponde intuitivamente a una columna.

$$X \in \mathbb{R}^{n \times m}; x_i \in \mathbb{R}^m$$

Queremos obtener una nueva matriz Z que tenga la misma cantidad de filas, pero un número de columnas d mucho menor que m.

$$Z \in \mathbb{R}^{n \times d}; d \ll m$$

Principal Component Analysis (PCA)

- Método algebraico (no depende del conocimiento de dominio).
- Calcula un conjunto de direcciones llamadas componentes principales:
  - Son ortogonales (independientes)
  - Están ordenados de acuerdo a la varianza de los datos originales que capturan.
- Se proyecta la matriz X en las direcciones de sus componentes principales
- Se seleccionan las primeras k dimensiones de la nueva matriz proyectada.

El proceso básico del PCA

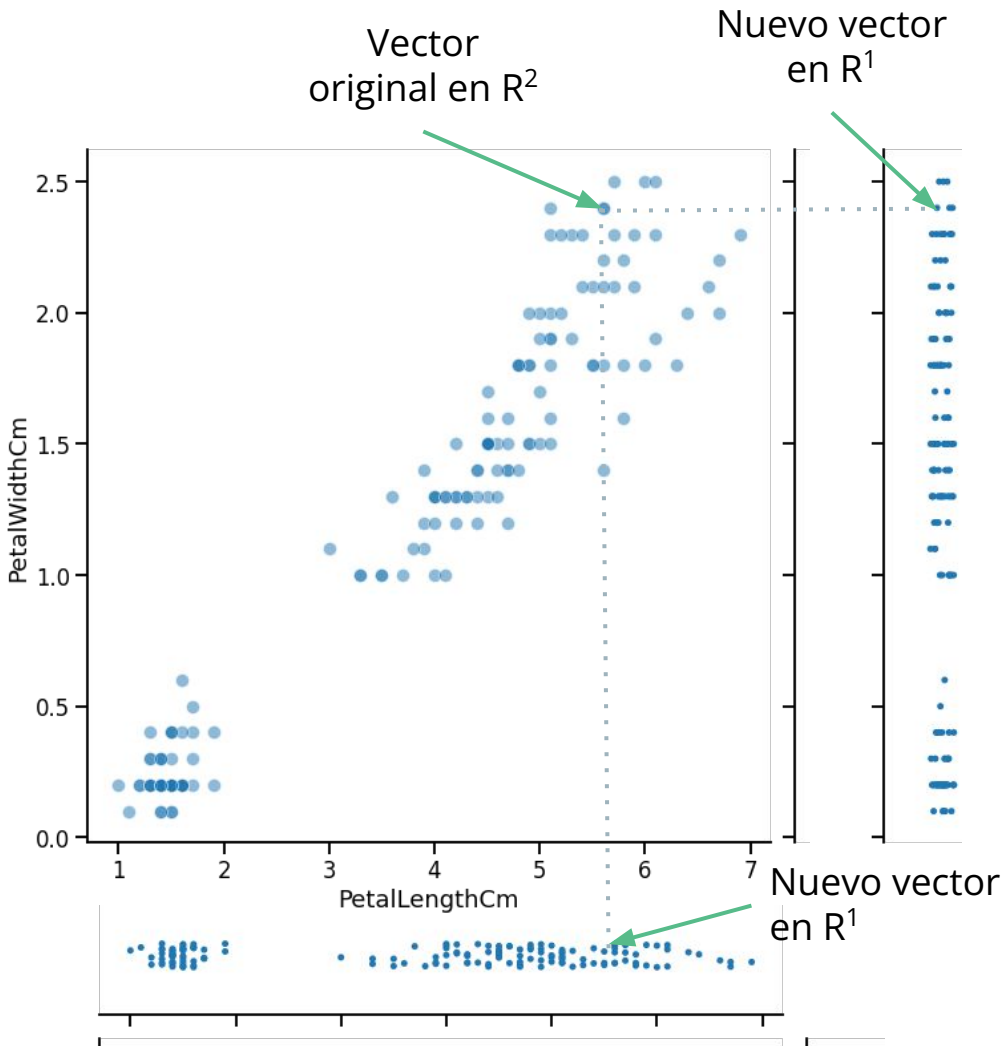
1. **Normalización**: Si las variables en el conjunto de datos tienen diferentes escalas, es importante normalizarlas para que todas tengan la misma importancia durante el análisis.
2. **Cálculo de la matriz de covarianza o matriz de correlación**: Se calcula la matriz que representa las relaciones entre las variables originales.
3. **Cálculo de los componentes principales:** Mediante técnicas matemáticas, se obtienen los componentes principales, que son combinaciones lineales de las variables originales.
4. **Selección de componentes principales:** Se seleccionan los primeros componentes principales que expliquen la mayor parte de la varianza en los datos. Normalmente, se establece un umbral o un porcentaje mínimo de varianza explicada para determinar cuántos componentes principales se retienen.
5. **Transformación de los datos:** Los datos originales se proyectan en el espacio de los componentes principales seleccionados, lo que reduce la dimensionalidad del conjunto de datos

