Aprendizaje Automático Relacional - Predicción de delitos en una banda londinense

Juan Manuel García Criado
Dpto. Ciencias de la Computación e Inteligencia Artificial
Universidad de Sevilla
Sevilla, España
juagarcri@alum.us.es

Fernando Miguel Hidalgo Aguilar
Dpto. Ciencias de la Computación e Inteligencia Artificial
Universidad de Sevilla
Sevilla, España
ferhidagu@alum.us.es

***Resumen*—El presente documento presenta un problema de clasificación relacional, aprendiendo a elaborar una función que dados los atributos de un nodo de una red, esta sea capaz de clasificar correctamente dicho nodo entre una serie de clases.**

En concreto, el problema radica en conocer si nuevos miembros de una banda conflictiva de Londres acabará yendo o no a la cárcel, partiendo de un grafo de los participantes de un secuestro.

Las herramientas usadas serán las librerías sklearn y networkx en conjunto con atributos relacionales de los grafos como la centralidad, el coeficiente de clustering o el grado de entrada y salida de cada nodo.

I. Introducción

El ser humano es social por naturaleza, y desde los albores de su creación, ha buscado asociarse con otros similares, aquellos con los que comparte gustos, ambiciones y misma visión de la vida.

Por desgracia, estos rasgos en común no tienen por qué ser necesariamente positivos, y ser el crimen y sus derivados la razón de union de las personas.

Y ese es el caso que nos ocupa: Una banda criminal de Londres. Vamos a realizar una investigación en el periodo entre 2005 y 2009, mediante datos sacados de informes de arrestos y condenas policiales de los miembros de la banda.

Los datos de los que disponemos son:

1) Edad
2) Lugar de nacimiento
3) Residencia
4) Arrestos
5) Arrestos
6) Previamente en prisión
7) Música

Además de las relaciones entre ellos que pueden ser:

1) Pasar el rato
2) Cometer delitos
3) Ser parientes

El objetivo del estudio es conocer que características son relevantes y que grado de relación con los otros miembros debe estar un individuo para acabar en prisión. Gracias a estos resultados, tendremos una herramienta que permitirá predecir si un nuevo miembro dará con sus huesos en el calabozo.

Para conseguirlo, contamos con 2 ficheros .csv, uno de ellos la matriz de adyacencia y otro con los atributos de cada nodo.

II. Preliminares

A. Métodos empleados

- Validación Cruzada: Técnica que evalúa los resultados de un análisis estadístico, garantizando la independiencia entre datos de entrenamiento y prueba. Usamos esta técnica para calcular diferentes medidas de evaluación sobre diferentes particiones para así obtener un conjunto de datos y modelo adecuado.

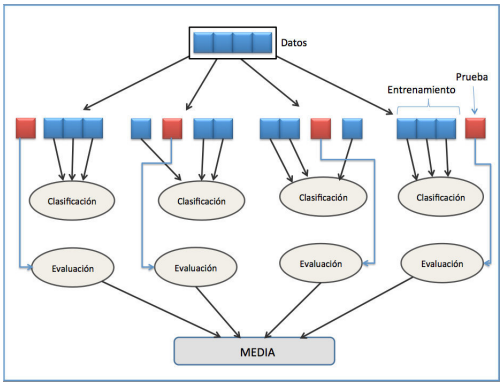


Fig. 1. Funcionamiento Validación Cruzada

- Naive Bayes: Basado en el teorema de Bayes, es un clasificador probabilístico que supone independencia entre las variables predictoras, es decir, es ingenuo (naive).

  Para trabajo que nos ocupa, se usará Naive Bayes para estudiar si los nodos del grafo creado (miembros de la banda) fueron a prisión.

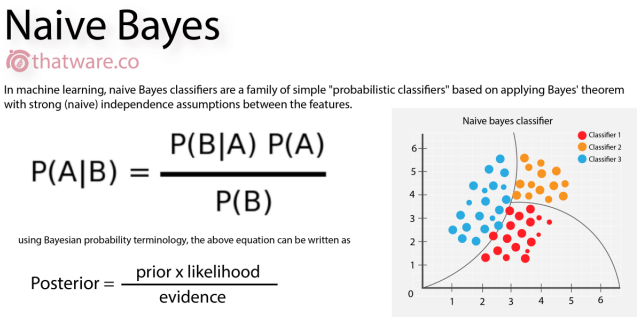


Fig. 2. Funcionamiento Naive Bayes

- KNN: También conocido como "k-vecinos más próximos", nos servirá para construir un modelo probabilístico a partir de los ejemplos que dispone el modelo de datos.

