Tarea 5. Aprendizaje no supervisado

Universidad Autonoma de Nuevo León
Beas Ham, Nelson Alfonso

11 de junio de 2024

1. Introducción

Se pretende analizar datos de tumores en diferentes partes del cuerpo de un grupo de pacientes, para los cuales se pretende crear una clasificación en dos grupos no supervisada, los datos se conforman por registros unicamente númericos para distintos atributos relevantes, como lo es el radio del rumor, un indice de textura, un perimetro, grado de compactación, entre otros.

2. Descripción de los datos

Los datos con los que se trabajaran constan de 9 dimensiones principales y 22 sub dimensiones, por lo que él análisis se centrará principalmente en las siguientes 9 características, una de las transformaciones más relevantes que se le realizaran a los datos es que los valores categóricos binarios de la columna çal"serán cambiados de (M, B) a (1, 0), todos los datos son idealmente números reales.

- cal
- radius_mean
- texture_mean
- perimeter_mean
- area_mean
- smoothness_mean
- compactness_mean
- concavity_mean
- concave points_mean

2.1. Origen de los datos

Los datos fueron obtenidos de un registro público de conjuntos de datos, los cuales provienen de estudios de múltiples hospitales, a continuación se encuentra el enlace al repositorio.

https://www.kaggle.com/datasets/erdemtaha/cancer-data/data

2.2. Estadística descriptiva básica

A continuación podemos observar las características estadísticas principales de nuestro conjunto de datos, el cual está formado por 569 registros, es importante tomar en cuenta que los conjuntos no están perfectamente balanceados.

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean
count	569	569	569	569	569	569	569	569	569
mean	0.372583	14.1273	19.2896	91.969	654.889	0.0963603	0.104341	0.0887993	0.0489191
std	0.483918	3.52405	4.30104	24.299	351.914	0.0140641	0.0528128	0.0797198	0.0388028
min	0	6.981	9.71	43.79	143.5	0.05263	0.01938	0	0
25%	0	11.7	16.17	75.17	420.3	0.08637	0.06492	0.02956	0.02031
50%	0	13.37	18.84	86.24	551.1	0.09587	0.09263	0.06154	0.0335
75%	1	15.78	21.8	104.1	782.7	0.1053	0.1304	0.1307	0.074
max	1	28.11	39.28	188.5	2501	0.1634	0.3454	0.4268	0.2012

Figura 1: Resumen estadístico de los datos

2.3. Preprocesamiento

La principal transformación que se aplicará a los datos será modificar el dominio de la variable "diagnosis"de la siguiente manera: M se tomará como 1 y B como 0.

Ademas, se recortará del conjunto original las sub Categorias, por lo que nos quedaremos con las 9 listadas en el apartado 2.

3. Antecedentes

1. Década de 1980 y 1990: - Durante este período, se comenzaron a explorar las primeras aplicaciones de la inteligencia artificial en el campo del cáncer. - Se utilizaron técnicas de redes neuronales artificiales y sistemas expertos para el diagnóstico y pronóstico del cáncer.

2. Década de 2000: - Se vio un aumento significativo en la aplicación de algoritmos de aprendizaje automático, como máquinas de vectores de soporte (SVM), redes neuronales convolucionales (CNN) y árboles de decisión, para tareas de diagnóstico y detección temprana del cáncer. - Se desarrollaron sistemas de apoyo a la toma de decisiones basados en inteligencia artificial para ayudar a los médicos en la interpretación de imágenes médicas, como mamografías y resonancias magnéticas, para la detección de cáncer de mama, cáncer de próstata, etc. - Se comenzaron a utilizar algoritmos de minería de datos para analizar

grandes conjuntos de datos clínicos y genómicos en busca de patrones y biomarcadores que puedan predecir la predisposición al cáncer, la progresión de la enfermedad y la respuesta al tratamiento.

