Análisis de Datos Masivos para el Negocio

Tema 7. Técnicas de aprendizaje no supervisado

Índice

Esquema

TÉCNICAS DE APRENDIZAJE NO SUPERVISADO

Aprendizaje no supervisado

- Técnicas de aprendizaj eautomático para crear modelos sin ejemplos previos. Sólo se dispone de observaciones de las variables independientes.
- El ANS se utiliza principalmente para crear modelos de clusterización o agrupamiento y de reducción de la dimensión.

TÉCNICAS DE CLUSTERIZACIÓN

Un modelo de clusterización toma un conjunto de observaciones de **variables independientes** y las divide en subconjuntos de forma que:

- las observaciones de un subconjunto estén muy relacionadas entre sí.
- los subconjuntos tengan poca relación entre ellos.

Técnicas basadas en partición:

- Utiliza centroides para representar cada agrupación.
- El número de centroides es información proporcionada por el usuario
- Cada observación se asigna a la agrupación con el centroide más cercano

Técnicas basadas en densidad:

- Las agrupaciones se forman por crecimiento o unión de nuevas observaciones.
- Genera de forma automática el número de agrupaciones.

Técnicas basadas en jerarquía:

- Utilizan el denominado dendograma

TÉCNICAS DE REDUCCIÓN DE LA DIMENSIÓN

Un modelo reductor de la dimensión toma un conjunto de observaciones de **variables independientes** y proporciona una estimación de las **variables latentes** o subyacente de forma que:

- se intenta minimizar el número de variables latentes.
- se intenta minimizar la cantidad de información perdida.

Análisis factorial:

- Agrupa las variables que están muy relacionas entre sí, altamente correladas. Ese grupo se sustituye por una única variable latente.
- Mantiene desagrupadas las variables no relacionadas.

Análisis de componentes principales (PCA)

- Define nuevas variables en función de la variabilidad de las observaciones.
- Cada nuevo componente principal (variable latente) es la combinación de las variables originales que explica mayor cantidad de variación en las observaciones.
- Los últimos componentes principales se pueden eliminar pues explican poca variabilidad de las observaciones.

7.1. Introducción y objetivos

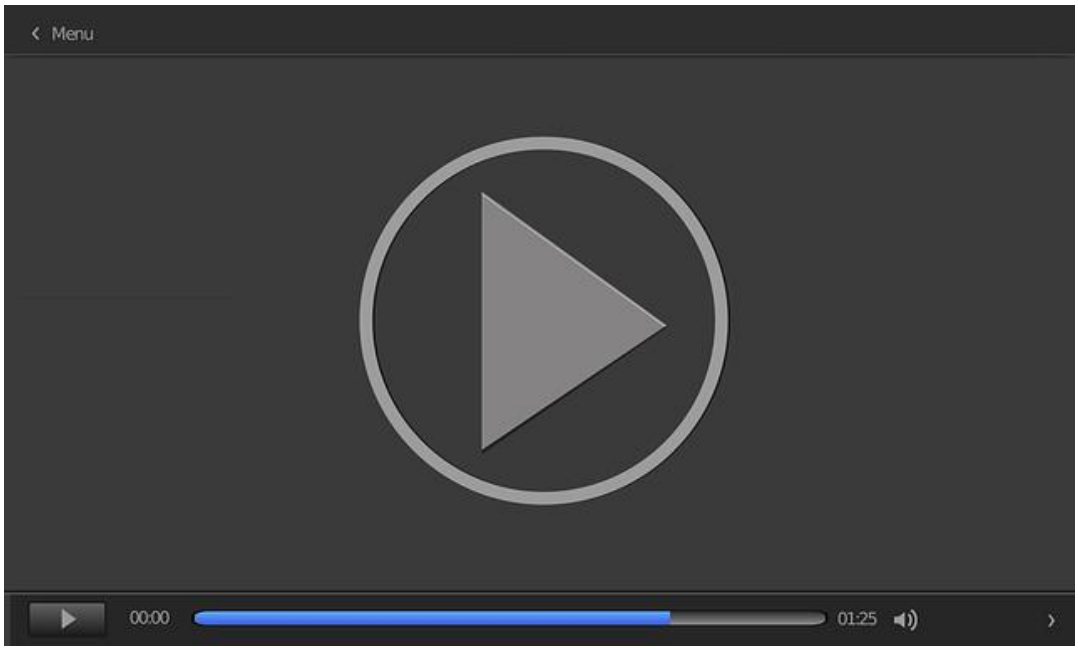
El tema «Modelo de un proceso orientado a datos» de esta asignatura estaba dedicado a presentar las distintas fases de los proyectos que tratan de extraer conocimiento de los datos disponibles. El tema «Proceso del modelado de datos» de la asignatura está dedicado a la fase de modelado de los datos. Cuando los datos están adecuadamente etiquetados, es conocido el resultado que se quiere obtener y en el modelado se pueden utilizar técnicas de aprendizaje supervisado. Por otro lado, cuando los datos no tienen etiquetas previas, pero es necesario averiguar las relaciones latentes o internas que contienen, se pueden obtener modelos basados en técnicas de aprendizaje automático no supervisado. El apelativo de no supervisado se refiere a que no se dispone de ejemplos previos en base a los que entrenar el modelo. Estas técnicas deben crear un modelo estudiando la propia estructura de los datos disponibles. Desde otro punto de vista, estas técnicas se aplican cuando en los datos disponibles, no existen variables dependientes (etiquetadas) y solo disponemos de variables dependientes. Existen principalmente dos familias de técnicas de aprendizaje no supervisado, las técnicas de agrupamiento o clusterización y las técnicas de simplificación de los datos o reducción de la dimensión.