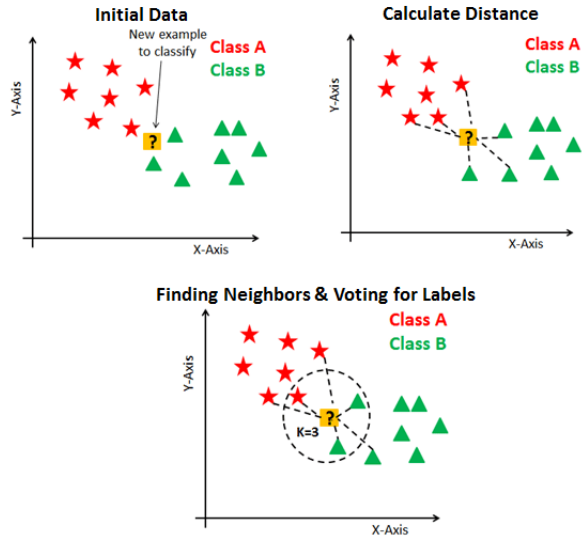


Fig. 3. Funcionamiento KNN

Clasificaremos los nuevos ejemplos en función de las categorías de los ejemplos más cercanos, según la distancia entre los mismos. En concreto, hemos usado la distancias euclídea, mediante de su fórmula.

$$d_{ij} = \sqrt{\sum_{p=i}^{k} (x_{ip} - x_{jp})^2}$$

Fig. 4. Distancia Euclídea

- Árbol de decisión: Algoritmo de clasificación cuyo objetivo es crear un modelo que prediga el valor de una variable objetivo, mediante el aprendizaje de reglas de decisión simples extraídas de los atributos de los datos.

  Mediante la librería sklearn se puede exportar el árbol resultante, en el que se puede distinguir de manera visual ciertos atributos relacionados con los datos con los que se trabaja. El árbol está formado por nodos interiores (atributos), arcos (posibles valores del nodo del que se origina) y hojas (valores resultantes).

  Importante mencionar que cuanto más profundo es el árbol, más complejas son las reglas de decisión usadas y más ajustado será el modelo, aunque árboles demasiado complejos puede provocar una incorrecta generalización de los datos, lo que se traduce en un sobreajuste.

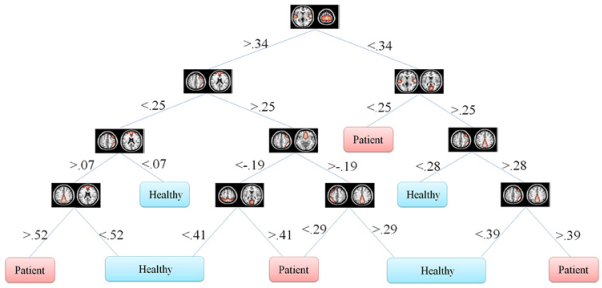


Fig. 5. Funcionamiento Arboles de Decisión

- Regresión Lineal: Técnica de modelado estadístico que se emplea para describir una variable de respuesta continua como una función de una o varias variables predictoras. Puede ayudar a comprender y predecir el comportamiento de sistemas complejos o a analizar datos experimentales

$$Y = \beta_0 + \sum \beta_i X_i + \epsilon_i$$

Fig. 6. Regresión Lineal

- Regresión Logística: Tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo en función de otros factores

