Tareas de Introducción a la Minería de Datos (c. 14/15)

JT Alcalá

Introducción

Se debe seleccionar un proyecto a realizar en la tarea correspondiente que he abierto en Moodle. Algunas son individuales y otras en parejas. Cualquier duda sobre la selección podeis preguntarla (preferentemente via Moodle, para que yo no la olvide). He dejado una opción sin límite de votos asociada a datos personales (de momento sólo hay una persona con itención manifiesta de trabajar sus datos, pero cualquier otra deberiamos valorar si son válidos para ser analizados con las técnicas vistas en el curso)

1. Estructura General

El informe debería estructurarse más o menos, en:

- Análisis preliminar.
  Aquí se debe discutir la naturaleza/tipología de las variables, comentar la varible de clasificación, analizar la coherencia del resto de variables (ausencia de datos disparatados o aberrantes en la base de datos), valorar el nivel de datos ausentes (si los hay) y cómo se puede minorar su efecto. Es interesante presentar un análisis descriptivo asociado a las variables, e incluso separado/estratificado por la variable de clasificación que nos permita una primera valoración de la importancia de cada variable en relación con la varible de clasificación.

  También puede ser interesante, en el caso de muchas variables (contínuas), visualizar el grado de correlación entre ellas y algún tipo de Análisis de Componentes Principales que permita prever la estructura de asociación entre las variables antes de ajustar o aplicar otras técnicas.

Al final de esta etapa se debe terminar con una división del conjunto original en un conjunto de entranimiento/aprendizaje y otro de validación. La proporción o la selección estratificada de los conjuntos puede depender de cada problema concreto. Si el conjunto de datos no lo aconseja, justificar el uso de otras técnicas como validación cruzada o bootstrap para evitar el sobreajuste y el optimismo en los resultados de clasificación. Todo el proceso deberá ser comentado.

- Técnicas de clasificación

Como norma general, se deberán **ajustar, valorar** y **comparar** al menos dos clasificadores en los trabajos individuales. En los trabajos con dos personas, al menos tres clasificadores o, si los datos no lo permiten, dos clasificadores con bastante variación en opciones y alternativas de esos dos clasificadores.

Al menos de los dos clasificadores, uno debe ser árboles de clasificación (con algunas de sus variantes: bagging, boosting,...) o Redes Bayesianas (donde se deben ensayar con diferentes algoritmos de búsqueda: Hill Climber, K2 y con diferentes número de padres). Obviamente, las redes bayesianas deberían ajustarse mediante Weka. También recordaros que con Weka y estas técnica se necesita que todas las varables sean de tipo cualitativo nominal (en caso de no serlo, deberían discretizarse en la fase de pre-procesado, antes de su uso). En el caso de usar tres clasificadores, uno de ellos deben ser Árboles y el otro del grupo: Redes Bayesianas o SVM. El tercer clasificador puede ser Regresión logística/multinomial ó bien A. discriminante (en alguna de sus variantes, incluido A. Discri. Penalizado/Regularizado) o bien k-vecinos próximos.

En todos los clasificadores, hay que comentar las estrategias utilizadas para selección de parámetros (de complejidad) y la prevención del sobreajuste.

En la parte de comparar los clasificadores, deberá hacerse uso no sólo de tablas de confusión, sino que también deberían aparecer otros elementos que permitan la comparación (curvas ROC, otros...) Por ejemplo, sería interesante valorar la concordancia en la predicción sobre el conjunto de entrenamiento que aportan los diversos clasificadores.

En los casos donde las variables o la temática lo permita, debería hacerse un esfuerzo de interpretación de los modelos finales que se han ajustado.

- Reproducibilidad. Como en algunos casos se usan métodos de aleatorización en diversas fases del estudio, se recomienda el uso de órdenes que fijen semillas para poder obtener un cierto grado de reproducibilidad de los resultados. También se recomienda que se entregue un script con las órdenes utilizadas o al menos que aparezcan las principales en el documento escrito.