Objetivos que se pretenden conseguir en este tema:

- Entender qué son las técnicas de clusterización.
- Conocer las principales técnicas de clusterización.
- Entender qué son las técnicas de reducción de la dimensión.
- Conocer las principales técnicas de reducción de la dimensión.

Ideas clave

Vídeo *Técnicas de aprendizaje no supervisado*.



Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=1ca244ab-71f9-4dda-a1bb-b15d00aa2656

7.2. Introducción a las técnicas de clusterización

Una de las principales técnicas de aprendizaje no supervisado es el agrupamiento o clusterización. Dado un conjunto de datos, estas técnicas intentan identificar qué relaciones internas hay entre las distintas observaciones. Por tanto, estas técnicas tienen como entrada un conjunto de observaciones y se obtiene como salida distintos subconjuntos o clústers de observaciones. En principio, las observaciones pertenecientes al mismo clúster están altamente relacionadas, mientras que observaciones pertenecientes a clústers distintos tienen un nivel bajo de relación. Desde un punto de vista práctico, **la clusterización** es una herramienta muy potente para realizar de forma automática segmentación de clientes o empresas. Esto es, crear categorías en los clientes o en las empresas de interés. Dichas categorías estarían basadas en características comunes, compartidas por sus miembros, pero diferenciadas entre distintas categorías. La segmentación es una herramienta potente para técnicas de ventas, marketing, publicidad, etc.

El primer paso que vamos a dar en esta sección es explicar de forma sencilla qué entendemos por clusterización. Para ello vamos a utilizar un pequeño ejemplo.

	Datos
	Variable
Observación 1	6
Observación 2	2
Observación 3	4
Observación 4	3
Observación 5	1010
Observación 6	1000
Observación 7	1015
Observación 8	1002

Tabla 1. Conjunto de datos disponibles.

Qué es la clusterización

Supongamos que disponemos del conjunto de datos incluidos en la tabla 1. La tabla muestra en cada fila las observaciones de las que disponemos. La figura 1 muestra los mismos datos.

Figura 1. Observaciones disponibles.

El objetivo en un problema de clusterización es crear subconjuntos de las observaciones. Como es una técnica de aprendizaje no supervisado, las observaciones no están etiquetadas por lo que hay que descubrir posibles relaciones latentes entre los datos.

Una posible división es la que presenta la figura 2.

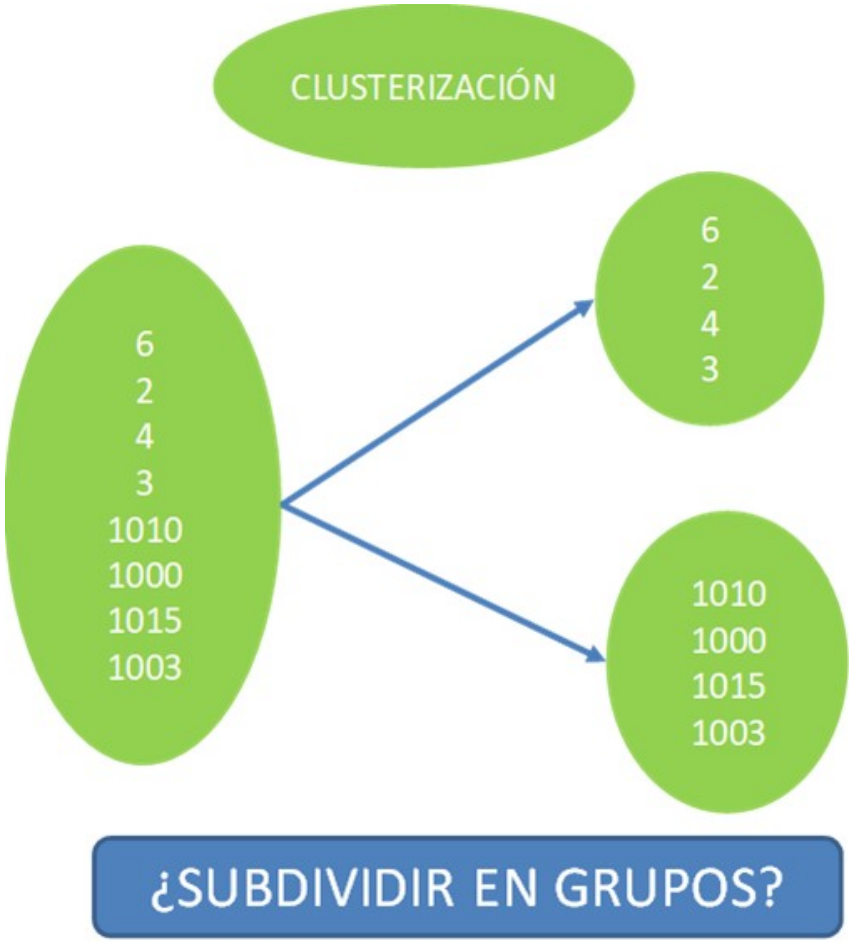


Figura 2. Predecir una nueva observación.

