UNIVERSIDAD DE GRANADA

UNIVERSIDAD DE GRANADA

Departamento de Ciencias de la Computación e Inteligencia Artificial

Inteligencia de Negocio

Guión de Prácticas

Práctica 1:
Resolución de problemas de clasificación y análisis experimental.

Curso 2020-2021

Cuarto Curso del Grado en Ingeniería Informática

Práctica 1

Resolución de problemas de clasificación y análisis experimental

1. Objetivos y Evaluación

En esta primera práctica de la asignatura Inteligencia de Negocio veremos el uso de algoritmos de aprendizaje supervisado de clasificación como herramienta para realizar análisis predictivo en la empresa. En ella el alumno adquirirá capacidades para abordar problemas reales donde la minería de datos puede aportar valor en forma de conocimiento para ayudar en la toma de decisiones para gestión empresarial. Concretamente, se trabajará con un conjunto de datos real sobre el que se emplearán diferentes algoritmos de clasificación (para su comparación) y a la luz del conocimiento descubierto se podrán concluir estrategias para resolver el problema. Para ello, se deberán crear informes de resultados y análisis lo suficientemente profundos para resultar de utilidad.

La práctica se calificará hasta un **máximo de 2 puntos**. Se valorará el acierto en los recursos de análisis gráficos empleados, la complejidad de los experimentos realizados, la interpretación del conocimiento extraído, la organización y redacción del informe, etc.

2. Descripción del problema: Identificación del tipo de tumor en mamografías

En un estudio sobre cáncer de mama (*Mammographic Mass dataset*[1]) se poseen datos de distintos pacientes. Se desea predecir el tipo de tumoe (benigno o maligno) en tu tratamiento, a partir del resto de datos. Hay datos de 961 pacientes, y los atributos característicos que se describen a continuación:

1. Código **BI-RADS**: sistema de control de calidad, su uso diario implica una evaluación en categorías numéricas de una mamografía, asignado por el médico radiólogo e image-

[1] http://bml.io/RILp7d

nologo radiólogo después de interpretar la mamografía. Puede ver más información en https://es.wikipedia.org/wiki/BI-RADS.

2. Edad del paciente, valor numérico entero.
3. Forma de la masa anormal detectada, identificado mediante una letra:

   R redondeada.

   O ovalada.

   L Lobular.

   I Irregular.

   N No definida.

4. Margen de masa: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5 (nominal)
5. Densidad de la masa tumoral, valor entero ordinal:

   1. Alta.
   2. Media.
   3. Baja.
   4. Contenido graso (no tumoral).

6. Severidad (objetivo a predecir), puede tomar los valores:

   benigno El tumor es de tipo benigno.

   maligno El tumor es de tipo maligno (cáncer).

Los datos se encuentran en el fichero **mamografias.csv**, disponible en la web de la asignatura y en PRADO.

3. Tareas a Realizar

La práctica consiste principalmente en que el alumno estudie el comportamiento de distintos algoritmos de clasificación mediante el diseño experimental apropiado y el análisis comparado de resultados. Además, también deberá extraer conclusiones a partir del conocimiento aprendido mediante estos algoritmos para comprender las relaciones entre las variables (también llamadas *características* o *predictores*) que favorecen una determinada clase. El trabajo se realizará sobre el software Scikit-Learn.

Concretamente, se deberán resolver adecuadamente las siguientes tareas:

1. Se considerarán al menos cinco algoritmos de clasificación distintos. Se valorará la selección justificada de estos algoritmos en función de las características del conjunto de datos así como la elección de variedad de tipos de representación (árboles, reglas, redes neuronales, Naïve-Bayes, k-NN, etc.). En los casos en los que el algoritmo dependa fuertemente de algún parámetro, se podrá realizar un estudio de dicho parámetro en la medida de los resultados.

2. Toda la experimentación se realizará con validación cruzada de 5 particiones. Para sustentar el análisis comparativo se emplearán tablas de errores (precisión), matrices de confusión, y alguna otra medida como el AUC. Además de la precisión, pueden considerarse otras medidas de rendimiento y de complejidad del modelo (número de hojas, número de reglas, número de nodos, etc.).

3. Todos los análisis de resultados serán comparativos, de forma que se estudien los pros y contras de cada representación y/o de cada algoritmo. La documentación deberá incluir al menos una tabla resumen que incluya los resultados medios de todos los algoritmos analizados. El análisis no podrá reducirse a una simple lectura de los resultados obtenidos. El alumno deberá formular y argumentar hipótesis sobre las razones de cada resultado. En este problema, ¿por qué el algoritmo X funciona mejor que el Y? ¿Por qué la representación X presenta ciertas ventajas respecto a la Y?