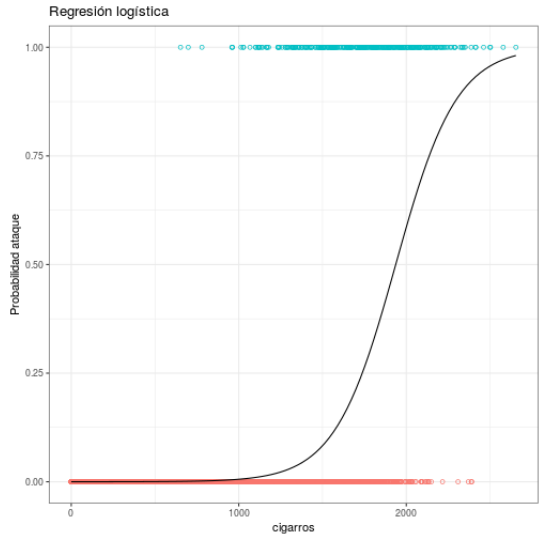
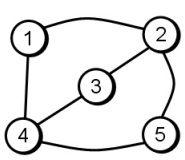


Fig. 7. Regresión Lineal

B. *Datos empleados*

- Edad: Marca los años que tienen el miembro.
- Lugar de nacimiento:
  - 1: África Oeste
  - 2: Caribe
  - 3: Gran Bretaña
  - 4: Africa Este
- Residencia: Marca si el miembro cuenta o no con residencia propia.
- Arrestos: Marca el número de arrestos del miembro.
- Condenas: Marca el número de condenas del miembro.
- Prisión: Marca si el miembro ha ido o no a prisión.
- Música: Marca si el miembro escucha o no música.
- Ranking: Posición del miembro dentro de la banda. Existen valores entre 1 y 5, siendo dicha posción mejor cuanto menor sea el valor de este atributo. Es decir, 1 mejor que 5

- Matriz de adyacencia: Matriz cuadrada conformada por el número de nodos del grafo. Cada celda de la matriz puede tener valores del 0 al 4, representando la relación entre los diferentes miembros de la banda.
  - 0: Sin relación
  - 1: Compartían su tiempo
  - 2: Cometían fechorías menores
  - 3: Ejecución de delitos más graves
  - 4: Delitos graves o los miembros son parientes



M	1	2	3	4	5
1	0	1	0	1	0
2	1	0	1	0	1
3	0	1	0	1	0
4	1	0	1	0	1
5	0	1	0	1	0

Fig. 8. Matriz de Adyacencia

C. *Métricas relacionales empleadas*

- Grado: Llamado ”valencia de un nodo”, es el número de aristas que inciden sobre el nodo, lo que traducido a el caso que nos ocupa, el número de miembros con los que este se relaciona.

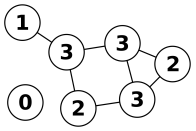


Fig. 9. Grado

- Clustering: Cuantifica que tanto está interconectado un nodo con sus vecinos.

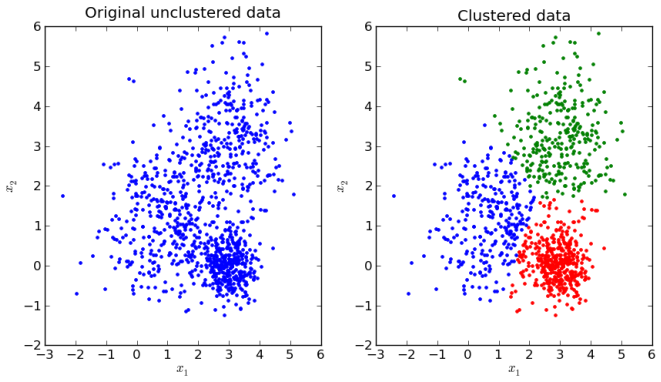


Fig. 10. Clustering

- Centralidad: Atributo relacional que realiza una medida estimada de un nodo por la cual se puede conocer la importancia o relevancia del nodo en ese grafo. Conocer la centralidad de un nodo ayuda a determinar el impacto que este causa dentro del conjunto del que forma parte.

