Indicaciones para los Informes de Prácticas

Análisis Automático de Datos para las Ciencias Biomédicas, Medioambientales, Agroalimentarias
Curso 2023-2024

A continuación se enumeran los casos y supuestos prácticos que deberá adjuntar en Moodle en la fecha que se indique por el profesorado. Cuando se disponga en dicha plataforma el enlace de envío para ello se darán instrucciones más detalladas.

El **formato** de entrega es PDF:

- El documento que presente debe ser formal y estar bien organizado: índice de contenidos, paginación, figuras y gráficas con leyenda y suficiente resolución, margenes correctos, encabezados y pie de páginas, texto en formato justificado, espaciado constante, etc.
- La portada debe contener nombre de la asignatura, titulación, universidad, curso académico, nombre del alumno/s y correo electrónico.

Contenido del informe. Ayúdese de las capturas de pantalla de Weka que considere oportunas para explicar mejor los casos y supuestos prácticos. No indique en el documento afirmaciones que no se pueden concluir de donde vienen o porque se ha llegado a ellas.

Alternativa a Weka. Si ya sabes programar y conoces entornos de ciencia de datos de R o Python puedes hacer este guión práctico con estos entornos y usar un notebook de Jupyter como forma de entrega de la práctica.

Primeros practicaremos la carga y preprocesado de un fichero ya procesado. Después debe elegir un caso práctico de los propuestos o utilizar datos propios de tu investigación. En este caso debes escribir al profesor.

1.1. Preprocesado de Audiology_soft

Cargue la base de datos **Audiology_soft** de Moodle **(carpeta "Datasets de ejemplo")**. *Audiology_soft* es una base de datos de la UCI que ya tiene algún preprocesamiento realizado. Trata sobre clasificación de enfermedades de oído (24 enfermedades) a partir de 10 atributos. Aplique los siguientes filtros, explicando cómo queda la base de datos en cada paso. Haga uso de capturas de pantallas y de lo que necesite para poder explicar el resultado:

1) Eliminar con el filtro ***Remove*** (o desde la pestaña ***Preprocess***, elige los que quieres borrar y pulsa en *Remove*) aquellos atributos que tengan demasiados valores perdidos como para hacer imputaciones a partir de medias o moda. Muestra cómo queda la base de datos al eliminarlos.

2) Usar el filtro ***filters → unsupervised → attributes → ReplaceMissingValues*** para reemplazar valores perdidos de atributos que sean factibles para imputar. Muestra cómo queda la base de datos al aplicar el filtro.

3) Usar el filtro para pasar de ***filters → unsupervised → attributes →* nominalToBinary,** haciendo que los nominales binarios (con dos valores) se pasen también a dos numéricos y no queden como un solo atributo numérico con valores 0 o 1. Muestra cómo queda la base de datos al aplicar el filtro.

1.2. Elección caso práctico

Escoja solo una base de datos, de **clasificación** o de **regresión,** de la siguiente lista.

Clasificación:

- ILPD (Indian Liver Patient Dataset): En ella hay valores perdidos en algunos patrones que debes procesar.
- Glass Identification. Base de datos del Servicio de Ciencias Forenses de EE.UU. donde se identifican 6 tipos de vidrio en función de su contenido en óxidos. En esta base de datos debes eliminar la variable del identificador de la muestra.
- Wine. Uso del análisis químico para determinar el origen de los vinos. En esta base de datos el atributo de clase está en la primera columna.

Regresión:

- Auto MPG: consumo urbano de automóviles (millas por galón). Para preprocesar esta base de datos al cargarla en una hoja de cálculo tendrás que activar el espacio y el tabulador como delimitador y "Combinar delimitadores". La columna "car name" debe eliminarse.
- Bike Sharing: este conjunto de datos contiene el recuento horario y diario de bicicletas de alquiler entre los años 2011 y 2012 en el sistema Capital bikeshare con la correspondiente información meteorológica y estacional. Elige la predicción por horas y/o días. Debes eliminar las variables casual y registered ya que la variable objetivo cnt es la suma de ambas. También debes eliminar dteday. Finalmente debes normalizar/estandarizar todas las variables, incluyendo la dependiente.

1.3. Importar y preprocesar datos

Si la base de datos está en formato CSV la puedes importar a weka con los pasos de Tema 4 - Conversión de bases de datos de la UCI a formato Weka o abrir directamente desde el Explorer. En otros casos los datos y los nombres de las variables están separados en ficheros .data y .names. Puedes abrir el .data como un CSV con LibreOffice Calc, Excel, etc. (ya que estará separado por espacios, tabuladores, comas, etc.) y añadir la fila de nombres de variables desde el .names a la primera fila.

Resumen de estos pasos:

1. Descargar en formato de texto o CSV.
2. Crear CSV a partir del fichero ejemplo.data ejemplo.names usando una hoja de cálculo.
3. Exportar a CSV. **Nota**: Al exportar a veces encontraremos problemas con la expresión de decimales según el separador de estos (. o , según idioma). Prácticamente todas las herramientas de análisis de datos asumen la codificación inglesa para expresar los decimales. A menudo lo más práctico es configurar nuestro entorno para trabajar con documentos en inglés o al menos la codificación de números en este idioma.
4. Desde el Experimenter de Weka abrir el CSV.
5. Seleccionar el atributo de clase y aplicar el filtro *Unsupervised* → *Attribute* → *NumericToNominal* seleccionando el índice de la clase (recuerde pinchar en Apply).
6. Pulsar en *Edit* para lanzad el editor de ficheros ARFF. Ahí se puede seleccionar el atributo de clase con el botón derecho (*attribute as class*). Pulsar OK.
7. A partir de aquí puedes exportar tu fichero a ARFF y trabajar en Weka.

