



Nebrija
Universidad

PRÁCTICA DE INGENIERÍA DEL CONOCIMIENTO

CONSTRUCCIÓN DE UN SISTEMA RECOMENDADOR DE PELÍCULAS

Prof. Constantino Malagón Luque
Departamento de Ingeniería Informática

CONSTRUCCIÓN DE UN SISTEMA CLASIFICADOR DE PELICULAS

1. Descripción del problema

Tenemos una base de datos con 223 películas descritas por varios atributos, como la década en que fue estrenada, el género, actores, duración, etc... El objetivo de la práctica es construir un sistema basado en algoritmos de aprendizaje automático que extraiga las regularidades que presentan las diferentes películas según sea la puntuación que se les ha asignado, es decir, en qué se parecen aquellas películas que son buenas, malas etc.

2. Descripción de los atributos:

- Los atributos y su descripción de detallan en el fichero con extensión “names”:
- Como etiqueta de clase se utilizará el atributo puntuación, con sus diferentes valores.

3. Descripción de la tarea :

1. Construir una base de datos MySQL que almacene estos datos. Para ello debemos diseñar un modelo relacional para la base de datos de forma que este conjunto de datos y posibles futuros datos sean fácilmente almacenados. Incluir este modelo y su explicación en la memoria.

2. Importar los datos de la práctica a la base de datos MySQL (se recomienda para ello la herramienta phpmyAdmin, aunque puede hacerse como se quiera).

NOTA: Los puntos 1 y 2 obviamente no son necesarios para la realización de esta práctica, en que tenemos pocos datos. Se pide esta tarea porque es lo que normalmente, en casos más grandes y reales, suele hacerse.

3. Construir un programa de forma que, a partir del fichero de datos (exportado en formato csv) le agregue la cabecera necesaria según los atributos y el formato de archivo arff requerido por WEKA (<http://www.cs.waikato.ac.nz/ml/weka/>) Se puede construir uno desde cero usando vuestro lenguaje favorito o modificar este ejemplo creado en perl (http://www.hakank.org/data_mining/badge_problem.html) que usa el conjunto de datos badge (<ftp://ftp.ics.uci.edu/pub/machine-learning-databases/badges>)

4. Entrenar al sistema con estos archivos usando los siguientes algoritmos:

- a.Arboles de decisión (algoritmo C4.5, llamado en Weka J48)
- b.Algoritmo Naïve Bayes.
- c.Algoritmo de los K vecinos más próximos (K Nearest Neighbours)
- d.Algoritmos de combinación de clasificadores: Boosting y Bagging, usando como clasificador individual los árboles de decisión y también con Naïve Bayes.
- e.Perceptrón multicapa. En este caso, identificar los parámetros (pesos, factor de aprendizaje, etc.) En cuanto a su arquitectura, identificar el número óptimo de neuronas en la capa oculta. Para ello, construye una tabla que refleje pruebas de comparación en el porcentaje de aciertos con perceptrones que tengan de 2 a 6 unidades para obtener así la configuración óptima.

- **NOTA:** Utiliza la técnica de dividir el conjunto de entrenamiento en dos partes (60%-40%) de forma aleatoria para entrenar y testear respectivamente.

5.Explica en un párrafo de 10 líneas como máximo para cada uno en qué consiste cada algoritmo, sin entrar en detalles de cálculos. **NO SE PUEDE COPIAR NADA DE INTERNET. DEBEIS USAR VUESTRAS PROPIAS PALABRAS.**

6.Interpreta todos los datos relevantes del fichero de salida (matriz de confusión, porcentajes de error y clasificación,...) para cada uno de los algoritmos.

7.Haz una tabla comparativa del rendimiento obtenido por cada uno de los algoritmos usados: para ello ten en cuenta como parámetros de comparación los porcentajes de error y el tiempo de realización del experimento. Señala cuál es el algoritmo que creas que es el mejor de ellos.

8. ¿En qué se parecen las películas consideradas como buenas según estas valoraciones?

9. Elige diez películas que hayas visto recientemente (en cine o en DVD) y dime cómo las clasificaría el mejor algoritmo que hayas obtenido. Dime según este modelo cuál me recomendarías ver y cuál no.

10. Estudiar cómo influye la forma de clasificar los datos en la precisión. Para ello modifica la clase de 15 películas del conjunto de entrenamiento para que estén de acuerdo con vuestros gustos (también podéis añadir y/o quitar las instancias que os parezcan)

- Como documentación para el tema de las películas os podéis basar en vuestro propio conocimiento o echar mano del Internet Movie Database (<http://www.imdb.com>). Un sistema interesante que hace crítica de películas basándose en opiniones de los usuarios es el MovieLens (<http://movielens.umn.edu/login>)

9. ¿Cómo usarías esta práctica para construir un sistema web que recomiende películas y que sirva como red social, de forma que ponga en contacto a usuarios con los mismos gustos? Describe la arquitectura hardware/software del sistema haciendo uso de programas como el Visio. ¿Qué algoritmos de aprendizaje automático utilizarías para las dos tareas de recomendación, tanto de películas como de usuarios con intereses parecidos?

INSTRUCCIONES

1. La extensión de la memoria, en formato opendocument (odt) o pdf, no podrá superar las 25 hojas (siendo la extensión mínima la que se considere oportuna). En dicha memoria deberá ir:
 - Una portada
 - Un índice general
 - Un índice de tablas y figuras
 - Referencias de Internet y/o bibliográficas que hayáis usado para la resolución de la práctica.

AVISO: No se aceptarán memorias que sean entregadas en formato doc (MS Word)

2. La práctica podrá realizarse de forma individual o preferiblemente, en grupos de dos alumnos. En el caso de que se reciba una práctica hecha por tres alumnos se repartirá la nota entre los tres a partes iguales. La fecha límite de entrega será el 29 de Mayo a las 23 horas, y debe realizarse únicamente por e-mail a la dirección del profesor (cmalagon@nebrija.es). En el asunto del mensaje deberá ir la frase *Práctica de Ingeniería del Conocimiento*. La respuesta se dará en un fichero comprimido cuyo nombre debe ser el de los componentes del grupo separados por guión. En dicho fichero irán incluidas la memoria y los ficheros que se crean necesarios. Si alguna de estas restricciones no se cumpliesen serán penalizadas con 0.25 puntos a restar de la nota final de la práctica.

NOTA MUY IMPORTANTE: Cualquier parte de la práctica que se detecte que esté copiada directamente de Internet supondrá una nota automática de 0. En general, en un trabajo de investigación se pueden copiar literalmente pequeños extractos, frases o figuras, pero siempre debe ir entrecorillados o en cursiva, y se debe a su vez citar la fuente de donde se ha extraído.

CRITERIOS GENERALES DE CORRECCIÓN DE LA PRÁCTICA

1. Se valorará la capacidad de analizar los resultados de acuerdo al modelo propuesto, no sólo en cada una de las tareas individuales sino también en cuanto a los mejores o peores resultados comparados entre ellas. Una forma de comparación entre los resultados de las diferentes tareas podría hacerse mediante tablas o gráficos.
2. Por supuesto una parte importante de la práctica es la obtención de resultados adecuados.
3. Así mismo se valorarán las aportaciones personales hechas a modo de ampliación, mejoras o aspectos que no se hayan tenido en cuenta.
4. La adecuada presentación de los documentos se da por supuesta. Una mala presentación implica una bajada de nota de hasta un 40%. Dicha presentación no sólo se refiere a la estética de la memoria, sino también, y sobre todo a la aparición de faltas ortográficas o a la mala redacción del texto.