

Aplicación de algoritmos supervisados y no supervisados en la clasificación de objetos celestes *

Mendoza Escareño Luis Ramón¹

Asesora :Navarro Jimenez Silvana Guadalupe¹

¹*Departamento de Física, CUCEI, Universidad de Guadalajara*

Bvd. Marcelino García Barragán 1421, Col. Olímpica, Guadalajara Jal., C. P. 44430, México

Resumen

Se pretende utilizar métodos de machine learning para la clasificación de objetos celestes a partir de la información capturada por el proyecto Sloan Digital Sky Survey (SDSS) . Los métodos implementados fueron del tipo supervisado y no supervisado, siendo estos k-vecinos más cercanos y k-means respectivamente.

En el caso de los métodos no supervisados el algoritmo se ejecutó 100 veces, debido a la naturaleza del método, la precisión se sostuvo alrededor del 69 % y en ciertas ocasiones bajó alrededor de 30 %, dando una media de 62.12 % con una desviación estándar de 17.40

Este modelo posee una precisión considerable, abierta a mejoras, se aborda en la sección de discusiones dos posibles razones de la caída de precisión en el modelo.

Por otro lado, el método supervisado de k-vecinos realizado de igual manera 100 veces obtuvo una precisión promedio de 97.19 % y una desviación estándar de 0.0032 en datos destinados a las pruebas.

El modelo generado tiene una excelente precisión para esta base de datos, lo cual hace pensar que la implementación de este modelo pudiera aplicarse para distintas bases de datos con mayor diversidad de objetos.

Introducción

A lo largo de la historia, el humano al observar el firmamento se ha preguntado por la naturaleza de los objetos celestes que lo rodean, únicamente en nuestro vecindario sideral el 'grupo local' existen alrededor de 40 galaxias, la nuestra teniendo un estimado de 300 mil millones de estrellas.

Hoy en día, gracias a la tecnología para captar y analizar la información sobre los cuerpos celestes es posible realizar modelos que permiten el avance en el análisis de diversos fenómenos

*

físicos, promoviendo con ello el avance de la ciencia.

El generar un modelo predictivo no supervisado a partir de los datos pudiera ser de interés para crear y analizar nuevas categorías de objetos celestes, ya que este tipo de modelos predictivos únicamente toman en cuenta la similitud que existe entre los datos que se proveen a partir de una métrica que define su semejanza.

A pesar de la existencia de modelos predictivos para categorizar cuerpos celestes, la implementación de nuevas tecnologías permite contrastar la eficiencia con la que lo hacen. Y debido a que el machine learning es una disciplina en constante crecimiento permite que el modelo sea susceptible a mejoras.

0.1. Proyecto Sloan Digital Sky Survey

El Sloan Digital Sky Survey o (SDSS), es un telescopio óptico dedicado de 2.5 metros de alto situado en el observatorio Apache Point en Nuevo México. Este telescopio ha mapeado un cuarto del cielo en detalle, determinando posiciones, brillo, distancias de más de un millón de galaxias y cuásars.

0.1.1. Filtros del SDSS

Cuando se observa un objeto particular en el telescopio, se puede mejorar el contraste del mismo utilizando filtros apropiados. El filtro bloquea ciertas regiones del espectro electromagnético, midiendo únicamente determinada región (o banda) dentro del espectro electromagnético, cada telescopio detecta el espectro electromagnético y se suele fraccionar en partes con ayuda de estos filtros.

En el caso del SDSS, este tiene 5 filtros (u' , g' , r' , i' y z'), cada uno permite el paso de cierta región (o banda) del espectro electromagnético, estas regiones están dadas de la siguiente forma:

Filtro	Longitud de onda central (Angstroms)	FWHM (Angstroms)
Ultravioleta (u')	3557	599
Verde (g')	4825	1379
Rojo (r')	6231	1382
Infrarrojo cercano (i')	7625	1535
Infrarrojo (z')	9134	1370

Cuadro 1: Longitud de onda central y ancho a potencia media (FWHM) de cada uno de los filtros del SDSS [1]

El **FWHM**, es un parámetro usado comúnmente para describir el ancho del relieve en la curva de una función (en este caso la transmitancia) , está dado por la distancia entre los puntos de la curva en las cuales la función alcanza la mitad del valor máximo.