2. Conjuntos de Datos

Se proponen un total de 17 proyectos (algunos individuales, otros para realizarse en pareja, algunos pueden ser abordados de forma individual o en pareja). Cuando el proyecto es marcado por el número máximo de personas que se indican, ya no puede ser seleccionado por otros alumnos.

Los proyectos son:

1.) vinos blancos portugueses (2 personas)

2.) vinos rosados portugueses (1 persona)

   `http://archive.ics.uci.edu/ml/datasets/Wine+Quality`

   Son dos conjuntos de datos (vinos blancos y rosados de la variedad "Vihno verde"). La variable de clasificación es quality (puntuación de 0 a 10). Debería recodificarse en un número menor de categorías para facilitar el análisis como un problema de clasificación.

3.) fragmentos de cristal (1 persona)

   `http://archive.ics.uci.edu/ml/datasets/Glass+Identification`

   Un clásico de problemas de clasificación. La variable de clasificación presenta un número elevado de categorías (tipos de cristal) y en un primer análisis puede ser conveniente recodificar la variable en una nueva con un menor número de categorías por agrupamiento de algunas categorías minoritarias.

4.) selección de marcas de zumos (1/2 personas)

   `http://azzalini.stat.unipd.it/Book-DM/data.html`(juice.data)

   Los datos hacen referencia a un número elevado de compras entre dos marcas de zumo competidoras. Se desea construir un sistema automático que determine en función de la 'lealtad' de cliente y las ofertas del momento la elección final de una marca u otra.

5.) satisfacción clientes banco brasileño (1/2 personas)

   `http://azzalini.stat.unipd.it/Book-DM/data.html`(brazil.csv)

   Los datos corresponden a campañas de marketing de un banco brasileño entre sus clientes y el resultado final es si el cliente está satisfecho o no con el banco (variable binaria `ok`). Además de las características más personales del cliente, también se dispone de información sobre una lista seleccionada de productos que posee con el banco y con la competencia (otros bancos).

6.) datos fidelidad telefónica (1/2 personas)

   `http://azzalini.stat.unipd.it/Book-DM/data.html`(telekon-www.zip)

   La muesrtra recoge durante 10 meses consumos/tráfico de los clientes de la compañia telefónica (activos). Al decimotercer mes, el cliente recibe el status de activo o desactivado (dos meses después del último dato de tráfico).

7.) enfermedad coronaria (2 personas)

   Conjunto de datos formado por 4 bases de datos unificadas (correspondientes a datos de hospitales de Clevelan, Hungría, Zurich y Bassel).

   `http://archive.ics.uci.edu/ml/datasets/Heart+Disease`

   El objetivo es poder predecir en base a las 14 variables fundamentales (procesadas) la presencia o no (variable recodificada) de enfermedad coronaria. En este trabajo, se deberá procurar usar los 4 ficheros.

8.) evaluación de riesgo crediticio (2 personas)

   `http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)`

   Los datos corresponden con una entidad de crédito alemana que desea construir un sistema automático para medir el riesgo de impago de un futuro cliente que demande un crédito. En la carpeta hay al menos dos conjuntos equivalentes, pero es algo más cómodo trabajar con el conjunto numérico donde han convertido las etiquetas en códigos numéricos en las variables nominales. Un aspecto especialmente interesante en este proyecto es utilizar la matriz de costes por mala clasificación a la hora de construir y comparar diferentes técnicas de aprendizaje.

9.) identificación de setas peligrosas (1 persona)

   `https://archive.ics.uci.edu/ml/datasets/Mushroom`

   Conjunto de datos con 22 atributos (nominales) y una variable de clasficación (comestible/venenosa). El objetivo es construir reglas que permitan diferenciar a las setas venenosas de las comestibles. No resulta sencillo encontrar un conjunto de reglas que sea eficaz con el objetivo.