4. Se probarán configuraciones alternativas de los parámetros de los algoritmos empleados justificando los resultados obtenidos. Por ejemplo, ¿puedo evitar o paliar el sobreaprendizaje ajustando los parámetros? ¿Puedo obtener modelos más fácilmente interpretables sin sacrificar excesiva precisión? Para realizar este análisis, se incluirán tablas comparativas con los resultados del algoritmo con parámetros o configuración por defecto y con las distintas variaciones estudiadas. Si el análisis es suficientemente completo, no es necesario estudiar todos los algoritmos analizados, se pueden escoger solo algunos de ellos.

5. A la luz de este análisis, se deberá estudiar un procesado básico de los datos que mejore la predicción (por ejemplo, eliminar alguna característica por razón justificada, agrupar los valores posibles de una característica, eliminar ciertas instancias del conjunto de entrenamiento que se consideren erróneas, convertir una característica categórica en varias binarias, imputar valores perdidos, equilibrar el balanceo de clases...). Deberán justificarse las acciones tomadas y analizar por qué determinado procesado funciona mejor en un determinado tipo de algoritmo. Si no se consigue mejorar la predicción, se podrá al menos describir los procesados que se han probado y los resultados obtenidos. De nuevo, se requiere una tabla resumen que muestre los resultados antes y después de los diferentes procesados de datos.

6. Basado en todo lo anterior, se deberán extraer conclusiones sobre los factores que determinan cada clase. Para llegar a estas conclusiones, se pueden analizar los modelos legibles generados (por ejemplo, árboles de decisión, conjuntos de reglas o regresiones lineales)

así como visualizar los resultados de predicción de los modelos sobre diferentes casos de entrada (*What-If Analysis*).

4. Esquema de la Documentación

La documentación entregada deberá ajustarse al siguiente esquema (debe respetarse la numeración y nombre de las secciones):

1. **Introducción**: se hablará sobre el problema abordado y todas las consideraciones generales que se deseen indicar.

2. **Procesado de datos**: se describirá el procesado analizado, y los distintos tipos de procesamiento (ej: distintas formas de abordar los valores nulos) que se vayan a estudiar. El procesamiento puede ser común a todos los algoritmos, o específico para determinados algoritmos, tal y como se describe en el punto 5.

3. **Configuración de algoritmos**: se incluirá un apartado para cada algoritmo cuya configuración y parámetros hayan sido estudiados. Si se estudian varios, se incluirá una tabla con los resultados y se realizará el correspondiente análisis como se describe en el punto 4.

4. **Resultados obtenidos**: incluirá un apartado x por cada algoritmo estudiado (si se probaron varios parámetros, se pondrá únicamente los valores de la mejor combinación). En cada apartado se mostrará el código de la creación del modelo, sus parámetros finales, y una tabla con los resultados obtenidos por el algoritmo por cada tipo de procesamiento como se describe en el punto 2 del apartado anterior.

5. **Análisis de resultados**: incluirá la tabla resumen de todos los algoritmos analizados así como su interpretación y análisis mencionados en el punto 3 del apartado anterior. Se hará un apartado por cada tipo de procesamiento global. También se podrá hacer un apartado adicional indicando cómo mejorar los más prometedoras, o sugerencias para mejorar los resultados. Se podrán añadir gráficas y visualizaciones que apoyen el análisis.

6. **Interpretación de los datos**: como se describe en el punto 6. Se identificarán los atributos que identifican mejor cada clase. Además, se analizarán los modelos interpretables (ej: visualizándolos) para sustentar la interpretación de resultados.

7. **Contenido adicional**: cualquier tarea adicional a las descritas en este guión puede presentarse en esta sección.

8. **Bibliografía**: referencias y material consultado para la realización de la práctica.

No se aceptarán otras secciones distintas de estas. Además, la primera página de la documentación incluirá una portada con el nombre completo del alumno, grupo de prácticas y dirección email. También se incluirá una segunda página con el índice del documento donde las diferentes secciones y páginas estarán enlazadas en el `pdf`.

5. Entrega

La fecha límite de entrega será el martes **3 de noviembre** de 2020 hasta las **23:59**. La entrega se realizará a través de la web de la asignatura en PRADO. En un único fichero `zip` se incluirá el árbol de directorios completo que contiene el código python, la documentación de la práctica realizada en `pdf` y cualquier otro archivo que el alumno considere relevante o sea necesaria para ejecutar el código. El nombre del archivo `zip` será el siguiente (sin espacios): `P1-apellido1-apellido2-nombre.zip`. La documentación tendrá el mismo nombre pero con extensión `pdf`. Es decir, la alumna "María Teresa del Castillo Gómez" subirá el archivo `P1-delCastillo-Gómez-MaríaTeresa.zip` que contendrá, entre otros, el archivo `P1-delCastillo-Gómez-MaríaTeresa.pdf`. En ningún caso se aceptan ni envíos por email o mensajes en PRADO.

Todas las referencias a archivos de entrada o salida deberán referirse a direcciones dentro de esa carpeta, de forma que el proyecto sea autocontenido.