Exploración y Curación de Datos

Diplomatura CDAAyA 2021

¿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
- Reducción de dimensionalidad con PCA
- [Si tenemos tiempo] Reducción de dimensionalidad con LDA

Encodings

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

Es necesario transformar nuestras variables categóricas a algún formato numérico

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				
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				
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.

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.

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

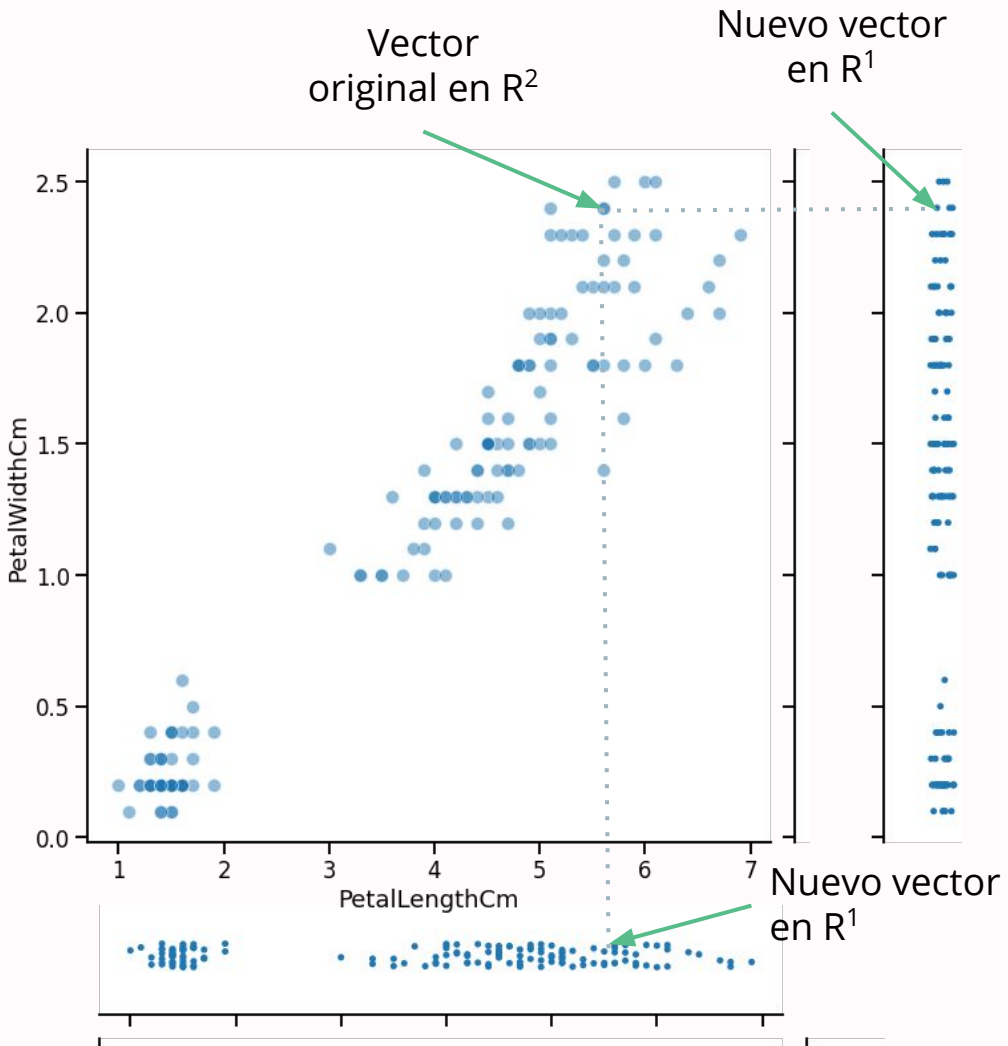
Queremos obtener una nueva matriz Z que tenga la misma cantidad de filas, pero un número de columnas d mucho menor que m.

