PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE | la clase ejecutiva

# Instrucciones Miniproyecto 3

## Aprendizaje no supervisado en Python

Dentro del aprendizaje no supervisado, los algoritmos de *clustering* son de especial importancia, principalmente por la gran variedad de aplicaciones que estos ofrecen. Dentro de estas aplicaciones podemos encontrar: la segmentación de clientes, procesamiento de imágenes, detección de objetos anómalos, entre muchas otras (Jain, 2010).

En esta actividad, a través de ejemplos, usted aplicará el algoritmo *k-means* utilizando librerías de Python.

En particular, a través de ejemplos simples,

- Implementar en Python *k-means, Gaussian Mixture Model y cluster jerárquico.*
- Visualizar los resultados de algoritmos de *clustering.*
- Interpretar los resultados obtenidos.

## Trabajo a realizar

### Implementación y visualización de k-means en 2D

- Cargue la base de datos *kmeans1.csv* utilizando pandas.
- Por medio de matplotlib (o seaborn) genere un gráfico de dispersión (scatter plot) de las variables A y B. ¿Cuántos clusters visualiza?
- Aplique el algoritmo k-means disponible en Scikit-Learn y observe los outputs disponiles (centroides, clusters asignados y distancia dentro de las clases).
- Ajuste k-means considerando 1,2,...10 clusters, guarde la distancia intra clases en un diccionario.
- Utilizando matplotlib (o seaborn) genere un gráfico que presente la distancia intra clases. Utilizando el criterio del codo, defina el número de clusters.

## Análisis de clusters

- Cargue base de datos *k-means2.csv* y genere un análisis descriptivo de la base de datos.
- Visualice y preprocese las variables disponibles dentro de ella.
- Aplique el algoritmo k-means a la base de datos. Determine el número de clusters adecuados con el criterio visto en la actividad anterior.
- Interprete los resultados de los centroides.
- Aplique el clustering jerárquico. Visualice el dendograma y determine el número de clusters.
- Estudie el clasificador GMM desde las lecturas obligatorias de este módulo y use su implementación en Scikit-Learn.
- Utilizando visualizaciones y métricas adecuadas, compare los resultados de K-means y GMM. Comente las principales ventajas de GMM respecto de K-means.

## Referencias


- Jain, A.K. (2010). Data clustering: 50 years beyond k-means.Pattern recognition letters, 31(8):651–666.