El conjunto de observaciones se ha dividido en dos subconjuntos. Los subconjuntos cumplen las siguientes propiedades, las observaciones de cada subconjunto se parecen mucho entre ellas, están muy relacionadas. Y además las observaciones de conjuntos diferentes están muy poco relacionadas. Por tanto, se ha intentado que las relaciones internas del conjunto sean grandes, pero que las relaciones entre conjuntos sean pequeñas.

La clusterización se puede usar como un paso previo a la tarea de

clasificación. Una vez que hemos dividido las observaciones, estas se pueden etiquetar y una vez etiquetadas se puede crear un clasificador. El clasificador permitirá decidir a qué grupo pertenece una nueva observación.

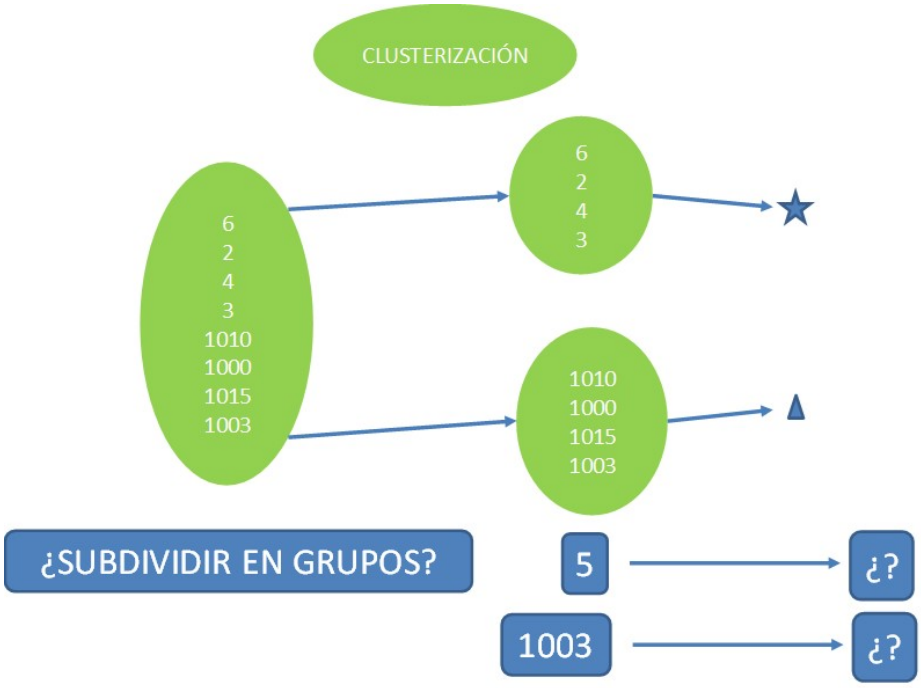


Figura 3. Predecir una nueva observación.

La figura 3 muestra ese proceso. El conjunto original de datos se ha subdividido en dos conjuntos de datos. Cada conjunto se ha etiquetado, en este caso, con clase estrella y clase triángulo. Una vez diseñado el clasificador, se le puede asignar una clase a cada una de las nuevas observaciones.

Resumiendo el ejemplo anterior, dado un conjunto de observaciones, el objetivo de las técnicas de clusterización es obtener un conjunto de clústers que describan las relaciones internas de las observaciones.

Medidas de calidad de una clusterización

Uno de los problemas más interesantes en la clusterización de un conjunto de observaciones es intentar medir su calidad. En un intento de objetivar la idoneidad de la división en clústers o subconjuntos se proponen un conjunto de medidas.

Entre los índices que miden dicha calidad se tienen:

- Índice de Davis-Boulding.
- Índice de Dunn.
- Coeficiente de Silhouette.

Los coeficientes anteriores se pueden usar para seleccionar la mejor división en clústers de los datos entre varias opciones candidatas. Es importante tener en cuenta que a la hora de utilizar los índices de calidad se tienen que usar técnicas de normalización u homogeneización para que ninguna variable concreta domine el resto. Además, hay que definir el concepto de *similitud entre observaciones* para que semánticamente tenga sentido en cada proyecto concreto.

7.3. Técnicas de clusterización

Una vez entendido cuál es el objetivo de las técnicas de clusterización, esta sección presenta las principales técnicas de clusterización y explica cómo funcionan.

Métodos basados en partición

Dado un conjunto de observaciones y un número *k* de clústers, un método basado en particiones **divide el número total de observaciones en k particiones o subconjuntos.** La idea básica es que un clúster de puntos debe estar compuesto por un conjunto de puntos muy parecidos al centro del clúster. Por tanto, cada conjunto está representado por su centro o centroide. Cada una de las observaciones se asigna al conjunto o clúster que tenga el centroide más cercano.

El método basado en particiones más conocido es el denominado *k-means*. En el método *k-means,* la *k* denota el número de subconjuntos en los que deseamos dividir el conjunto inicial de observaciones. El número *k* es una información a priori proporcionada por el usuario. La calidad de los subconjuntos obtenidos es muy dependiente del valor de *k* utilizado. En observaciones de dimensión 2 o 3, que se pueden representar en una gráfica, una inspección visual nos puede dar una estimación de ese valor *k*. En observaciones de dimensión mayor, se suelen utilizar varios valores de k y seleccionar el valor que mejor resultado proporcione mediante una de las métricas presentadas en la sección anterior.