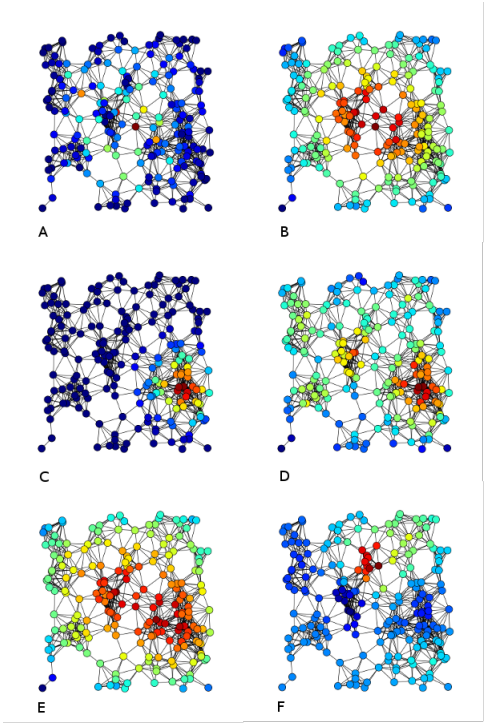


Fig. 11. Centralidad

III. Metodología

1) Primero leemos los datos del archivo LONDON-GANG.csv, quitando la primera fila y columna, ya que corresponden a los índices de la hoja, no son datos reales. Destacar que no se ha aplicado procesado a los datos de entrada.

2) Creamos un grafo con la matriz de adyacencia y comprobamos que se ha creado correctamente, numerado en nuestro caso concreto.

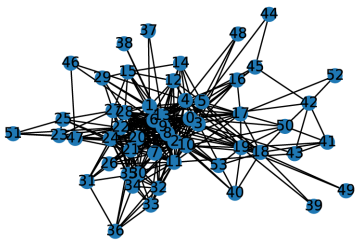


Fig. 12. Grafo Londres

3) Obtenemos los atributos relacionales, dados en un diccionario, asi que cogemos los valores y los ponemos en forma de lista. Esto se aplica a las 3 centralidades.

- De Grado:Numero de nodos unidos a ese nodo

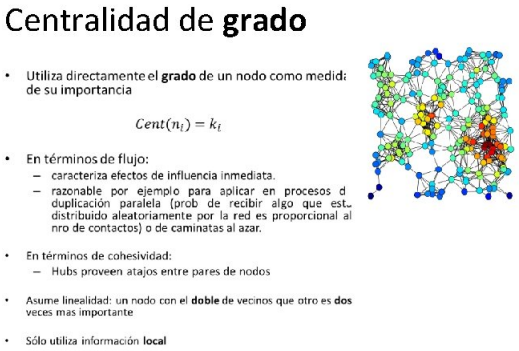


Fig. 13. De Grado

- Intermedia: Número de nodos que se unen mediante el nodo que se estudia.
- Katz: Participación de un nodo a la hora de generar las relaciones de una red.

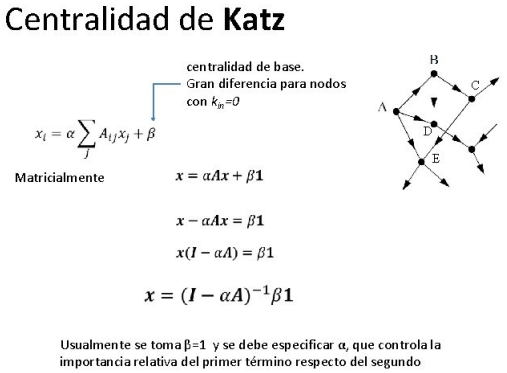


Fig. 14. Katz

4) Procedemos ahora a emplear los métodos elegidos.
   - *Naive Bayes*: Naive Bayes asume que la presencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra característica.

- *Regresión Lineal*: Permite aproximar la relación de dependencia entre una variable. Destacar que este método no está preparado para dar una clasificacion basada en etiquetas, es decir, no devuelve resultados exactos. Para adaptarlo a nuestras necesidades, se ha aplicado un suavizado a los resultados, dejando los resultados en solo 1 o 0.
- *Regresión Logística*: La regresión logística puede usarse para tratar de correlacionar la probabilidad de una variable cualitativa binaria (”0” y ”1”) con una variable escalar x. La idea es que la regresión logística aproxime la probabilidad de obtener ”0” (no ocurre cierto suceso) o ”1” (ocurre el suceso) con el valor de la variable x.
- *Arboles de Decisión*: En el contexto de este trabajo, los arboles de decisiones busca reducir la entropía para cada nivel del arbol. Se aplicado tanto el criterio ”entropia” como ”gini”, siendo ”entropia” el que presenta resultados mas satisfactorios.
- *K Vecinos Mas Cernanos (K Nearest Neighbors, KNN)*: Dado el numero de vecinos y una distancia, muestra como de conectada se encuentra la red. Hemos optado por usar distancia euclídea, dado que la distancia de Hammng no se ajusta al problema en cuestión.