10.) enfermedad torácica (1 persona)

    `http://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data`

    Los datos hacen referencia a la expectativa de supervivencia a un año de pacientes operados de cancer de laringe.

11.) pacientes hepáticos (1 persona)

    `https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)`

    Datos referentes a pacientes con enfermedad hepática y sin ella en una región concreta de la India. El objetivo es caracterizar y clasificar a futuros pacientes como enfermos hepáticos o sanos.

12.) daños de láminas de acero (1 persona)

    `https://archive.ics.uci.edu/ml/datasets/Steel+Plates+Faults`

    Los datos corresponden con diferentes tipos de deterioros sobre planchas de acero y una serie de 27 características asocidas. El objetivo es construir modelos que permitan anticipar el tipo de fallo. La varible respuesta es nominal con 7 categorías diferentes.

13.) disfonía en enfermos de Parkinson (1/2 personas)

    `https://archive.ics.uci.edu/ml/datasets/Parkinsons`

    Los datos son adecuados para ajustar modelos de clasificación que permitan diferenciar a enfermos de Parkinson de pacientes sanos, en base a registros de voz. Aunque hay un total de 197 registros de voz, corresponden a 31 pacientes, de los cuales hay 23 enfermos. Las variables son fundamentalmente de tipo continuo.

14.) reconocimiento óptico de números manuscritos (1/2 personas)

    `https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits`

    Los datos corresponden a intensidades registradas sobre una retícula por un dispositivo óptico correspondientes a números manuscritos. Hay un conjunto de entrenamiento formado por 30 personas y uno de validación formado por 13 personas diferentes. El problema es disponer de modelos que predigan correctamente el dígito. Observar que la variable respuesta es nominal con 10 categorías.

15.) reconocimiento de spam (2 personas)

    `https://archive.ics.uci.edu/ml/datasets/Spambase`

    El objetivo final sería construir un filtro anti-spam. Los datos ya están preprocesados y se han calculado básicamente la frecuencia con la que aparecen ciertas palabras o carácteres en el texto del correo o secuencias de letras mayúsculas.

16.) Marketing bancario (2 personas)

    `https://archive.ics.uci.edu/ml/datasets/Bank+Marketing`

    Los datos son un conjunto de 20 características: personales, propias de la campaña y socieconómicas. La mayoría son de tipo cualitativo, ordinal y las menos de tipo numérico continuo. Tal vez sea necesario proporcesar los datos, descartando alguna variable. El objetivo último es anticipar si un cliente va a comprar el producto bancario o no.

17.) Recurrencia en Cáncer de Pecho (3 personas)

    `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Prognostic)`

    Los datos recogen el seguimiento de pacientes con cáncer de pecho. El problema al que nos vamos a dedicar es aprender a clasificar si un tumor es recurrente o no a un cierto umbral de tiempo (p.ej. 24 meses). Esta variable no está en el conjunto tal cual, hay que constuirla a partir de la información que hay en los primeros campos del fichero de datos y que hacen referencia a si la paciente ha desarrollado recurrencia y en qué momento desde la operación. Las características numéricas se calculan para los núcleos celulares y se dan tres valores para cada característica (p.ej. radio): media, SE y media de los tres mayores radio observados. El conjunto no es muy numeroso en cuanto a casos, deberían valorarse si conviene retener un conjunto de validación o es mejor hacer CV o boostrap a la hora de valorara la eficiencia de los clasificadores.

18.) datos propios

    Cualquier conjunto que tenga sentido plantear un problema de clasificación (binaria o múltiple). Al menos debe haber en torno a 10-15 variables que tengan una mayor o menor relación con la variable de clasificación. El número de casos o instancias debería ser al menos en una relación 10:1 por cada variable, es decir no menos de 100 o 150 casos, además las categorías de la variable de clasificación deberían no estar extraordinariamente descompensadas. La existencia de un alto porcentaje de datos ausentes puede invalidar o dificultar notablemente la aplicación de las técnicas si no se corrige de antemano dicha situación.