La entrada inicial del algoritmo *k-means* es el conjunto completo de observaciones y el número *k* de *clústers* deseados. El algoritmo prueba inicialmente con *k* ejemplos de centros de clúster. Estos centros iniciales los puede proporcionar el usuario o bien ser generados aleatoriamente. El algoritmo, iterativamente, mueve los centros de clúster en las direcciones cercanas donde aumenta la cantidad de puntos vecinos cercanos. El resultado final es que cada centro de clúster se coloca en la zona de

máxima vecindad local de puntos.

Es importante indicar que la forma de los clústeres obtenidos mediante *k-means* está muy influenciada por la utilización de un centro y el concepto de distancia usado. A un nivel genérico, se puede decir de *k-means* obtiene clústeres con forma esférica o elipsoidal.

Métodos basados en densidad

A diferencia de los métodos basados en partición que modelan cada clústers en función de la distancia de cada componente al centro del clúster, los métodos basados en densidad **modelan los clústers en función de la densidad local de puntos.** La idea básica es que un clúster es una aglomeración densa de puntos donde cada observación debe tener muchos vecinos cercanos. Así mismo los clústeres deben estar separados por regiones con poca densidad de observaciones. Qué cantidad de observaciones definen un clúster y qué significa que una región es densa en observaciones son parámetros que debe definir el usuario para cada proyecto concreto.

Una de las técnicas basadas en densidad es la denominada Density-Based Spatial Clústering of Applications with Noise (DBSCAN). Este algoritmo tiene tres entradas: las observaciones a clústerizar, un parámetro positivo de distancia denominado épsilon y un número mínimo de vecinos denominado minPoints. La técnica devuelve el conjunto de observaciones clusterizado, esto es dividido en clústeres. Es importante indicar que en este caso el número de clústers no está predefinido, es función del proceso de creación.

Si una observación tiene como mínimo minPoints vecinos con una cercanía menor al parámetro épsilon definido entonces se considera que todas esas observaciones son parte de un mismo clúster. El clúster se expande incorporando vecinos que cumplan la condición antes expuesta. Es importante entender que la expansión del clúster se realiza punto a punto sin considerar ningún centroide, lo que permite crear clústeres que no tienen un centro interno y por tanto pueden adquirir formas conexas pero complejas.

Cuando un clúster ha terminado de expandirse, porque no hay observaciones en la vecindad, se considera la creación de otro clúster con el resto de observaciones.

Es posible que algunas observaciones no se incluyan en ningún clúster. En ese caso se consideran observaciones que son ruido o sin información relevante.

Método jerárquico

Los métodos de clusterización jerárquica están **basados en los denominados dendogramas.** Dado un conjunto de observaciones, un **dendograma** es una representación gráfica en forma de árbol que muestra las relaciones entre las observaciones. Cada una de las observaciones se representan en las hojas del árbol. La interrelación entre observaciones se realiza mediante ramas y nodos. Una rama une dos nodos o bien una hoja con un nodo. Un nodo conecta dos ramas inferiores con una superior. El nivel de relación entre dos observaciones viene dado por el nivel del nodo que las conecta dentro del dendograma.

La figura 4 muestra un conjunto de observaciones, las cuales están numeradas. Se pueden ver 10 observaciones diferentes. El objetivo es crear un dendograma que proporcione información de las relaciones entre observaciones. El correspondiente dendograma se puede observar en la figura 5. Las observaciones se pueden ver en el eje horizontal de la gráfica, formando las hojas del dendograma. Claramente las observaciones 2 y 3 son las que mayor cercanía geográfica tienen y por tanto están unidas por el nodo de menor nivel. La altura del nodo, altura en el eje vertical es la

distancia medida entre dichas observaciones. El siguiente nodo en nivel es el que une la observación 4 con el nodo que unía las observaciones 2 y 3. El tercer nodo en jerarquía es el que une las observaciones 7 y 8. El resto del dendograma presenta todos los nodos y ramas de unión. En este ejemplo sencillo es fácil observar gráficamente que el dendograma tiene dos grandes ramas de observaciones. Cada rama está relacionada con nodo de altura pequeña, lo que significa que tienen observaciones muy cercanas. Sin embargo, el nodo principal que une esas dos grandes ramas tiene una gran altura, lo que representa mucha distancia entre ellas. Por tanto el dendograma nos muestra la presencia en las observaciones de dos grandes clústeres.

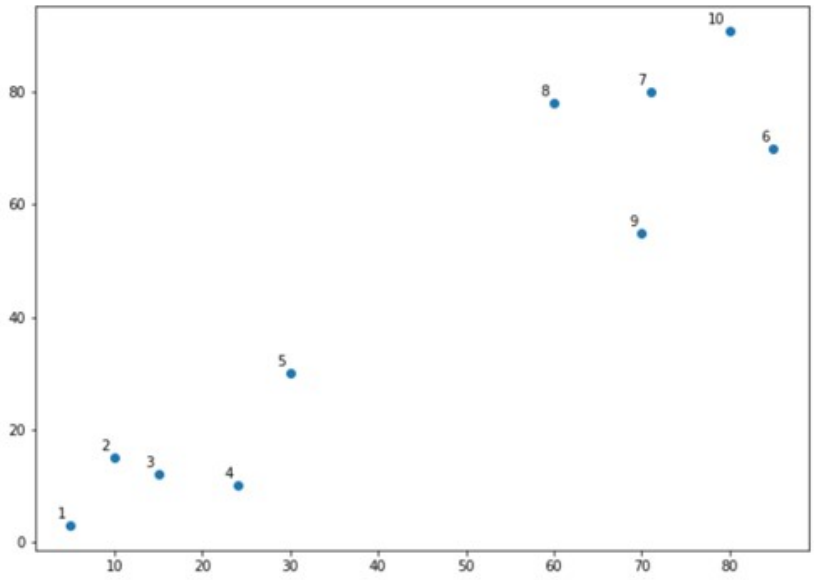


Figura 4. Conjunto de observaciones. Fuente: https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/