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

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



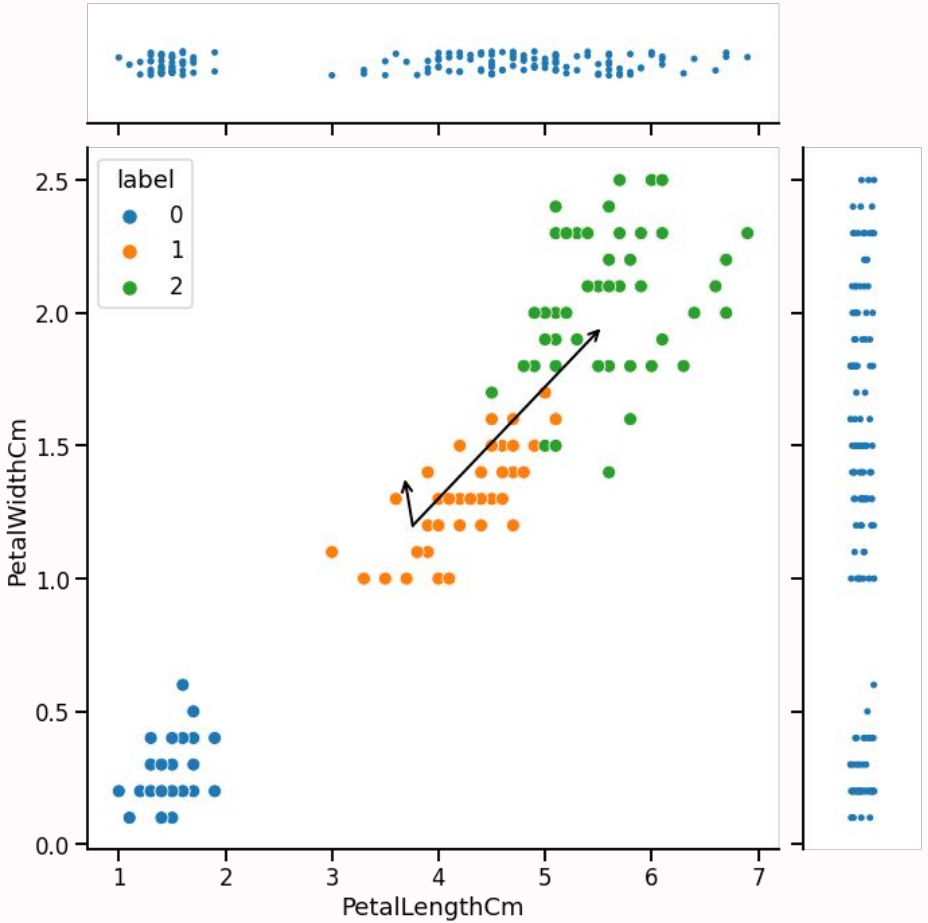
Principal Component Analysis (PCA)

- 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.

Componentes principales

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

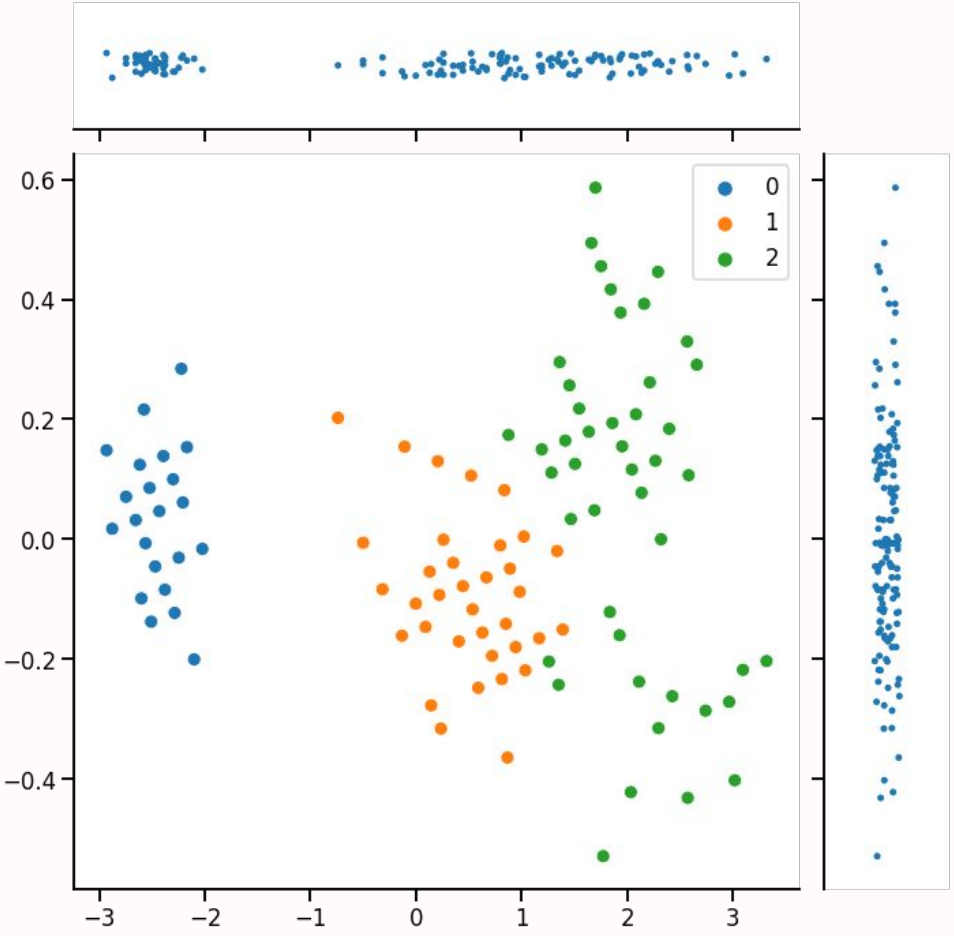
¿Por qué no se “ven” ortogonales?



