

Tarea 2. Laboratorio

Subida: 27 de abril de 2023

Entrega: 5 de mayo de 2023

1. Usando K-Means y el datasetblow.csv (Proporcionado junto con la tarea)
 - (a) Para $k = 2, \dots, 16$. Calcula el WCSS para cada k y gráfica los WCSS de cada k en una gráfica de líneas.(Revisar la referencia [3]).
$$WCSS = \sum_{P_i \in Cluster1} distance(P_i, C_1)^2 + \sum_{P_i \in Cluster2} distance(P_i, C_2)^2 + \dots + \sum_{P_i \in Clusterl} distance(P_i, C_l)^2$$
 - (b) Usando la gráfica anterior justifica que valor de k sería "óptimo" según el criterio del "codo" y gráfica su Diagrama de Voronoid Generado por los centroides de dicho k .
 - (c) Gráfica el comportamiento de la tolerancia relativa, i.e., la norma de Frobenius de la diferencia de los centroides como matriz para cada iteración.
2. Con el mismo k implementa K-MEDOIDS. Usa lo para el mismo dataset(datasetblow.csv).
 - (a) Comprueba que los centroides pertenecen al dataset.
 - (b) Genera el Diagrama de Voronoid Generado por los centroides.
3. Usando el archivo MNIST(mnist_train_small.csv)
 - (a) Realiza una reducción dimensional a dos componentes principales. De todo el conjunto de datos.
 - (b) Separen los datos a cuyas etiqueta correspondan el 0 y 1. De tal forma que a cada etiqueta correspondan 50 muestras. Para crear un conjunto de entrenamiento.(100 Muestras)
 - (c) Gráfica la malla de la clasificación y los puntos de los datos completos de ambas etiquetas. Usando K-NN.
 - (d) Usando sklearn.metrics.classification_report() cree un informe que muestre las principales métricas de clasificación(Precision, Recall, etc). Usando como X_test los datos completos del conjunto de datos y como X_train El conjunto de datos con las 100 muestras.
 - (e) Discutan en equipo y den una conclusión de lo que se observa en los datos.¿Qué puedes decir acerca de las muestras si se usaran 10 muestras por clase? ¿Son muchas o pocas?

Referencias

- [1] Kevin A. Beyer et al. *When Is "Nearest Neighbor" Meaningful?* Springer Science+Business Media, ene. de 1999, págs. 217-235. DOI: [10.1007/3-540-49257-7_{_}15](https://doi.org/10.1007/3-540-49257-7_{_}15).
- [2] Mark De Berg et al. *Computational Geometry: Algorithms and Applications*. Ene. de 1997.
- [3] Saptarsi Goswami. «Clustering using k-Means with implementation - Towards Data Science». En: (dic. de 2021). URL: <https://towardsdatascience.com/clustering-using-k-means-with-implementation-40988620a973>.
- [4] Hae-Sim Park y Chi-Hyuck Jun. «A simple and fast algorithm for K-medoids clustering». En: *Expert Systems With Applications* 36.2 (mar. de 2009), págs. 3336-3341. DOI: [10.1016/j.eswa.2008.01.039](https://doi.org/10.1016/j.eswa.2008.01.039).