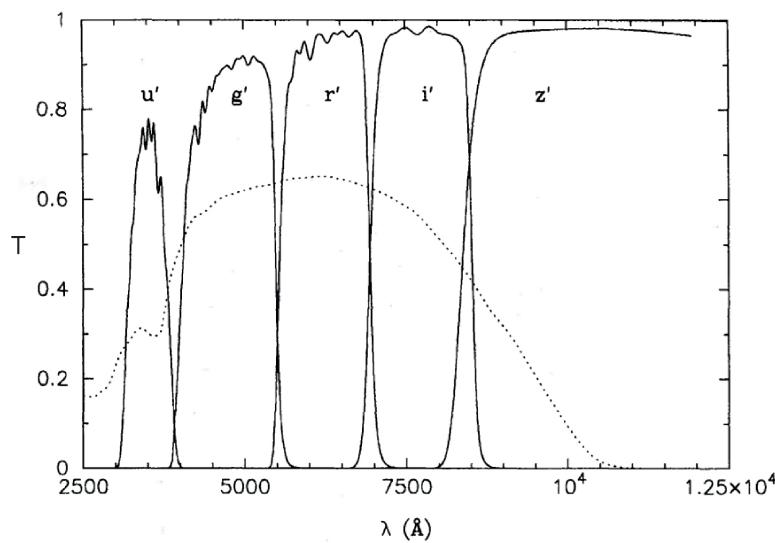


Figura 1: Curva de la transmitancia de los filtros del SDSS, como su nombre lo dice describe la transmisión en los filtros (ópticos o electrónicos) como una función del la longitud de onda o frecuencia de la onda electromagnética.

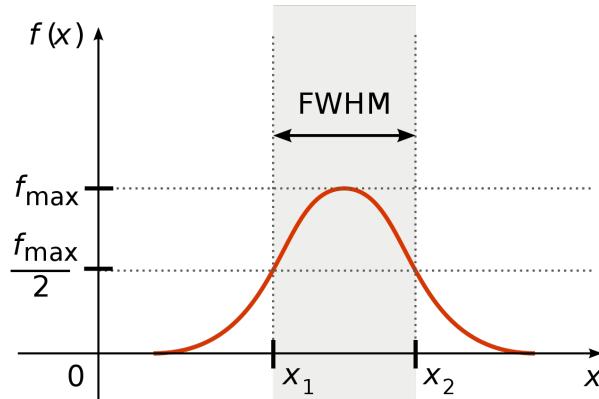


Figura 2: Representación del parámetro FWHM en una curva gaussiana.

0.2. Flujo radiante y el color de los objetos celestes

La "magnitud", es un número que está relacionado con el brillo de una estrella o una galaxia, los valores más altos de este corresponden a objetos con menor brillo y viceversa. Esta magnitud se puede determinar en diferentes bandas de los filtros que tenga el telescopio.

El **color** se define como la diferencia de dos magnitudes, así, existe el color r-i ó el g-i, estas variables son de interés para caracterizar un objeto astronómico.

La magnitud en una banda está caracterizada por el flujo radiante, esto es la cantidad de luz por unidad de tiempo y de área que incide en la Tierra a través de esa banda, esto es:

$$m_x = -2.51 \log_{10} \left(\frac{F_x}{F_{x,0}} \right) \quad (1)$$

Ecuación 1: Definición de la magnitud en la banda x en términos de los flujos radiantes observados en misma banda x. [2]

Con F_x el flujo radiante, el término $F_{x,0}$ se utiliza para definir el cero en la magnitud de la escala, por convención suele usarse el flujo radiante de la estrella Vega.

Entonces la ecuación nos define que para estrellas más brillantes corresponden a valores menores de 'm'. Así es como para cada una de las bandas se puede definir la magnitud en términos de su flujo.

Como se mencionó anteriormente los colores están dados por la diferencia de las magnitudes y gracias a las propiedades de los logaritmos, los colores corresponden al logaritmo del cociente de dos flujos radiantes pertenecientes a cada filtro, esto permite comparar la magnitud de los flujos de cada filtro, lo cual lo hace una variable muy útil para categorizar los objetos celestes.

0.3. Corrimiento al rojo

El corrimiento al rojo o 'redshift' es definido como un incremento en la longitud de onda de radiación electromagnética recibida por un detector comparado con la longitud de onda emitida por la fuente.

En la astronomía y cosmología, hay 3 principales causas del corrimiento al rojo electromagnético:

- Debido a que la radiación viaja entre objetos que están en movimiento relativo.
- La radiación viaja hacia un objeto en un potencial gravitacional más débil.
- La radiación viaja a través del espacio en expansión (corrimiento al rojo cosmológico). La medición en galaxias suficientemente lejanas muestran el corrimiento al rojo correspondiente a su distancia de la tierra, a esto se la llama la ley de Hubble.

