# Álgebra Lineal Computacional

Segundo Cuatrimestre 2023

## Trabajo Práctico N° 2

## Sistemas de Recomendación

### Indroducción

Un sistema de recomendación es una herramienta que utiliza datos sobre las preferencias y comportamientos de sus usuarios y usuarias para sugerirles elementos que le puedan interesar. Se aplican comúnmente en una variedad de entornos, como plataformas de compras en línea, servicios de transmisión de video y sitios de redes sociales.
El objetivo es mejorar la experiencia del usuario proporcionando sugerencias personalizadas que sean relevantes y útiles de forma indivicualizada. En muchos casos, los sistemas de recomendación utilizan algoritmos de aprendizaje automático para analizar los datos y hacer predicciones sobre lo que le puede interesar a cada persona. Estos algoritmos pueden tomar en cuenta factores como las compras anteriores, los artículos que le han gustado o en los que ha hecho clic, y los artículos que son populares entre otros usuarios con intereses similares.

### Los Datos

Se cuenta con el conjunto de datos wine.csv, descargado de https://doi.org/10.24432/C5PC7J (Aeberhard, Stefan and Forina, M.. UCI Machine Learning Repository. 1991. CC by 4.0).
Este conjunto de datos se compone de 178 vinos diferentes y, para cada uno se indican 13 características específicas, desde el porcentaje de alcohol hasta el contenido de prolina. Cada característica proporciona información relacionada con un vino determinado. Cada fila corresponde a un vino, y cada vino tiene características diferentes. La última columna, $'Segmento\_Cliente'$, representa diferentes segmentos de vino, que como se puede ver están clasificados en: 1, 2 y 3. Por tanto, son 13 las variables independientes y la clasificación es la variable dependiente.

### Objetivo

Supongamos que este conjunto de datos pertenece a un comercio que tiene diferentes botellas de vino para vender y además cuenta con una gran base de clientes.
El dueño de la tienda de vinos quiere reducir la complejidad del conjunto de datos y desea construir un sistema de recomendación de vinos para sus clientes.
El objetivo de este problema es identificar a qué segmento de clientes pertenece cada vino para poder recomendar el vino al cliente adecuado. En pocas palabras, hay que crear un sistema de recomendación de vinos, con el objetivo de optimizar las ventas y aumentar los beneficios de la tienda de vinos.

**Ejercicio 1.** Seguir los siguientes pasos para implementar un sistema de recomendación de vinos:

(a) Descargar los datos.

(b) Separar los datos en variables dependientes (X) e independiente (Y).

(c) Normalizar y centrar los datos. Como puede verse el conjunto de datos no está normalizado y puede estar desplazado del origen. Se pide construir una función en python que normalice y centre los valores respecto del promedio. Es decir, aplicar a cada variable (dependiente e independiente):

$x_i = (X_i - \overline{X})/s,$

$\overline{X} = (\sum_{i=1}^{N} X_i)/N$

$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2}$

¿Por qué es importante este paso?

(d) Calcular la matriz de covarianza. ¿Qué representa la matriz de covarianza? ¿Qué dimensiones tiene y qué propiedades cumple?

(e) Encontrar el máximo de los autovalores, y su correspondiente autovector, de la matriz de covarianza hallada en el item anterior por el método de la potencia. Implementar una función que dada un matriz encuente el máximo autovalor y su correspondiente autovector.

(f) Modificar la función anterior para encontrar (además del máximo) los siguientes n autovalores sucesivos y sus respectivos autovectores. Además de la matriz, n debe ser parámetro de entrada de la función.

¿Cómo se relaciona la cantidad de autovalores con la cantidad de propiedades de los vinos? ¿Qué representan cada unos de los autovectores correspondientes?

(g) Finalmente, dado un conjunto de características de un vino en particular, obtener su clasificación, es decir, a qué segmento de cliente pertenece. Para ello, se utiliza el método de PCA a partir de los componentes hallados en los items anteriores. La función debe tomar como dato de entrada el o los vinos a clasificar y el/los autovectores principales hallados con anterioridad.

En primera aproximación utilizaremos el conocido algoritmo de k vecinos más cercanos o kNN [1], por su sigla en inglés, para asignar la clasificación. En su versión más simple, este algoritmo considera a cada objeto de la base de entrenamiento como un punto en el espacio euclídeo m-dimensional (que se corresponde con la cantidad de características de cada vino), para el cual se conoce a qué clase corresponde (en nuestro caso, $'Segmento\_Cliente'$) para luego, dado un nuevo objeto, asociarle la clase del o los puntos más cercanos de la base de datos.