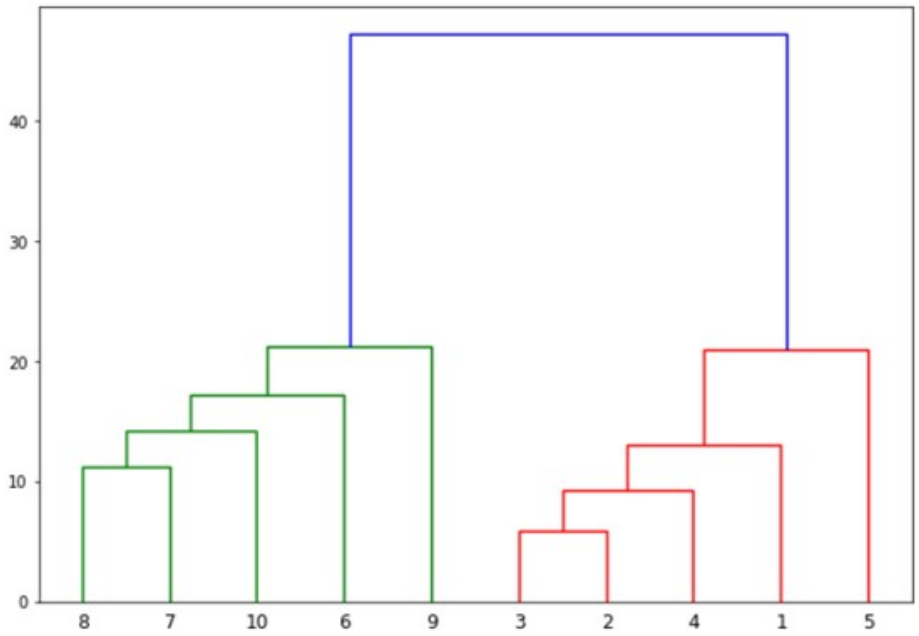


Figura 5. Dendograma. Fuente: https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/

7.4. Introducción a las técnicas de reducción de la dimensión

Las técnicas de reducción de la dimensión se utilizan principalmente para simplificar la cantidad de datos disponibles, pero sin disminuir la información que contienen. Disminuir la dimensión no significa disminuir el número de observaciones disponibles. El objetivo es disminuir el número de variables que componen una observación. Por ejemplo, supongamos que disponemos de un conjunto de observaciones mensuales. Las variables observadas son ingresos, gastos y cantidad ahorrada de una persona. En este caso estamos en dimensión 3 pues son tres las variables observadas. Sin embargo, una de las variables es claramente redundante, pues la cantidad ahorrada es la resta entre los ingresos y los gastos. Por tanto, con observaciones de dimensión 2 (ingresos y gastos) tenemos la misma información, pero con menos dimensiones y por tanto menos datos. Las técnicas de reducción de la información basadas en aprendizaje automático buscan reducir la dimensión de los datos disponibles buscando relaciones latentes, ocultas y posiblemente complicadas de descubrir de forma directa. El conjunto de variables iniciales se reduce a un conjunto de variables latentes que son las que contienen la información relevante de los datos. Las variables latentes están ordenadas de mayor a menor importancia, indicando la cantidad de información que contienen. Por otro lado, hay que indicar que una variable latente puede ser una combinación de varias de las variables originales, indicando que esas variables originales contienen información muy similar que se puede resumir en la latente.

Desde el punto de vista práctico, las técnicas para reducir la dimensión tienen dos funciones bien definidas.

- Por un lado, simplifican el conjunto de datos, pero manteniendo la cantidad de información.

- La otra utilidad es identificar qué variables latentes son las importantes.

Para ilustrar la última utilidad veamos un ejemplo práctico. Supongamos que deseamos averiguar cuáles son las razones más importantes que llevan a un cliente a comprar un producto. Para ello hemos de identificar las variables originales que pensamos pueden influir en la decisión del comprador a la hora de comprar ese producto (por ejemplo, precio, tamaño, estética, etc.). Una vez tenemos esas variables tendremos que obtener observaciones, que pueden ser, por ejemplo, encuestas donde el posible comprador puntúe de 0 a 10 la importancia que le da a cada una de esas variables a la hora de comprar el producto. Una vez recopiladas todas las observaciones se puede aplicar un método de reducción de la dimensión para descubrir qué variables latentes, no observables directamente, tienen esas observaciones. Las variables latentes más importantes son las que contienen la información relevante de las preferencias de los posibles compradores. Dos son las informaciones importantes que podemos usar. La primera es que como cada variable latente estará relacionada principalmente con un subconjunto las variables originales, ese nos indica que ese subconjunto de variables contiene variables muy relacionadas entre sí, o sea, que expresan más o menos el mismo concepto. Por otro lado, como las variables latentes están ordenadas por importancia, las variables originales relacionadas con las variables latentes importantes son las principales razones valoradas en general por los compradores del producto. Esta información puede ser utilizada en la comercialización del producto.