Así, el corrimiento al rojo (denotado por z) está descrito por la siguiente ecuación:

$$z = \frac{\lambda_{obs} - \lambda_{emit}}{\lambda_{emit}} \quad (2)$$

Con λ_{obs} el valor de la longitud de onda medida y λ_{emit} la longitud de onda emitida.

0.4. Algoritmo K-Means

El problema de agrupación del tipo no supervisada consiste en lo siguiente: teniendo un conjunto de datos $x^{(1)}, \dots, x^{(m)}$ se quiere agrupar los datos en diferentes grupos. Teniendo cada valor de una observación como un vector $x^{(i)} \in \mathbb{R}^n$ con 'n' la cantidad de variables de cada dato. La meta es encontrar k centroides y asignarle a cada punto x^i un grupo $c^{(i)}$.

Para poder asignar correctamente el grupo de cada punto x^i , se define la **inerzia** de un grupo como la sumatoria de la distancia cada punto al centroide de un grupo al cuadrado

(hacer notar que existen diferentes métricas para definir la distancia entre los puntos y el centroide, en el caso del algoritmo se utilizará la euclídea).

$$Inercia = \sum_{i=1}^m \|x^{(i)} - \vec{\mu}_j\|^2 \quad (3)$$

Con m , la cantidad de puntos en nuestro grupo, y μ_j el centroide del grupo j .

Este valor es el que nuestro algoritmo intentará minimizar para cada uno de los grupos, entonces el algoritmo es el siguiente:

Inicializar los centroides de cada cluster $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$ de manera aleatoria.

Repetir hasta la convergencia:

Para cada i :

$$c^{(i)} := argmin_j \|x^{(i)} - \mu_j\|^2 \quad (4)$$

Se define el grupo $c^{(i)}$ para cada punto x^i en base a la menor 'distancia' de los centroides μ_i con el punto.

Se reasigna el punto del centroide, esto es que para cada j :

$$\vec{\mu}_j := \frac{\sum_{i=1}^m \hat{x}^{(i)}}{m} \quad (5)$$

Con \hat{x} todos los vectores de x que pertenecen al grupo $c^{(j)}$ y m el número de elementos de esta categoría.

0.5. Índice de pureza Gini

El índice Gini, o también conocido como impureza gini, calcula la probabilidad de que una característica específica sea clasificada incorrectamente cuando se selecciona de manera aleatoria. Si todos los elementos del grupo son de un único grupo entonces se le llama 'puro'.

Este índice tiene valores entre 0 y 1, donde el 0 expresa la pureza máxima del grupo y más cercano al 1, significa un grupo más homogéneo (igual cantidad de valores de cada categoría en el grupo). El índice Gini se define de la siguiente manera:

$$Gini = 1 - \sum_{i=1}^m P_i^2 \quad (6)$$

Con P_i la proporción de elementos de la categoría 'i' en el grupo, esto es el número de elementos clasificados correctamente entre la cantidad total de elementos.

Metodología

Se utilizó la base de datos del 'Sloan Digital Sky Survey DR14' provista por la página Kaggle [3] , en esta base de datos se tienen 10,000 datos de objetos estelares, estos son 4998 galaxias, 4152 estrellas y 850 cuásares. Cada objeto cuenta con 18 variables que lo caracteriza.

Con el uso del lenguaje de programación python, acompañado de la librería pandas, se procedió a la extracción de datos, en esta se observaron las columnas en las cuales se tiene información como: códigos de identificación de los objetos, fechas de avistamiento, clasificación de los objetos, etc. De las variables provistas, las que resultaron de interés fueron:

- Los valores de la magnitud de los 5 filtros del telescopio
- Los valores del corrimiento al rojo de cada observación.
- Las clasificaciones de cada objeto.

Las clasificaciones al estar escritas en las categorías "STAR", "GALAXY", "QUASAR" y , se cambiaron por los números 0, 1 y 2 respectivamente, para así poder utilizarlas en los siguientes pasos.

Se utilizaron las variables de la magnitud de los filtros del telescopio para crear las variables de colores antes mencionada a partir de la diferencia entre magnitudes.

0.6. K-means

Se procedió a re-escalar las variables de interés, esto es dividir cada valor de la variable entre el máximo para que los valores se encuentren entre 0 y 1. Seguido a esto se manipularon los datos a través de la librería pandas de la siguiente manera:

Se tomaron las características de los colores (las diferencias de las magnitudes de cada filtro) y la del corrimiento al rojo como variables y a través de la librería sci-kit learn de python se ejecutó el algoritmo de k-means.