3. Década de 2010: - La integración de la inteligencia artificial y el análisis de big data condujo a avances significativos en la medicina de precisión y la oncología personalizada. - Se desarrollaron algoritmos de aprendizaje profundo (deep learning) para la interpretación automatizada de imágenes médicas, como tomografías computarizadas (TC), resonancias magnéticas (RM) y biopsias digitales, mejorando la precisión en el diagnóstico y la detección de cáncer. - Se utilizaron técnicas de aprendizaje automático para identificar biomarcadores genómicos y moleculares que pueden ayudar en la selección de tratamientos específicos y predecir la respuesta del paciente a la terapia.

4. Década de 2020 en adelante: - Se espera que la inteligencia artificial continúe desempeñando un papel fundamental en el avance de la investigación y el tratamiento del cáncer. - Se están explorando nuevas aplicaciones de la inteligencia artificial, como la medicina de sistemas, que integra datos clínicos, genómicos, proteómicos y de imágenes para comprender mejor la biología del cáncer y desarrollar terapias más efectivas y personalizadas. - Se espera que la inteligencia artificial siga evolucionando para mejorar la precisión en el diagnóstico temprano, la predicción del riesgo de recurrencia, la identificación de nuevas dianas terapéuticas y la optimización de los regímenes de tratamiento del cáncer.

4. Metodología

Para este análisis, el atributo "Diagnosis"no le será proporcionado al algoritmo, haciendo uso de la librería pandas y sklearn del lenguaje de programación python a continuación describiremos las bases del algoritmo candidato para la clasificación.

5. Descripción del algoritmo K-Means

El algoritmo K-Means es un método de agrupamiento (clustering) que divide un conjunto de datos en K grupos (clusters) basados en la similitud de las características de los datos. Funciona de la siguiente manera:

1. **Inicialización:** Selecciona aleatoriamente K puntos del conjunto de datos como los centroides iniciales de los clusters.

2. **Asignación de puntos:** Para cada punto en el conjunto de datos, calcula la distancia entre el punto y todos los centroides. Asigna el punto al cluster cuyo centroide está más cercano.

3. **Actualización de centroides:** Recalcula los centroides de los clusters como el promedio de todos los puntos asignados a ese cluster.

4. **Repetición:** Repite los pasos 2 y 3 hasta que no haya cambios significativos en la asignación de puntos a los clusters o se alcance un número máximo de iteraciones.

6. Fundamento matemático

El algoritmo K-Means se basa en minimizar la función objetivo conocida como la suma de los cuadrados de las distancias intra-cluster. Esta función se define como:

$$\sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2$$

donde:

- K es el número de clusters.
- C_i es el conjunto de puntos asignados al cluster i.
- μ_i es el centroide del cluster i.
- $\|x - \mu_i\|$ es la distancia euclidiana entre el punto x y el centroide μ_i.

El objetivo del algoritmo es encontrar la asignación de clusters que minimice esta función. Esto se logra mediante la actualización iterativa de los centroides y la reasignación de puntos a clusters, hasta que se alcanza una convergencia o se alcanza un número máximo de iteraciones.

El algoritmo K-Means es conocido por su eficiencia y simplicidad, pero su rendimiento puede verse afectado por la sensibilidad a la inicialización de los centroides y su dependencia de la elección del número de clusters K.

7. Resultados

Se generó un modelo de Kmeans con un número de clústers igual a 2, además de un estado inicial igual a 42, con esto logramos que solo 92 de 569 datos no tuvieran una predicción correcta de su etiqueta, con un porcentaje de aciertos de un 83.83 %, Podemos observar como se comporta la agrupación en función de diferentes campos en los adjuntos de la última página.

8. Discusión

La maldición de la dimensionalidad hace que el espacio de búsqueda se vuelva más disperso a medida que aumenta el número de dimensiones, dificultando la identificación de agrupaciones coherentes. La interpretación de los resultados puede ser difícil, ya que los clusters pueden estar dispersos en múltiples

dimensiones, lo que dificulta su comprensión. Las características irrelevantes o redundantes pueden afectar negativamente el rendimiento de los algoritmos de agrupación, introduciendo ruido o aumentando la complejidad del modelo. El costo computacional aumenta con el número de dimensiones, especialmente en algoritmos que dependen de cálculos de distancia o densidad.