Esta sección comienzaexplicando mediante un ejemplo y de forma sencilla qué entendemos por reducción de la dimensión.

		Datos	
	Variable 1	Variable 1	Variable 3
Observación 1	1	2	3
Observación 2	2	1	3
Observación 3	1	3	4
Observación 4	2	1	3
Observación 5	3	3	6
Observación 6	4	1	5

Tabla 2. Conjunto de datos disponibles.

Qué es la reducción de la dimensión

Supongamos que disponemos del conjunto de datos incluidos en la tabla 2. La tabla muestra en cada fila las observaciones de las que disponemos. Las columnas representan las variables que componen cada observación. En este caso tenemos datos de dimensión 3. Si realizamos una observación atenta de los datos, podemos constatar que los datos de la variable 3 son la suma de los datos incluidos en la variable uno y dos. Por tanto, la variable 3 contiene información redundante pues es deducible de las otras dos variables. Una forma de reducir la dimensión de los datos es eliminar en todas las observaciones una de las tres variables, pues esa información está contenida en las otras dos variables

7.5. Técnicas de reducción de la dimensión

Análisis factorial

El objetivo del análisis factorial es **encontrar las variables latentes o factores comunes que hay en el conjunto de observaciones disponibles.** En la figura 6 se muestra un ejemplo gráfico para explicar cómo trabaja el análisis factorial. Se parte de un conjunto de observaciones de nueve variables. Cada variable se ha representado con un color. El algoritmo del análisis factorial agrupa aquellas variables que están muy relacionadas entre sí (por ejemplo, que tienen alta correlación). En este caso la relación entre variables se ha representado mediante colores de la misma familia. El análisis factorial asume que aquellas variables que están altamente relacionadas entre sí, pero, pobremente con el resto representan a una variable latente u oculta. En este caso, el análisis detecta tres variables latentes representadas con círculos que engloban a las variables originales que representan. La utilización de estas tres variables latentes disminuye la dimensión de las observaciones desde 9 a 3.

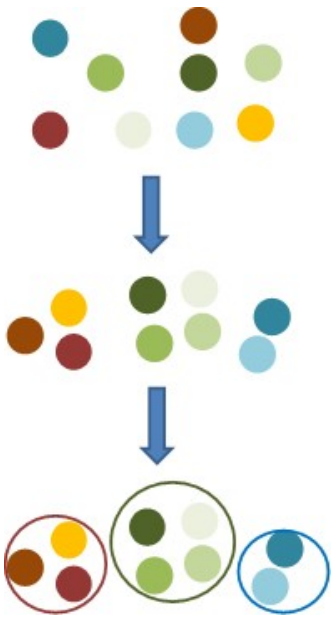


Figura 6. Variables latentes.

La figura 7 representa cómo el conjunto de las tres variables latentes genera la totalidad de las variables originales. Es importante destacar que aquellas variables que están muy relacionadas tienen como origen la misma variable latente o factor común.

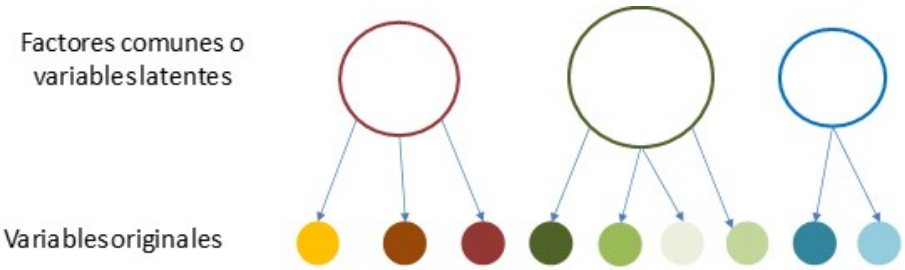


Figura 7. Variables latentes.

En datos reales, la relación entre variables latentes y originales no es tan directa. Es posible que una variable original sea generada a partir de varias variables latentes complicando la interpretación de los datos originales. La figura 8 representa gráficamente un ejemplo de esas relaciones. La gráfica muestra la relación entre un conjunto de 11 variables originales (representadas mediante cuadrados) y las 2 variables latentes o factores comunes más importantes (representadas mediante círculos). El grado de influencia de cada variable latente en las variables originales se representa mediante el grosor de las flechas que las unen. Por ejemplo, es claro que el factor común 1 influye mucho en los valores obtenidos por las variables P.198, P.199 y P.200. Otra interpretación es que esas tres variables se pueden resumir en la variable latente 1. Por otro lado, el factor común 2 influye principalmente en las variables Grw, Fod y Prp, aunque con menos intensidad que en el caso del factor 1. Además, hay que indicar que ambos factores influyen en mayor o menor medida en todas las variables originales, sinembargo, lo importante es la intensidad de esa influencia.