Ya que los clusters creados por el algoritmo inicializan los centroides al azar, a cada grupo se le asignó un valor de 0-2 el cual no tiene por qué coincidir con los valores que se asignaron inicialmente a cada categoría, para realizar la conexión entre categorías que se asignaron inicialmente (i) y las que asignó el algoritmo (a) se realizó lo siguiente:

Para cada elemento de cada grupo (i) se contaron las coincidencias con los grupos (a), se tomó que el grupo de (a) que tuviera más coincidencias con cada grupo (i) sería la relación entre los grupos y se pasó a cambiar el valor de (a) al respectivo de (i). Un ejemplo sería que si el algoritmo tomaba con el valor '2' el grupo de las estrellas y el que inicialmente escogimos es el '0', cambiamos la categoría del valor '2' al '0' para así poder calcular la **precisión** del algoritmo, esto lo definimos

como el número de clasificaciones correctas del algoritmo entre el número total de elementos.

Se generaron funciones dentro del programa para que se corriera el algoritmo, se re-clasificaran las categorías, se evaluaran los valores de precisión, índice de inercia de cada grupo y el índice Gini. Después se realizó un ciclo que ejecutara 100 veces la función antes mencionada y guardara los valores de interés (precisión, índice de inercia, etc.) para cada ejecución, esto con el fin de determinar un promedio y desviación estándar de cada valor.

0.7. K-vecinos más cercanos

Para el algoritmo supervisado, de igual manera que en el anterior provino de la librería sci-kit learn, se realizó 100 veces el algoritmo utilizando como variables los valores de las diferencias de los filtros y el de desplazamiento al rojo.

Se usó el 80 % de los datos en el entrenamiento y el restante para la validación del modelo, guardando cada uno de los valores de la precisión y los índices Gini de cada grupo, se promediaron y calcularon las desviaciones estándar de cada valor.

Resultados

0.8. Método K-Means

Los promedios junto a su respectiva desviación estándar fueron los siguientes

- Precisión: 62.12 % con $\sigma = 17.40$
- Inercia: 137.53 y con $\sigma = 4.73 * 10^{-5}$
- Índice Gini del grupo estrellas: 0.324 con $\sigma = 6.99 * 10^{-5}$
- Índice Gini del grupo galaxias: 0.498 con $\sigma = 4.32 * 10^{-5}$
- Índice Gini del grupo cuásares: 0.147 con $\sigma = 2.897 * 10^{-17}$

0.9. Método K-Vecinos más cercanos

Para este método únicamente se tomó el promedio y la desviación estándar de la precisión, la cual fue la siguiente:

- Precisión: 97.19 % con $\sigma = 3.2 * 10^{-3}$
- Índice Gini del grupo estrellas: 0.0224 con $\sigma = 3.46 * 10^{-18}$
- Índice Gini del grupo galaxias: 0.0621 con $\sigma = 6.93 * 10^{-18}$
- Índice Gini del grupo cuásares: 0.0746 con $\sigma = 1.38 * 10^{-17}$

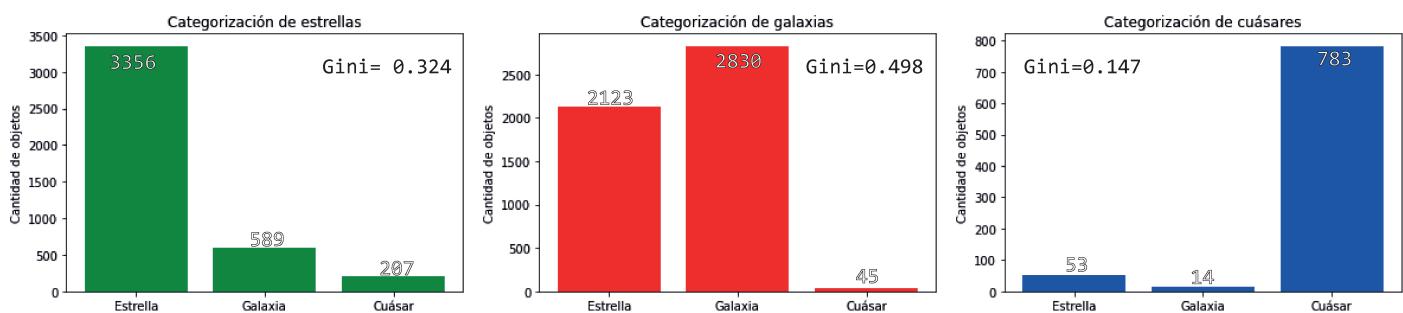


Figura 3: Categorización de los grupos por el método k-means, en este ejemplo tomado de una de las iteraciones se obtuvo una precisión del 69.69 %, la cual es la mayor precisión que se obtiene del modelo, se observa que en el grupo de las galaxias, se tiene el peor rendimiento ya que el número de categorizaciones de estrellas y galaxias es alto, esto se refleja en el valor del índice Gini el cual es el mayor de los grupos.