**Ejercicio 2.**

Experimentos y Graficación. Para estudiar la potencialidad del método de PCA, debemos poder comparar valores reales con los predichos con el modelo propuesto. Para ello usualmente se dividen los datos con los que se cuenta en dos subconjuntos. El primero que llamamos datos de 'entrenamiento', usados para ajustar nuestro modelo, y el segundo con datos de 'testeo', datos que usaremos para testear. Hay distintos criterios para seleccionar estos conjuntos pero algo fundamental a tener en cuenta es la cantidad de muestras que tendrá cada conjunto a partir del original, una sugerencia es usar 80% para el conjunto de entrenamiento y el 20% restante para testear.

(a) Para evaluar el sistema de recomendación del ejercicio anterior, dividir el conjunto de datos en dos, para entrenamiento y para testeo. Ejecutar el modelo realizado para el conjunto de entrenamiento utilizando de 1 a 4 componentes principales. Es decir que se obtendrán para predecir 4 modelos, cada uno de los cuales corresponderá a tomar de 1 a 4 componentes principales. Construir una tabla donde se muestre el número de componentes principales de cada modelo y los valores de la varianza explicada[2], la varianza explicada relativa y la acumulada de cada componente, es decir:

| Modelo PCA | Componente | Varianza explicada | Porcentaje | Acumulado |
|---|---|---|---|---|
| 1 Componente Principal | 1 | | | |
| 2 Componentes Principales | 1 | | | |
| | 2 | | | |
| 3 Componentes Principales | 1 | | | |
| | 2 | | | |
| | 3 | | | |
| 4 Componentes Principales | 1 | | | |
| | 2 | | | |
| | 3 | | | |
| | 4 | | | |

(b) Realizar graficos que puedan mostrar los resultados obtenidos para los casos de los modelos con 1, 2 y 3 componentes principales.

De acuerdo a los valores encotrados en la tabla del item anterior, cuál de los modelos sugiere utilizar para predecir el segmento al que pertenece un vino. Tener en cuenta que el comercio cuenta con una gran base de datos de clientes y se quiere poder hacer una recomendación rápidamente.

(c) Realizar la matriz de confusión con los explerimentos realizados en el item (a). ¿A qué conclusiones se puede llegar?

---

### Entrega y lineamientos

La entrega se realizará a través del campus virtual de la materia con las siguientes fechas y formato:

- Fecha de entrega: hasta el domingo **16 de noviembre** a las 23:59 hs.
- Fecha de re-entrega: hasta el **7 de diciembre** a las 23:59 hs.
- Formato: Jupyter Notebook (no se acepta entrega de un archivo .py).

Prestar especial atención a las siguientes indicaciones:

- El TP2 se realizará en los mismos grupos que para el TP1. En caso que algún integrante no continúe con la materia deberá avisar a la cátedra.
- Leer el enunciado completo antes de comenzar a generar código y sacarse todas las dudas de cada ítem antes de implementar. Para obtener un código más legible y organizado, pensar de antemano qué funciones deberán implementarse y cuáles podrían reutilizarse.
- El código debe estar correctamente comentado. Cada función definida debe contener un encabezado donde se explique los parámetros que recibe y qué se espera que retorne. Además las secciones de código dentro de la función deben estar debidamente comentados. Los nombre de las variables deben ser explicativos.
- Las conclusiones y razonamientos que respondan los ejercicios, o cualquier experimentación agregada, debe estar debidamente explicada en bloques de texto de las notebooks (markdown cells), separado de los bloques de código. Aprovechen a utilizar código LaTeX si necesitan incluir fórmulas.

**Importante:** Se deberá agregar un bloque de texto final en la notebook entregada. Este bloque de texto se titulará CONCLUSIONES FINALES y allí deberán escribir sus conclusiones generales que englobe a todos los ejercicios. Una entrega sin conclusiones generales se considerará incompleta.

- Gráficos: deben contener título, etiquetas en cada eje y leyendas indicando qué es lo que se muestra.

[1] Richard O Duda, Peter E Hart, and David G Stork. Pattern classification. John Wiley Sons, 2012.

[2] La varianza explicada puede representarse como la división entre cada uno de los autovalores y la suma de los mismos. Digamos que hay N autovectores, entonces la varianza explicada de cada autovector (componente principal) será:
$Varianza\_explicada = \lambda_i/(\lambda_1 + \lambda_2 + \ldots + \lambda_n)$