Práctica 8.
Aprendizaje Automático

Fecha de entrega: 13 de mayo de 2018

Esta práctica tiene como objetivo aplicar a distintos conjuntos de datos algunos de los algoritmos de aprendizaje automático disponibles en el entorno WEKA (www.cs.waikato.ac.nz/ml/weka/). Se elaborará una memoria en la que se incluirán las respuestas a las preguntas planteadas, los resultados obtenidos y la interpretación de dichos resultados.

Introducción: UCI Machine Learning Repository

En internet existen muchos repositorios de datos sobre los que probar algoritmos de aprendizaje automático. Quizás el más famoso de ellos sea el repositorio de la Universidad de California en Irvine: UCI Machine Learning Repository.

En esta práctica deberás elegir un conjunto de datos para cada apartado aunque podrías utilizar el mismo en varios apartados. Observa que existe una manera de filtrar los conjuntos de datos por la tarea a realizar (clustering, regression, classification,…) que te será especialmente útil para esta práctica:

http://archive.ics.uci.edu/ml/datasets.html

A modo de consejo, te recomendamos que no uses conjuntos de datos con un número muy elevado de atributos (más de 10), ni de instancias (más de 2.000).

Descripción del conjunto de datos

Por cada conjunto de datos que utilices deberás incluir una breve descripción del mismo.

- Nombre del conjunto de datos
- Breve descripción del problema que describe
- URL desde la que se descarga
- Tabla con el nombre y tipo de las variables
    - Si hay variables de salida (es un problema de regresión o clasificación) describir la variable de salida (distribución de frecuencias o histograma de la variable)

Apartado 1: Agrupamiento o clustering

Elige un conjunto de datos sobre el que realizar un agrupamiento o clustering y sigue el siguiente guión:

- Considera si debes normalizar o estandarizar las variables antes. Razona tu elección.
- Aplica un algoritmo de clustering de los que hemos visto en clase con una parametrización (el valor de *k* en el algoritmo de *k*-medias, o la forma en la que se agrupan clusters en el caso jerárquico).
- Determina el número de clusters que consideras adecuado para el conjunto de datos y justifica tu elección.

    - Para estar seguro, es posible que necesites repetir el proceso varias veces probando con otros parámetros u otras variables.
- Da un sentido a cada uno de los clusters que has obtenido en el contexto del problema que representa el conjunto de datos. Intenta contestar a estas preguntas (o a la mayoría de ellas):
    - ¿Qué valores toman las variables en cada cluster? Puedes usar la media, o añadir también la desviación típica.
    - ¿Qué cluster es más numeroso?
    - ¿Qué cluster es más homogéneo? ¿y menos?
- Documenta todo el proceso indicando los métodos utilizados y su parametrización, así como adjuntando capturas de pantalla de las soluciones y toda la información que consideres necesaria.

Apartado 2: Clasificación

Elige un conjunto de datos sobre el que resolver un problema de clasificación, concretamente utilizando un árbol de decisión. Para ello, elige un problema donde la variable de clasificación tenga al menos 3 clases. Sigue el siguiente guión:

- Considera si debes normalizar o estandarizar las variables antes. Razona tu elección.
- Ejecuta J48 (versión WEKA de C4.5). Utiliza para la validación el "training set".
- Vuelve a ejecutar J48 partiendo el conjunto de datos en entrenamiento y validación al 66%.
    - A la luz de los resultados es posible que desees cambiar los parámetros de generación del árbol. Haz pruebas hasta quedar satisfecho y documenta muy brevemente el proceso seguido.
- Incluye en la memoria los resultados obtenidos en ambas ejecuciones y las representaciones gráficas de los árboles correspondientes. En ambos casos crea la matriz de confusión de los datos de validación. Incluye en esa tabla "precisión" y "recall" para cada uno de los métodos. Comenta los resultados obtenidos.
- Incluye en la memoria y comenta los árboles de decisión obtenidos.
    - ¿Cuál de los dos árboles consideras mejor? ¿En qué ramas encuentras las principales diferencias?
    - ¿Cuál de las dos validaciones te parece más fiable? ¿Y si hubieras utilizado validación cruzada?
    - Interpreta el árbol que consideres más adecuado en el contexto del problema que estás abordando.
        - Interpreta someramente la pregunta que se realiza en el nodo raíz y los nodos hijos resultantes. Hazlo tanto en el contexto de un problema de clasificación (¿qué clases ha clasificado mejor?), como en el del problema representado en el conjunto de datos (¿qué sentido tiene esa pregunta y la clasificación que infiere dentro del problema?).
        - ¿Qué variables son más relevantes? ¿Se han elegido variables diferentes para ramificar nodos del mismo nivel? ¿A qué crees que se debe?
        - ¿Qué clases se confunden más entre sí en el árbol? ¿Esas clases se confunden habitualmente en todas las partes del árbol o solamente en

alguna rama? ¿Puedes darle algún sentido a esto dentro del problema de clasificación que estás abordando?
        - ¿Existe algún nodo-pregunta que tenga un gran poder discriminante y que merezca la pena destacar? ¿y algún nodo que consideres que no aporta mucho desde el punto de vista de la clasificación o del problema?
- Documenta todo el proceso indicando los métodos utilizados y su parametrización, así como adjuntando capturas de pantalla de las soluciones y toda la información que consideres necesaria.

Apartado 3: Regresión

Elige un conjunto de datos sobre el que resolver un problema de regresión, concretamente utilizando un perceptrón multicapa y el k-NN, y sigue el siguiente guión:

- Considera si debes normalizar o estandarizar las variables antes. Razona tu elección.
- Utiliza para las comparaciones la validación cruzada en p partes. Fija un valor de p que consideres adecuado.
- Elige la configuración automática del perceptrón multicapa y analiza el error cometido.
- A continuación, intenta ajustar un k-NN (clasificador “IBk” en el grupo “lazy”) de forma que obtenga un error parecido o inferior (compara tanto la raíz cuadrada del error cuadrático medio, como el error absoluto medio).
    - Seguramente te interese probar con diferentes subconjuntos de las variables de entrada y normalizarlas.
- Documenta todo el proceso indicando los métodos utilizados y su parametrización, así como adjuntando capturas de pantalla de las soluciones y toda la información que consideres necesaria.

Apartado 4: Opcional

En el problema de clasificación del apartado 2, prueba a ajustar un k-NN que obtenga iguales o mejores resultados que el árbol utilizado. Para que los resultados sean comparables debes utilizar la misma estrategia de validación. Compara los resultados de forma crítica. ¿Son similares la matriz de confusión y las medidas de error de clasificación? ¿A qué crees que se deben las diferencias?

Entrega

La entrega se realizará a través del campus virtual en un fichero pdf que contendrá la memoria elaborada. En la portada de la memoria debe aparecer el número de grupo y los nombres completos de sus integrantes. Además el nombre del archivo será P8GXX, siendo XX el número de grupo.