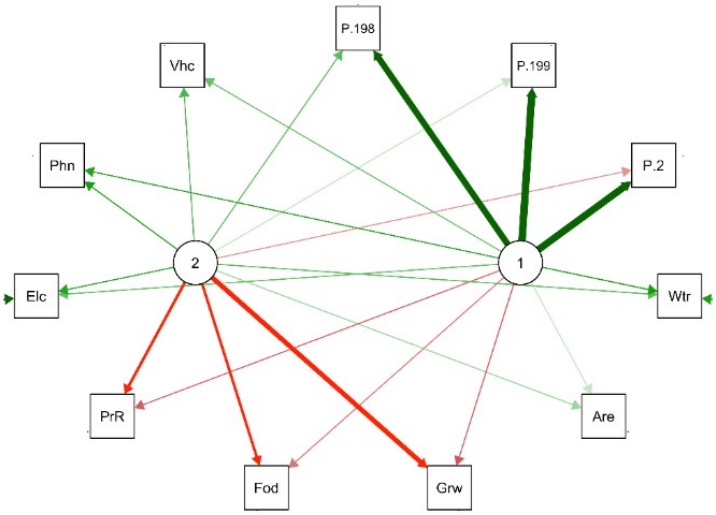


Figura 8. Variables latentes.

Finalmente hay que indicar que, si las variables originales no presentan correlaciones o bien todas están altamente correladas, entonces no es sencillo encontrar variables latentes que las agrupen en subgrupos.

Análisis de componentes principales (PCA)

El análisis de componentes principales es una técnica que permite descubrir las variables latentes que hay en un conjunto de datos. Es una técnica muy relacionada con el análisis factorial, aunque las matemáticas implicadas son diferentes, por lo que los resultados pueden diferir.

En PCA, a cada una de esas variables latentes se le denomina **componente principal.** Los componentes principales están ordenados desde mayor a menor relevancia. El objetivo es concentrar la mayor cantidad de información en los componentes principales más importantes, los primeros, mientras que los últimos apenas tienen información relevante de los datos. El número máximo de componentes principales o variables latentes es igual al número de variables que

componen las observaciones. Si el número de variables latentes o componentes principales importantes es menor al número de variables que componen las observaciones entonces se ha conseguido una reducción de la dimensión de los datos sin perder información.

Veamos visualmente cómo actúa el algoritmo PCA para encontrar los componentes principales de un conjunto de observaciones.

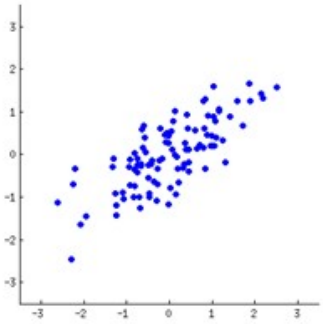


Figura 9. Conjunto de observaciones. Fuente: https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

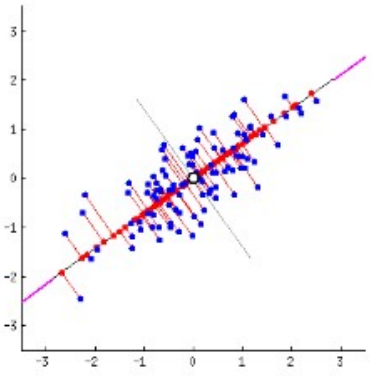


Figura 10. Conjunto de observaciones. Fuente: https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

En la figura 9 se muestran un conjunto de observaciones. Las observaciones provienen de 2 variables y por tanto están en dimensión 2. Se puede observar cierta estructura en los datos. La figura 10 concreta esa estructura. Existe una recta sobre la que se pueden proyectar las observaciones de forma que las observaciones proyectadas mantienen mucha de la información presente en las observaciones originales. Estas observaciones proyectadas en la recta serían la variable latente o primer componente principal. Nótese que las observaciones proyectadas sobre la recta son de dimensión 1 por lo que se ha logrado disminuir la complejidad de los datos sin perder demasiada información. Por supuesto, la información perdida es la segunda componente principal, que es la información mostrada como segmentos rojos en la figura 10. Sin embargo, esa información perdida es la mínima posible desde el punto de vista del algoritmo PCA.

En datos de mayores dimensiones hay qué decidir qué número de variables latentes o componentes principales utilizamos para representar los datos originales.

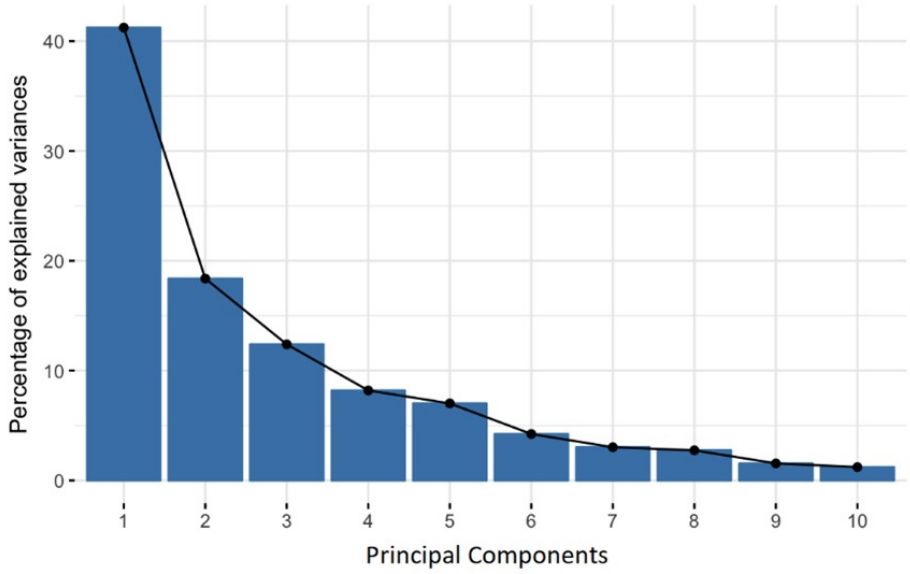


Figura 11. Componentes principales. Fuente: https://towardsdatascience.com/a-step-by-step-explanation-of-principal-component-analysis-b836fb9c97e2

