Algoritmo ID3 Mixto

QUÓRUM E INCERTIDUMBRE

Objetivo

Versión modificada del algoritmo ID3 que nos de como resultado un árbol de decisión.

Cada rama del árbol podrá ser:

- Rama sin incertidumbre
- Valor de clasificación con probabilidad asociada (En caso de no cumplir quórum)

Elementos del algoritmo

1) Algoritmo ID3

2) Quórum

3) Naive Bayes

Algoritmo ID3

- Generar árboles de decisión a partir de un conjunto de datos de entrenamiento

- Basado en la elección de mejor atributo. Será establecido mediante entropía

  - $Ent(D) = -\frac{|P|}{|D|}\log_2\frac{|P|}{|D|} - \frac{|N|}{|D|}\log_2\frac{|N|}{|D|}$

- Se irá eligiendo aquel atributo que tenga una mayor ganancia

  - $Ganancia(D, A) = Ent(D) - \sum_{v \in Valores(A)} \frac{|Dv|}{|D|} Ent(Dv)$

Quórum

Número entero positivo que nos indica la cantidad mínima de ejemplos que queremos imponer para considerar que una rama del árbol es fiable

Naive Bayes

- Clasificador simple de la familia de clasificadores probabilísticos .

- Basado en la aplicación del teorema de Bayes con ciertas modificaciones

  - $\text{argmax}_{vj \in V} P(vj) \prod_i P(ai|vj)$

  - $P(vj) = \frac{\#(V=vj)}{N}$

  - $P(ai|vj) = \frac{\#(Ai=ai, V=vj)+k}{\#(V=vj)+k|Ai|}$

Metodología

- **Representación árbol mixto**
- **Datos**
- **Algoritmo ID3 modificado**
- **Clasificación nuevos datos**
- **Naive Bayes**
- **Rendimiento**

Representación árbol mixto

- Estará compuesto por nodos, de esta forma el árbol será una lista de nodos.

- Nodos representarán: atributos del conjunto de datos, clasificaciones o llamadas al método Naive Bayes.

Atributos de los nodos

- **Nombre:** nombre del atributo al que representa.

- **Hijos:** lista de nodos que están relacionados con este.

- **Arista:** valor del atributo al que representa el nodo padre de este que habría que tomar.

- **Padre:** Nodo superior a él en el árbol.

Otros atributos de árbol mixto

- **Atributos:** lista de los atributos del conjunto de datos sin el atributo objetivo.

- **Datos entrenamiento:** representados como un dataframe de la librería Pandas.

- **Datos de evaluación:** representados como un dataframe de la librería Pandas.

- **Clasificaciones:** lista con las clasificaciones para los datos de evaluación (en caso de realizarse llamada a Naive Bayes también tiene la probabilidad).

Datos

- Para representar datos se ha utilizado la librería Pandas, concretamente el tipo de datos Dataframe.

- Deberán estar en formato .csv y los atributos deberán estar colocados de forma que la última columna contenga el atributo objetivo.

- Se le deberá indicar al programa si los datos contienen en la primera línea del archivo los nombres de los atributos. En caso de tenerlos se utilizarán como nombre de los nodos interiores, en caso contrario se usarán índices.

Algoritmo ID3 modificado

Para que la creación del árbol se ha implementado un método para que el usuario solo tenga que indicar si el archivo con los datos tiene en la primera línea los nombres de los atributos, la ruta del archivo y el valor del quórum. Este método inicializa los atributos necesarios del árbol y realiza la llamada inicial al método que realmente ejecuta el algoritmo ID3 modificado para construir el árbol.

Algoritmo ID3 modificado

- Entrada

- Salida

-Algoritmo

Entrada

- Datos

- Quórum

- Nodo anterior (nulo en primera llamada): se utiliza en la creación de nuevos nodos.

- Valor anterior (nulo en primera llamada): se utiliza para asignar arista en creación nuevos nodos.

- Atributos

Salida

- No tiene, ya que va creando los nodos y los va añadiendo al atributo nodos del árbol mixto.

Algoritmo

1) Si longitud(datos) < quorum:

- Crear nodo para llamada Naive Bayes

2) Si longitud(valores de ultima columna) == 1:

- Crear nodo con el valor que contenga la ultima columna

3) Si longitud(atributos) <= 0:

- Crear nodo con el valor más común de la última columna en datos

Algoritmo

4) En otro caso:

- - Obtener mejor atributo
- - Crear nodo con mejor atributo
- - nodo = nodo creado en paso anterior
- - Para valor en los valores del mejor atributo:
  - 1) nuevos datos = Obtener datos con el valor del mejor atributo
  - 2) Copiar atributos
  - 3) Eliminar el mejor atributo de la copia de atributos
  - 4) Si longitud(nuevos datos) == 0 y quorum = 0:
    - Crear nodo con el valor más común de datos
  - 5) En otro caso:
    - Hacer la llamada recursiva con datos = nuevos datos, quórum, nodo, valor, copia atributos.

Clasificación de nuevos datos

Se ha creado un método que recibe la información sobre los datos (si los atributos están especificados en la primera línea del archivo y la ruta del archivo) y es este el que realiza la llamada al método que obtiene la clasificación del nuevo ejemplo. El primer método obtiene los datos de la ruta indicada, inicializa las variables correspondientes y para cada fila en los datos llama al método que obtiene la clasificación.

Clasificación de nuevos datos

- Entrada

- Salida

-Algoritmo

Entrada

- Fila a clasificar

- Nodo actual (en la llamada que realiza el método mencionado anteriormente es el nodo raíz)

Salida

Clasificación de la fila, en caso de que se realice la llamada al método Naive Bayes devuelve una lista con la clasificación en la primera posición y la probabilidad en la segunda(Posiciones 0 y 1 en la lista).