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**



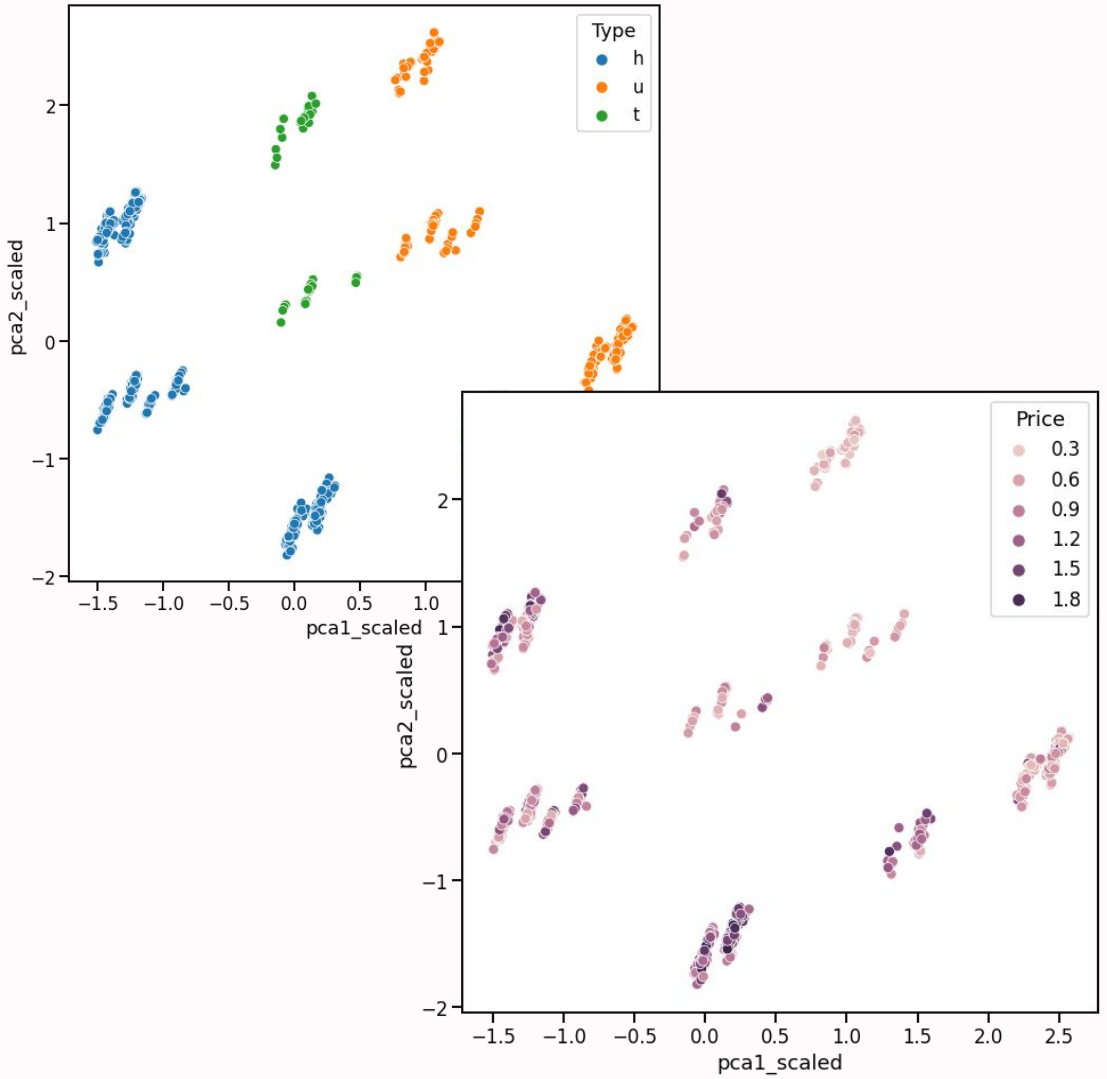
Demo notebook
06 PCA.ipynb

Demo notebook
07 Encodings y PCA en Melbourne.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

Otras proyecciones posibles

Análisis de texto libre

Suburb	closest_airbnb_neighborhood_overview
Melton South	Close to the CBD, 30-60 minutes from top Victorian beaches and suitable for day trips out to the beautiful Victoria countryside...
Oakleigh	Close to Chadstone Shopping centre, Oakleigh Centro, Walking distance approx 500m to Oakleigh and Huntingdale train station .Bus stops are easily available a couple of streets away...
Balwyn	Filled with gorgeous parks, award winning restaurants and shops and leading Deli's across Melbourne. It's close to the city- 15 minute tram ride into the city or 12 minutes into Richmond...

Codificación de texto en bolsas de palabras

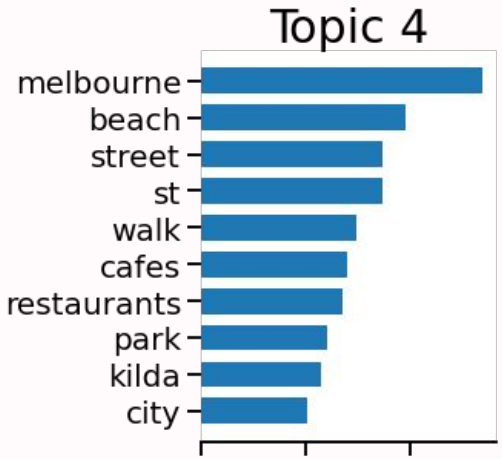
Id	Comentario
1	Nada de tráfico
2	Cerca del aeropuerto
3	Tráfico del aeropuerto
4	Cerca de la playa

Id	aeropuerto	cerca	de	del	la	nada	playa	tráfico
1	0	0	1	0	0	1	0	1
2	1	1	0	1	0	0	0	0
3	1	0	0	1	0	0	0	1
4	0	1	1	0	1	0	1	0

Topic modeling con LDA

LDA o Latent Dirichlet Allocation es un modelo que asume que cada texto habla de un tema o topic desconocido.

Encuentra los vectores que corresponden a los topics que mejor explicarían los datos



Proyección con LDA

Luego, LDA se usa para estimar la probabilidad condicional de que un texto esté hablando de cada uno de los topics.

Podemos representar ahora cada texto con una combinación de distintos temas

closest_airbnb_neighborhood_overview	topic0	topic1	topic2	topic3
Our house is located in a very small, quiet and safe court in the bayside suburb of Moorabbin, with no through traffic, so you are undisturbed by traffic noise. The local shopping centre and cafes is 10 minute's walk from the house The large Southland (Westfield) Shopping Centre is 2.6Km away and easily accessible by a bus which is a few minutes walk from our home. Chadstone is a bus ride away. Brighton Beach is 6Km from the house and easily accessed by public transport, where you can enjoy a walk or swim, or a meal of fish and chips on the foreshore.	0,001	0,001	0,934	0,062

Demo notebook
08 Encodings para texto y LDA.ipynb

Algunos links útiles

- Tutorial de Scikit-learn sobre distintos tipos de descomposiciones
- Video sobre PCA, lamentablemente solo en inglés
-