IV. Resultados

- *Naive Bayes*: Notamos que se obtienen el mejor resultado cuando todos los atributos relacionales y originales están presentes

```
NAIVE BAYES
Acierto con atributos originales para Prison : 0.643
Acierto con todos los atributos para Prison : 0.714
Acierto con los atributos relacionales solamente, para Prison:  0.5
Acierto con atributos originales  mas Centralidad_grado para Prison : 0.643
Acierto con atributos originales  mas Centralidad_intermedia para Prison : 0.643
Acierto con atributos originales  mas Centralidad_katz para Prison : 0.643
Acierto con atributos originales  mas Grados para Prison : 0.643
Acierto con atributos originales  mas Clusters para Prison : 0.571
```

Fig. 15. Resultados Naibe Bayes

- *Regresión Lineal Múltiple*: En este caso, el mejor resultado corresponde a usar solo los atributos originales

```
Regresión Lineal
Acierto con atributos originales para Prison : 0.786
Acierto con todos los atributos para Prison : 0.643
Acierto con los atributos relacionales solamente, para Prison:  0.571
Acierto con atributos originales  mas Centralidad_grado para Prison : 0.643
Acierto con atributos originales  mas Centralidad_intermedia para Prison : 0.643
Acierto con atributos originales  mas Centralidad_katz para Prison : 0.643
Acierto con atributos originales  mas Grados para Prison : 0.643
Acierto con atributos originales  mas Clusters para Prison : 0.643
```

Fig. 16. Resultados Regresion Lineal Múltiple

- *Regresión Lineal Logística*: Para CV, el mejor resultado corresponde a usar solo los atributos originales unido al atributo relacional de Grados.

```
Regresión Logistica Cross-Validation
Acierto con atributos originales para Prison : 0.786
Acierto con todos los atributos para Prison : 0.786
Acierto con los atributos relacionales solamente, para Prison:  0.571
Acierto con atributos originales  mas Centralidad_grado para Prison : 0.786
Acierto con atributos originales  mas Centralidad_intermedia para Prison : 0.786
Acierto con atributos originales  mas Centralidad_katz para Prison : 0.786
Acierto con atributos originales  mas Grados para Prison : 0.857
Acierto con atributos originales  mas Clusters para Prison : 0.786
```

Fig. 17. Resultados Regresion Lineal Cross Validation

Para este TS, se obtiene el mejor resultado en dos casos, esto es, usando sólo los atributos originales y tambien con los atributos originales más los relacionales

```
Regresión Logistica Test-Split
Acierto con atributos originales para Prison : 0.714
Acierto con todos los atributos para Prison : 0.714
Acierto con los atributos relacionales solamente, para Prison:  0.571
Acierto con atributos originales  mas Centralidad_grado para Prison : 0.643
Acierto con atributos originales  mas Centralidad_intermedia para Prison : 0.643
Acierto con atributos originales  mas Centralidad_katz para Prison : 0.571
Acierto con atributos originales  mas Grados para Prison : 0.643
Acierto con atributos originales  mas Clusters para Prison : 0.643
```

Fig. 18. Resultados Regresion Lineal Test-Split

- *K Vecinos Cernanos (K Near Neighbors, KNN)*: Por último, el mejor resultado se presenta usando los atributos originales en conjunción con la centralidad de Grado

```
KNN - Euclidea

Acierto con atributos originales para Prison : 0.563

Acierto con todos los atributos para Prison : 0.563

Acierto con los atributos relacionales solamente, para Prison:  0.473

Acierto con atributos originales  mas Centralidad_grado para Prison : 0.68

Acierto con atributos originales  mas Centralidad_intermedia para Prison : 0.66

Acierto con atributos originales  mas Centralidad_katz para Prison : 0.66

Acierto con atributos originales  mas Grados para Prison : 0.543

Acierto con atributos originales  mas Clusters para Prison : 0.66
```