A pesar de esto se pueden conseguir resultados excelentes bajo la experiencia de las transformaciones y la reducción de dimensiones, hoy en día es muy común encontrarnos con enfoques mixtos de transformaciones y algoritmos, además, existe la posibilidad de utilizar grupos diferentes de agrupación en competición.

Referencias

- Li, W., Cao, Y., Xu, J., Wang, M., & Hu, W. (2020). Deep learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomography images for radiotherapy planning. Radiation Oncology, 15(1), 1-10.

- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

- Wang, S., Zhou, M., Liu, Z., Liu, Z., Guo, H., & Ai, J. (2018). A multi-instance multi-label learning algorithm based on neural networks for protein subnuclear localization prediction. Bioinformatics, 34(10), 1704-1711.

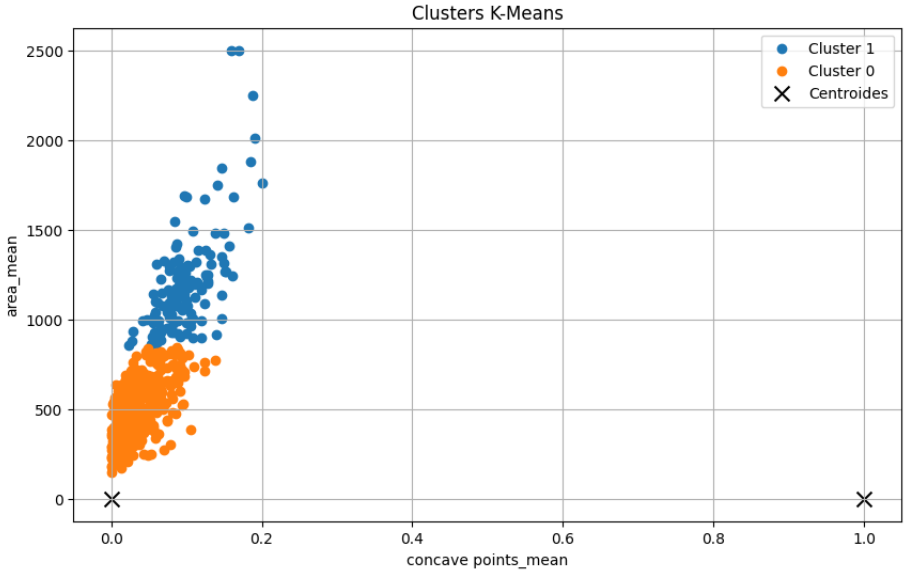

Figura 2: Representación gráfica de los closters generados en función de concave points_mean y area_mean

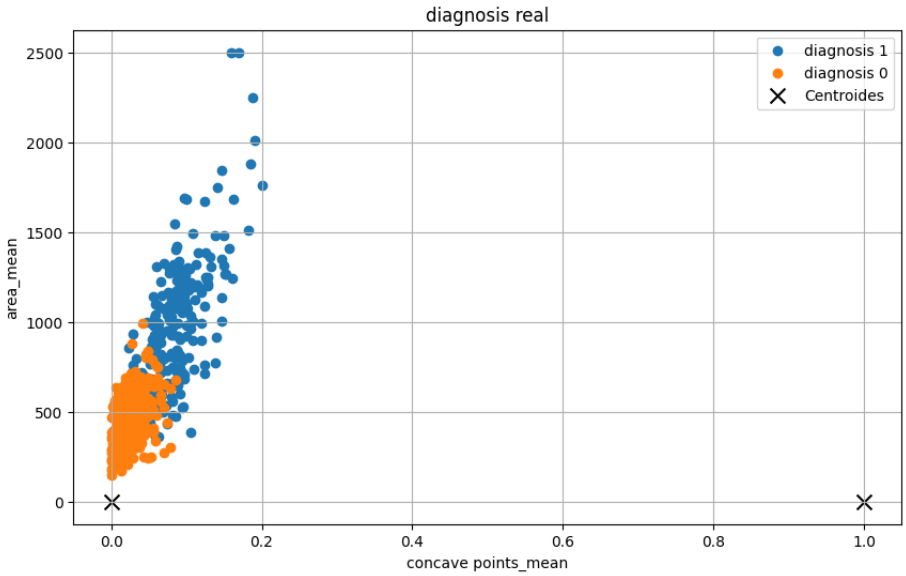

Figura 3: Representación gráfica de la diagnosis real en función de concave points_mean y area_mean