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

Eliminación de columnas

Cada fila es un vector x en R2,
es decir, tiene dos dimensiones.

Si sacamos cualquiera de ellas,
proyectamos los puntos a la
dirección del eje x o y



“Ver” las direcciones ortogonales

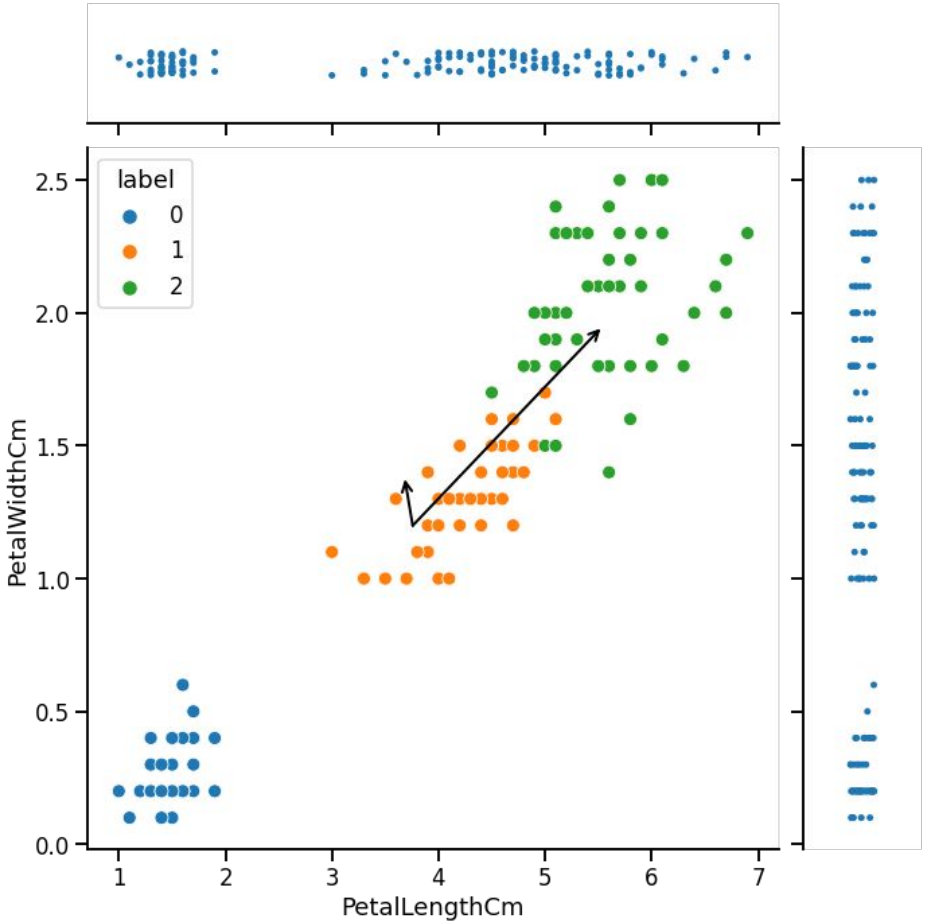
- Si estamos analizando un conjunto de datos multidimensionales y se calcula la matriz de covarianza, "ver" las direcciones ortogonales nos permitirá identificar qué variables están correlacionadas y cuáles no.
- Visualmente, podrías representar cada variable como un eje en un gráfico y observar cómo se relacionan entre sí. Si los ejes son perpendiculares (ortogonales), indicaría que las variables son independientes o no están correlacionadas.
- En el caso de una matriz de transformación, "ver" las direcciones ortogonales significa observar cómo los vectores de entrada se transforman en los vectores de salida.
- Se puede visualizar los vectores en un espacio tridimensional y ver cómo la matriz los rota, escala o traslada. Las direcciones ortogonales serían aquellas que se mantienen perpendiculares después de la transformación.

