UNIDAD TEMÁTICA 5: Aprendizaje No Supervisado, Clustering

Trabajo de Aplicación 3 – Análisis de Componentes Principales

EJERCICIO 1 – reducción de dimensionalidad

Comenzaremos con un dataset público sobre información nutricional.

Revisar la información disponible en https://dasl.datadescription.com/datafile/cereals/ y en https://www.kaggle.com/jeandsantos/breakfast-cereals-data-analysis-and-clustering/data y documentar el contexto del problema

El dataset puede descargarse de cualquiera de esos sitios, y también disponible como planilla electrónica en la webasignatura "cereals.xls".

El dataset incluye información sobre los ratings e información nutricional de 77 cereales de desayuno. Hay un total de 16 variables, incluyendo 13 parámetros numéricos.

El objetivo es reducir este conjunto de 13 predictores a una lista mucho menor, utilizando PCA.

Paso 1. Preparación de los datos

1. Crea un nuevo proceso en blanco.
2. Importa el dataset al repositorio de datos de RapidMiner
3. Retira los parámetros no numéricos "Cereal name," "Manufacturer," y "Type (hot or cold)," ya que PCA sólo puede trabajar con atributos numéricos. Son las columnas A, B y C (en RapidMiner se pueden convertir estos atributos en rol ID para usarlos como referencia más tarde).

Paso 2 . Operador PCA

Agrega un operador "PCA" y conéctalo con los datos (ya preparados). Analizar los parámetros que se pueden configurar en este operador.

En principio, para el parámetro "dimensionality reduction" selecciona "keep variance" y para "variance threshold" deja el valor por defecto 0.95. así el operador ha de seleccionar sólo los atributos que expliquen el 95% de la varianza total de los datos.

Conecta las salidas del operador PCA a los puertos de resultados.

Paso 3. Ejecución e Interpretación

- Al ejecutar el proceso RapidMiner crea varias pestañas en el panel de resultados. Al seleccionar la pestaña "PCA" veremos tres secciones relacionadas con PCA: "Eigenvalues" (valores propios), "Eigenvectors" (vectores propios) y "Cumulative Variance Plot" (gráfico de varianza acumulada).

- En la sección de "Eigenvalues" podemos obtener información sobre la contribución individual que cada componente principal aporta a la varianza de los datos.

- Si, como hemos configurado, nuestro umbral de varianza es 95%, entonces alcanza con tener en cuenta solamente los tres primeros componentes, ya que explican cerca del 97% de la varianza de los datos. PC1 contribuye mayoritariamente, con aprox. 55%.

- Luego podemos analizar en profundidad cómo cada uno de los componentes principales identificados se relaciona linealmente con los parámetros reales del dataset. Observa (y documenta) la composición de PC1, PC2 y PC3 (en la vista de "Eigenvectors").

- En este momento, deseamos tener en cuenta solamente los atributos reales que tienen peso significativo en la composición de los CPs.

¿Cómo seleccionar estos atributos?

- Podemos ordenar (en la vista de "Eigenvalues") por cada componente principal, y seleccionar entonces los dos o tres atributos reales que más le impactan.
- Para los PC1, PC2 y PC3, podemos seleccionar "calories", "sodium", "potassium", "vitamins", y "rating" para formar el dataset reducido (obtenidos seleccionando los 3 más importantes para cada PC). En este ejemplo entonces tenemos una reducción de 13 a 5 atributos, o sea, más del 50%.
- Considera por un momento lo que esto significa, cuando el dataset es muy grande, en términos del rendimiento computacional de diferentes algoritmos de ML.
- PCA ES UNA HERRAMIENTA MUY EFECTIVA Y AMPLIAMENTE UTILIZADA para reducción de dimensionalidad, especialmente cuando los atributos son numéricos.

Riesgos a considerar cuando se utiliza PCA:

1. Los resultados de PCA deben ser evaluados en el contexto de los datos
   - Si los datos tienen mucho ruido, PCA puede erróneamente seleccionar los atributos más ruidosos como más significativos.
   - Interpreta los resultados anteriores (los PC en función de los atributos reales para cada producto)
2. Agregar datos no correlacionados no siempre ayuda. Tampoco ayuda agregar datos que pueden estar correlacionados pero son irrelevantes
3. **PCA es muy sensible a los efectos de la escala en los datos!**

Observa los datos del ejemplo.

- ¿qué características estadísticas tienen los atributos reales que fueron identificados como más significativos?

Estos factores dominan los resultados de PCA porque contribuyen más a la varianza total de los datos

- ¿Cuál sería el efecto de otro posible atributo, "volumen de ventas", cuyo rango estuviera en los millones (de $ o cajas)? Claramente, enmascararía los efectos de cualquier otro atributo.
- Para minimizar los efectos de la escala, se recomienda NORMALIZAR LOS DATOS (rangos entre 0 y 1).
- Aplica esta normalización, ejecuta el proceso y analiza los nuevos resultados.