Fig. 19. Resultados KNN

- *Arboles de Decisión*: Aquí encontramos el mejor resultado usando los atributos originales en conjunción con la centralidad de Katz.

```
Arboles de decisión

Acierto con atributos originales para Prison : 0.5

Acierto con todos los atributos para Prison : 0.5

Acierto con los atributos relacionales solamente, para Prison:  0.429

Acierto con atributos originales  mas Centralidad_grado para Prison : 0.429

Acierto con atributos originales  mas Centralidad_intermedia para Prison : 0.5

Acierto con atributos originales  mas Centralidad_katz para Prison : 0.786

Acierto con atributos originales  mas Grados para Prison : 0.429

Acierto con atributos originales  mas Clusters para Prison : 0.357
```

Fig. 20. Resultados Arbol de Decisión

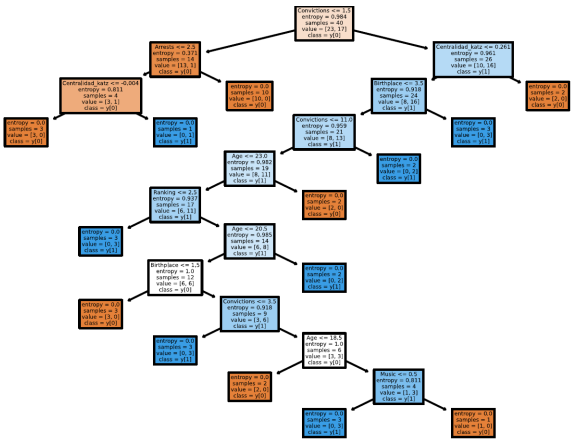


Fig. 21. Arbol resultado

V. Conclusiones

Las conclusiones que podemos sacar, es que no existe un atributo que marque como tal la diferencia sobre el resto, sino que dependiendo del método usado uno de los atributos despuntará en sus resultados.

Sí podemos marcar con certeza como KNN no presenta este comportamiento, siendo el método menos esclarecerdor a la hora de presentar que atributo considera de mayor relevancia.

Nos hemos percatado de que a pesar que el consumir música puede parecer trivial, hemos decido cercionarnos de que tanto peso aproximadamente podria tener dicho atributo en los datos mediante los coeficientes de decision.

Podemos observar que Music tiene un peso considerable, si bien retirar el atributo mejora algunos resultados en la mayoria de casos los empeora.

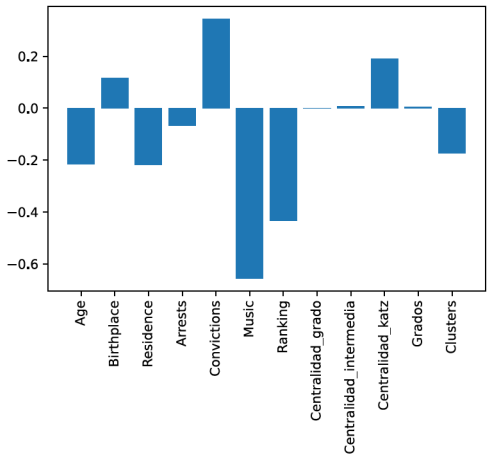


Fig. 22. Coeficiente de Music

Asímismo nos gustaría detallar que el código se ha desarrollado de forma que sea reutilizable si se decide cambiar de objetivo. Además, el desarrollo del código nos ha brindado la oportunidad de conocer más de las librerias networkx y scikit-learn.

Como comentario final, y en vista a nuestra experiencia, nos hubiese gustado recibir algún tipo de seguimiento obligatorio, para evitar sorpresas desagradables en el momento de recibir las calificaciones del trabajo.

Referencias

[1] Networkx Documentation [networkx.org/]
[2] Scikit-Learn Machine Learning in Python [scikit-learn.org]
[3] Tema 2 - Aprendizaje automático, Universidad de Sevilla
[4] UCINET Software, London Gang Dataset [https://sites.google.com/site/ucinetsoftware/datasets/covert-networks/londongang]
[5] UCINET Software, London Gang Dataset [https://sites.google.com/site/ucinetsoftware/datasets/covert-networks/londongang]
[6] Aprende con Alf Recursos Manual Python La librería Pandas: [https://aprendeconalf.es/docencia/python/manual/pandas/]
[7] Wikipedia:[https://es.wikipedia.org/wiki]
[8] Librería Pandas: [https://pandas.pydata.org/pandas-docs/stable/index.html]