Componentes principales

Los componentes principales de una matriz son las direcciones ortogonales de mayor variación de los datos.

Esto es útil para identificar componentes o factores independientes en un conjunto de datos.

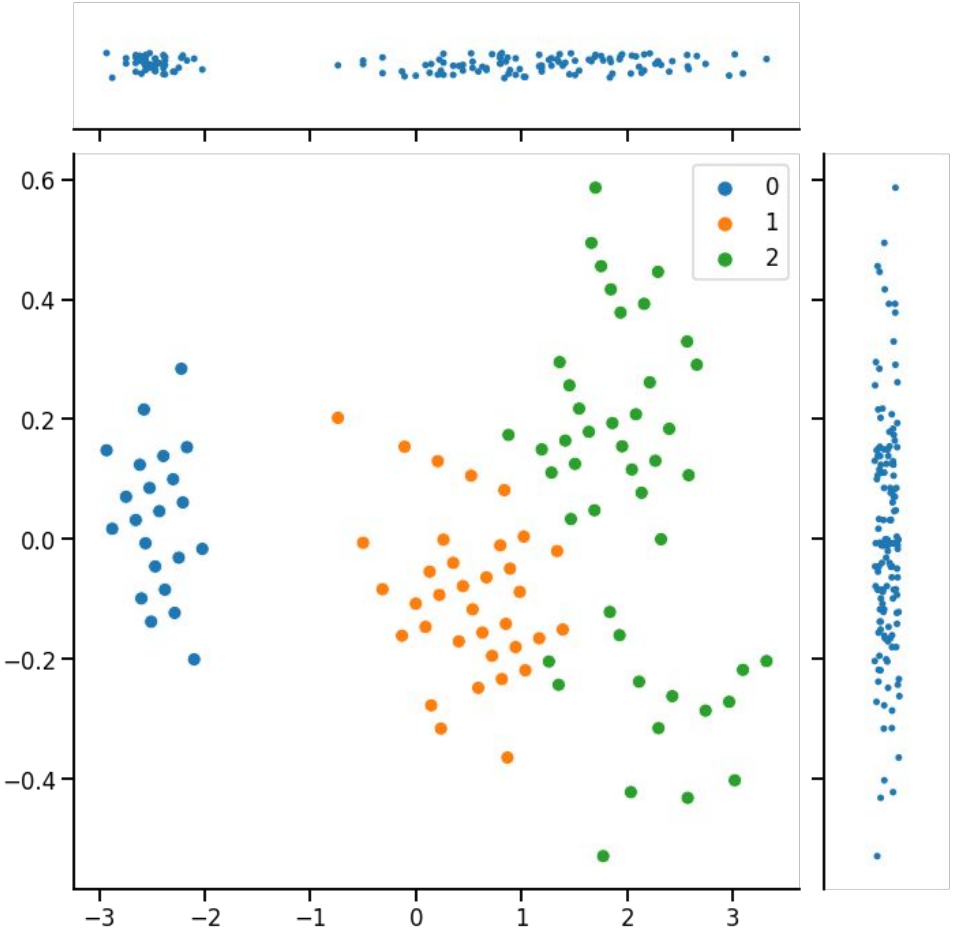
¿Por qué no se “ven” ortogonales?



Nueva proyección

Proyectamos cada una de las filas en las direcciones de los componentes principales.