En la figura 11 se representa un ejemplo de gráfico obtenido al aplicar el algoritmo PCA a un conjunto de observaciones de dimensión 10. El eje horizontal representa el número de la variable latente o componente principal. El eje vertical indica el porcentaje de información o varianza que contiene cada variable latente encontrada. Si se quiere retener el 100 % de la información, hay que tomar todas las variables latentes y por tanto no se produce reducción de la dimensión. Por otro lado, si se desea retener hasta un 80 % de la información, con las primeras 5 variables latentes puede ser suficiente, con lo que se habría pasado de dimensión 10 a dimensión 5.

7.6. Referencias bibliográficas

Han, J. y Kamber, M. (2006). *Data mining: concepts and techniques morgan kaufmann.* [Capítulo 7].

Suykens, J. A. K., Gestel, T. V., Brabanter, J. D., Moor, B. D .y Vandewalle, J. (2002). *Least Squares Support Vector Machines.* Singapore: World Scientific.

Tan, P. N., Steinbach, M. y Kumar, V. (2006). *Introduction to data mining Addison-Wesley.* [Capítulos 8 y 9].

Técnicas de clusterización

Naftali Harris (19 enero 2014). Visualizing K-Means Clustering y Naftali Harris (24 enero 2015). Visualizing DBSCAN Clustering.

Las dos páginas webs referenciadas presentan una animación en la que se puede observar cómo funcionan dos de las técnicas más conocidas de clusterización. Son muy didácticas para entender los resultados obtenidos al aplicarlas.

Accede a los artículos a través del aula virtual o desde las siguientes direcciones web: https://www.naftaliharris.com/blog/visualizing-k-means-clustering/ https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Análisis de componentes principales (PCA)

Powell, V. y Lehe, L. (s. f.). Principal component analysis. Setosa y Making sense of principal component analysis, eigenvectors & eigenvalues. Stackexchange.

En las páginas web referenciadas se puede observar un ejemplo interactivo de cómo funciona el algoritmo PCA para encontrar las variables latentes en los datos y una explicación con gráficos dinámicos.

Accede a los documentos a través del aula virtual o desde las siguientes direcciones web: http://setosa.io/ev/principal-component-analysis/ y https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

Análisis factorial y de componentes principales (PCA)

Department of Geography, University of Oregon (s. f). *Factorial and principal components analysis (PCA)*. Geographic Data Analysis.

En la página web referenciada se presenta un ejemplo de estudio de estas dos técnicas. El estudio incluye gráficos basados en grafos que enriquecen la explicación de los resultados obtenidos por los algoritmos.

http://geog.uoregon.edu/bartlein/courses/geog495/lec16.html

Test

1. El dendograma es un gráfico usado en:

A. Métodos basados en partición.

B. Métodos basados en densidad.

C. Métodos jerárquicos.

D. Ninguna de las anteriores es correcta.

2. La clusterización basada en movimientos de centroides y distancias a los centroides corresponde a:

A. Métodos basados en partición.

B. Métodos basados en densidad.

C. Métodos jerárquicos.

D. Ninguna de las anteriores es correcta.

3. En qué métodos de clusterización normalmente hay que indicar a priori el número de clústeres deseado:

A. Métodos jerárquicos.

B. Métodos basados en partición.

C. Métodos basados en densidad.

D. Ninguna de las anteriores es correcta.

4. En un dendograma, dos observaciones están muy relacionadas si:

A. Si una de las dos observaciones es un nodo.

B. Si hay un nodo que las une con nodos mediante otros nodos.

C. Sus nodos padres están al mismo nivel.

D. El nodo que las une está en un nivel bajo o es de poca altura.

5. El análisis factorial se basa:

A. En la búsqueda de variables latentes no observables directamente.

B. En la agrupación de observaciones por densidad.

C. En la creación de un dendograma.

D. En la eliminación aleatoria de las variables originales.

6. ¿Qué ventaja tiene el método DBSCAN frente a k-means?

A. Es más rápido en datasets grandes.

B. Requiere definir el número de clústeres previamente.

C. Detecta clústeres de forma libre y maneja ruido.

D. Solo funciona en datos de dos dimensiones.

7. ¿Qué índice se utiliza comúnmente para evaluar la calidad de una clusterización?

A. Coeficiente de Gini.

B. Índice de Dunn.

C. AUC.

D. Varianza explicada.

8. ¿Cuál de las siguientes es una técnica de reducción de la dimensión?

A. DBSCAN.

B. Análisis factorial.

C. Árbol de decisión.

D. Regresión logística.

9. En el análisis de componentes principales (PCA), el primer componente principal es:

A. El que tiene menor varianza.

B. El que menos contribuye a la reducción de la dimensión.

C. El que más varianza explica de los datos.

D. El que corresponde a la variable original más representativa.

10. ¿Qué técnica usarías para segmentar clientes sin saber previamente sus categorías?

A. Clasificación supervisada.

B. Árboles de decisión.

C. Clusterización no supervisada.

D. Regresión logística.