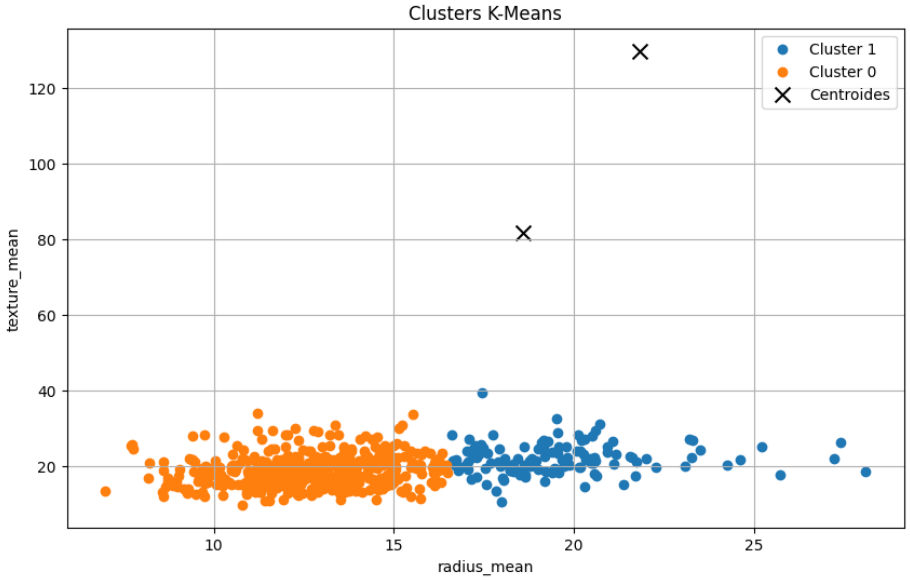

Figura 4: Representación gráfica de los closters generados en función de radius mean y texture mean

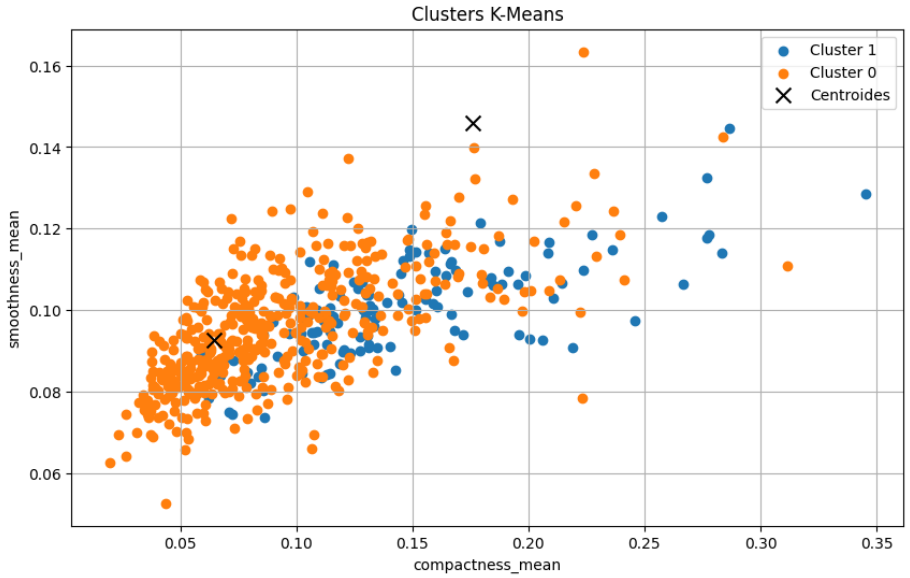

Figura 5: Representación gráfica de los closters generados en función de compactness mean y smoothness mean