Tener en cuenta que ambas representaciones de los datos tienen **exactamente la misma información**



Notebook

04 PCA: Ejemplo de juguete.ipynb

Clase 4.3 - Transformaciones

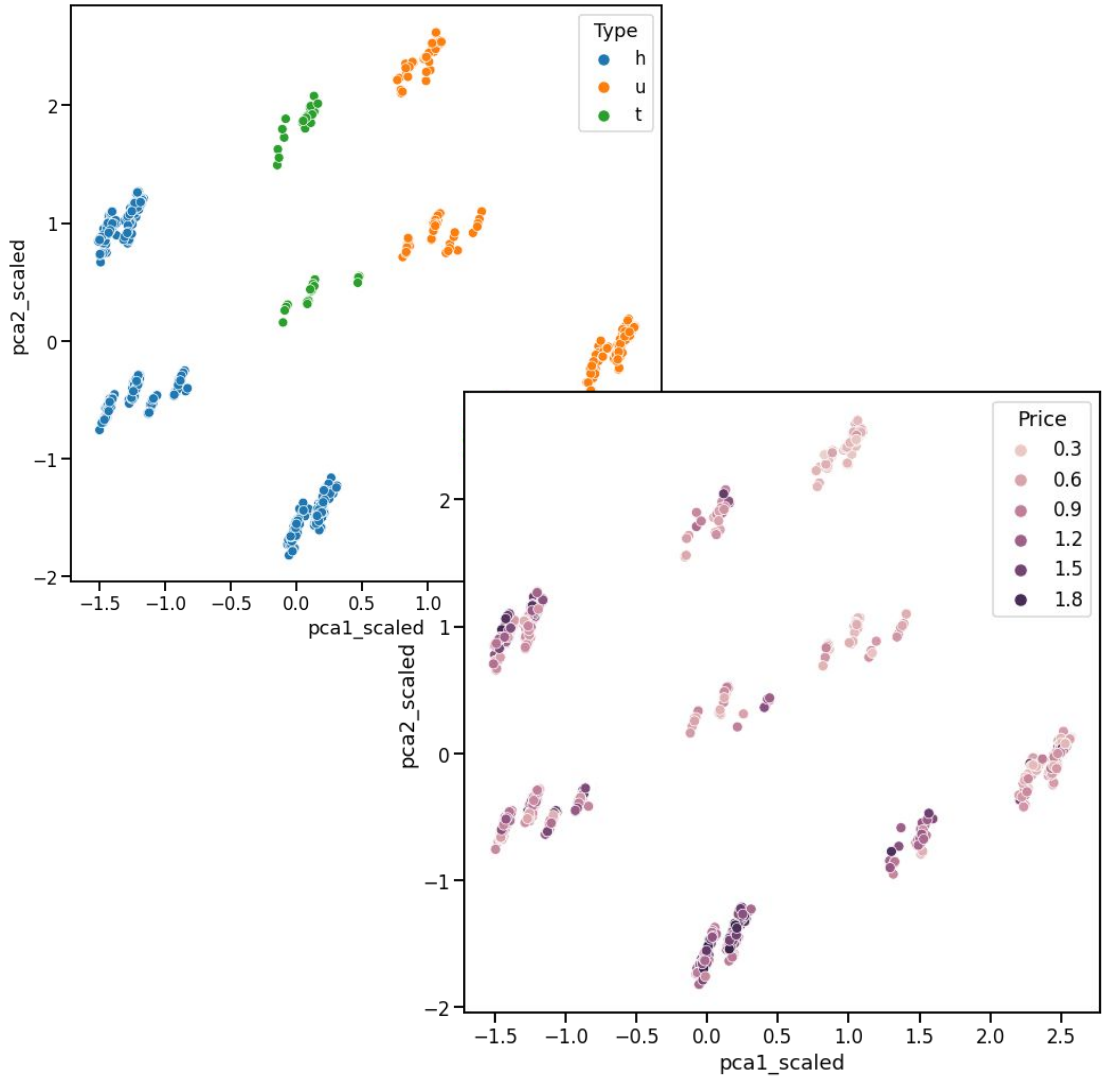
Notebook

04_Encodings_PCA_2024.ipynb

Resultado

En el conjunto de datos de melbourne, las componentes principales separan muy bien los tipos de propiedad, y en menor media el precio

¿Si el tipo está muy relacionado con los componentes del PCA, nos sirve agregar esta nueva información?



Cuando proyectamos cambiamos las propiedades de los datos, queremos proyectar de una forma que ayude a entender/clasificar

Algunos links útiles

- Tutorial de Scikit-learn sobre distintos tipos de descomposiciones
- Video sobre PCA, lamentablemente solo en inglés
- https://setosa.io/ev/principal-component-analysis/