1.4. Análisis preliminar

1. Desde el entorno *Preprocess*, cargue la base de datos en Weka e indiqué que tipo de problema está estudiando. Por cada variable, comente su significado o de qué se trata y cuál es su tipo (el uso de tablas agilizará su interpretación y lectura). Puede hacer también comentarios sobre las variables que desee y que quiera destacar. Además de la información que aporta Weka, puede ayudarse de la información disponible en los ficheros *.names* y *.txt*, que está recogida del propio repositorio de la UCI:

    1. Para todas las variables cuantitativas, recoge el valor medio, máximo, mínimo y la desviación típica de la misma, indicando de dónde obtiene esa información en Weka (con poner un ejemplo de de dónde se obtiene es suficiente). Dado el caso, puede comentar algo que le parezca interesante sobre esos valores.

    2. Para todas las variables cualitativas, recoge la frecuencia de cada categoría de la variable, indicando de dónde obtiene esa información en Weka (con poner un ejemplo de de dónde se obtiene es suficiente).

2. ¿Existen valores perdidos en la bases de datos? Si fuera así, analice en qué porcentaje y en qué variables (indique de dónde obtiene esa información). Sino fuera así, indique cómo podría analizarlo en Weka.

3. ¿Encuentra alguna relación entre pares de atributos, algunos valores de los mismos que permitan una cierta separación de clases, algún rango de valores donde se concentren datos?. Si localiza alguna relación significativa, explíquela (estudiando más a fondo el tipo de problema puede comprobar si puede tener o no sentido, por ejemplo, si es una base de datos de ILPD, puede mirar en la Web qué suele influir más en esa enfermedad). Si no ve relaciones indíquelo también. Además de determinados filtros de Weka, las opciones "*plot size*" y "*point size*" en la pestaña "*Visualize*" de Weka puede ayudarle en el proceso, al igual que la opción "Jitter" al pulsar sobre la gráfica de un par de atributos en concreto.

4. Si hay algún preprocesamiento que se pueda realizar a simple vista coméntelo. Es suficiente con que indique cómo realizar ese preprocesamiento o modificación de la base de datos en cuestión, pero si que debe comentar de dónde obtiene la información que afirme.

1.5. Análisis de correlación de atributos

Cargue la base de datos elegida y use el algoritmo ***CorrelationAttributeEval + Ranker*** (si es problema de regresión) o ***InfoGainAttributeEval + Ranker*** (si es problema de clasificación). Estos algoritmos indican un *ranking* de atributos con respecto a la variable de salida en función del coeficiente de correlación de *Pearson* o de la ganancia de información respectivamente. Use por defecto el conjunto de entrenamiento, que es la opción aplicada ("***Use full training set***"). ¿Qué atributos influyen más sobre la salida y cuáles menos? Se valorará si estudia el tipo de problema más a fondo y encuentra si tiene sentido o no la mayor o menor influencia de determinados atributos según muestra Weka,.

1.6. Normalización o estandarización

La mayoría de métodos de aprendizaje automático requieren o funcionan mejor con bases de datos normalizadas o escaladas.

Puede probar el efecto en el rendimiento con el filtro ***filters → unsupervised → attributes → Normalize*** y comentar como queda la base de datos.

1.7. Entrenamiento y prueba del modelo predictivo

Si es una base de datos de clasificación comente el modelo predictivo (y variables asociadas al mismo) que se obtiene al usar el clasificador ***functions → SimpleLogistic*** con un *10-fold crossvalidation*. Comenta los resultados, recoge los valores de CCR (*Correctly Classified Instances*), sensibilidad o CCR por clase (*TP Rate*), y área bajo la curva ROC (*ROC Area).*

Si es un problema de dos clases. Si es una base de datos de regresión comente el modelo predictivo (y variables asociadas al mismo) que se obtiene al usar el regresor ***functions → LinearRegression*** con un *10-fold crossvalidation*. Comenta los resultados, recoge el valor de R2 (*Correlation coefficient*), MAE (*Mean absolute error*) y RMSE (*Root mean squared error*).

Ambos métodos suelen usarse para obtener un rendimiento base en un conjunto de datos. Trata de mejorar el resultado comparando con dos métodos más de clasificación o regresión. Depende del problema que elijas métodos como SimpleLogistic suelen conseguir resultados muy competitivos y no ser fácil superarlo. Crea una tabla donde se recojan las métricas de rendimiento de los 3 métodos.

Nota de diseño experimental. En experimentación k-fold Weka creará una partición nueva de la base de datos en la evaluación de cada método. En una experimentación formal esto no serviría para hacer una comparación real de métodos ya que deberíamos tener los mismos 10 conjunto train-test para probar todos los métodos. Además, la normalización debería hacerse utilizando la media y desviación típica sólo del conjunto de train. Para estas prácticas obviaremos estos requisitos pero para un trabajo real debéis tenerlos en cuenta.