Figura 4: Categorización de los grupos por el método k-vecinos, en este ejemplo se obtuvo un 97.6 % de precisión, se observa que cada grupo es bastante homogéneo, lo cual se refleja en los valores bajos del índice Gini.

Discusión

Analizando los valores de la precisión en el modelo no supervisado se observó que en su mayoría son valores alrededor del 69 %, sin embargo en pocos casos pasa que la precisión baja hasta el 30 %, esto puede ser causado por la naturaleza del algoritmo ya que inicialmente se colocan arbitrariamente los puntos de los centros de los grupos, podría ser que hay configuraciones en las cuales no se agrupan adecuadamente antes de converger a un valor de la inercia en específico.

Otra hipótesis acerca del error, es que este radica en la asignación de los valores de los grupos generados por el algoritmo no supervisado, esto es en el paso para relacionar los valores de los grupos inicialmente categorizados (i) con los valores de categorización que provee el algoritmo k-means (a), se cree esto debido a el valor casi nulo de la desviación que tienen los parámetros Gini de cada grupo. Si este error de asignación de valores fuera cierto, podría ser el causante del descenso aparente de la precisión del modelo predictivo.

Conclusiones

Debido al buen rendimiento que se mostró en el caso supervisado, este método de generación de modelos predictivos pudiera ser de interés para desarrollar modelos con mayor diversidad de objetos celestes.

Hablando del caso no supervisado, excepto por los casos 'extraños' en los que la precisión baja, también se tiene un rendimiento considerable abierto a mejoras. El hecho de que en este sólo se cuente la semejanza entre los valores (descrita por una métrica) de las categorías nos podría permitir usarla como una herramienta para crear nuevas categorías estelares a partir de los datos sin clasificar.

Ambos casos son de interés ya que, estos son sólo un ejemplo de los algoritmos supervisados y no supervisados, con la inclusión de más variables de valor, nuevos algoritmos especializados, bases de datos con más ejemplos y diversidad de objetos podrían ser de ayuda para mejorar la precisión de los modelos.

Referencias

- [1] Fukugita, M., Ichikawa, T., Gunn, J. E., Doi, M., Shimasaku, K., Schneider, D. P. (1996). The Sloan Digital Sky Survey Photometric System. En The Astronomical Journal (Vol. 111, p. 1748). American Astronomical Society. <https://doi.org/10.1086/117915>
- [2] Karttunen, H. (2003, 13 agosto). Fundamental Astronomy (4th ed.). Springer.
- [3] <http://skyserver.sdss.org/dr1/en/proj/advanced/color/amounts.asp> (Página oficial del SDSS)
- [4] Full Width at Half Maximum – from Wolfram MathWorld. (s. f.). Recuperado 6 de octubre de 2022, de <https://mathworld.wolfram.com/FullWidthatHalfMaximum.html>
- [5] Wikipedia contributors. (2022, 13 septiembre). Redshift. Wikipedia. Recuperado 6 de octubre de 2022, de <https://en.wikipedia.org/wiki/Redshift>
- [6] <https://www.kaggle.com/datasets/lucidlenn/sloan-digital-sky-survey> (Dataset del SDSS)
- [7] <https://colab.research.google.com/drive/18fBOqDkIgIPjU8sMzHbCTfM-cAnkVwCG?usp=sharing> (Código en google colab)
- [8] CS221. (s.f.). Recuperado 2 de agosto de 2022, de <https://stanford.edu/%7Ecpiech/cs221/handouts/kmeans.html>
- [9] Tyagi, N. (2021, 13 diciembre). Understanding the Gini Index and Information Gain in Decision Trees. Medium. Recuperado 6 de agosto de 2022, de <https://medium.com/analytics-steps/understanding-the-gini-index-and-information-gain-in-decision-trees-ab4720518ba8>

Nombre y Firma del Asesor¹