Algoritmo

1) Si el nodo es nulo:

- Devolver nulo

2) Si el nombre del nodo == “Naive Bayes”:

- Realizar llamada al método Naive Bayes y devolver lo que devuelve este

3) Si longitud(hijos del nodo) == 0:

- Devolver el nombre del nodo

4) En otro caso:

- Encontrar el hijo del nodo actual que tiene como arista el valor de la columna que tiene el nombre igual al nombre del nodo actual en la fila.
- Realizar llamada recursiva con el nodo encontrado en el paso anterior.

Naive Bayes

- Se ha realizado un método que haga los cálculos en lugar de utilizar alguna librería.

- El método calcula la probabilidad con suavizado de Laplace para evitar probabilidades nulas.

- Se utilizan las fórmulas indicadas en la explicación de Naive Bayes realizada al inicio.

Rendimiento

- **Medida de rendimiento:** tasa de aciertos.

Resultados

- Experimento 1

- Experimento 2

- Experimento 3

Experimento 1

El primer experimento ha sido realizado con un conjunto de datos extraído del boletín de ejercicios de la asignatura.

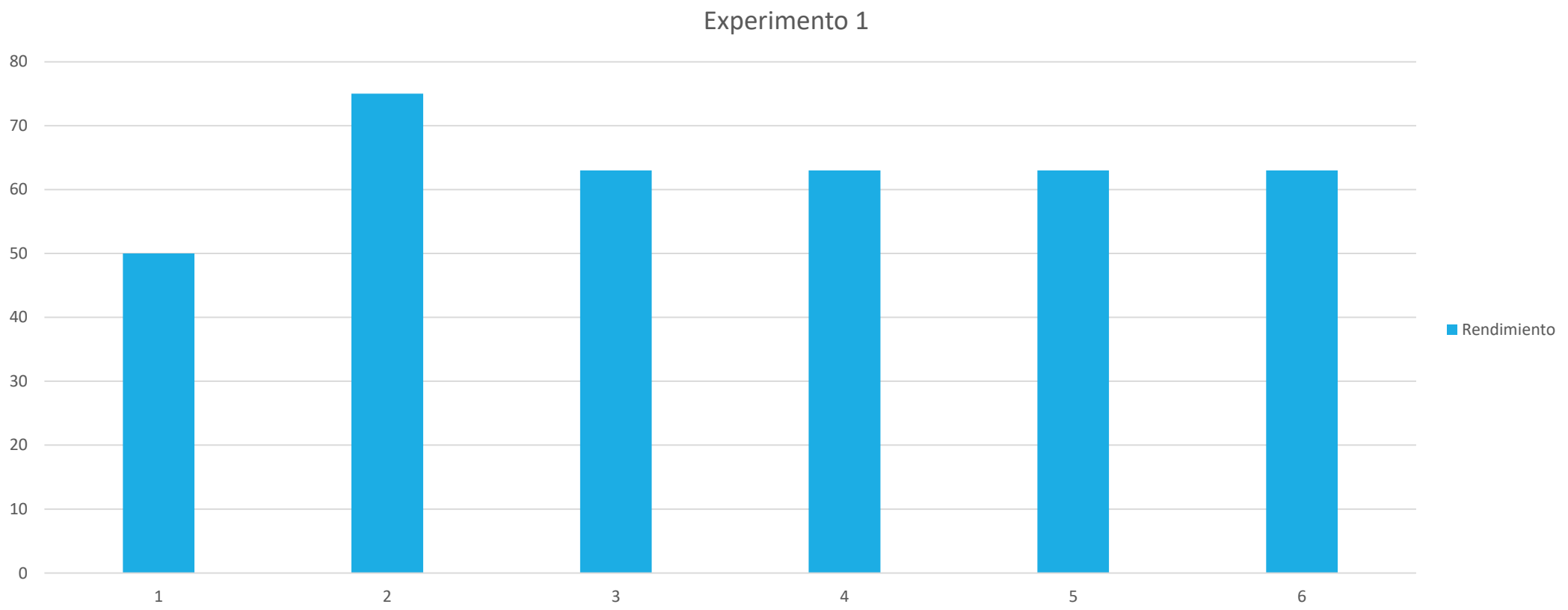
Nombre del conjunto de entrenamiento: ejemplo2.csv

Nombre del conjunto de evaluación: evaluación2.csv

Número de entradas en el conjunto de entrenamiento: 15

Número de entradas en el conjunto de evaluación: 8

Experimento 1



Experimento 2

El segundo experimento ha sido realizado con un conjunto de datos mayor obtenido en un repositorio de github.

Nombre del conjunto de entrenamiento: car-data-train.csv.

Nombre del conjunto de evaluación: car-data-test.csv.

Número de entradas en el conjunto de entrenamiento: 1296

Número de entradas en el conjunto de evaluación: 432

Experimento 2

Experimento 3

El último experimento ha sido realizado con un conjunto de datos aún mayor que el anterior obtenido de un repositorio de gihub.

Nombre del conjunto de entrenamiento: kr-vs-kp-train.csv.

Nombre del conjunto de evaluación: kr-vs-kp-test.csv

Número de entradas en el conjunto de entrenamiento: 2556

Número de entradas en el conjunto de evaluación: 640

Experimento 3

Conclusiones

En los experimentos realizados se observa que el mejor rendimiento se obtiene para valores del quórum de entre 1 y 7, es decir, valores pequeños. También se observa que no para todos los conjuntos de datos mejora el rendimiento, como por ejemplo el experimento 3.