UNIVERSIDAD DE GRANADA
E.T.S. de Ingenierías Informática y de Telecomunicación

Departamento de Ciencias de la Computación e Inteligencia Artificial

Inteligencia de Negocio

Guión de Prácticas

Práctica 3: Competición en DrivenData

Curso 2019-2020

Cuarto Curso del Grado en Ingeniería Informática

Práctica 3

Competición en DrivenData

1. Objetivos y Evaluación

En esta tercera práctica de la asignatura Inteligencia de Negocio veremos el uso de métodos avanzados para aprendizaje supervisado en clasificación sobre una competición real disponible en DrivenData (`https://www.drivendata.org/`). El estudiante adquirirá destrezas para mejorar la capacidad predictiva del modelo mientras se familiariza con una de las plataformas de competición en ciencias de datos que está ganando gran interés por dirigir la resolución de problemas al bien social.

La evaluación de la práctica se dará en función de la posición final (relativa al conjunto de estudiantes participantes) que ocupe el resultado propuesto por el estudiante con el siguiente criterio:

puesto	1°	2°	3°	...	10°	...	último
puntuación	2	1,93	1,85	...	1,33	...	0,5

Las posiciones serán linealmente proporcionales entre los 2 puntos del primero y los 1,33 puntos del décimo, y a su vez linealmente entre estos 1,33 puntos del décimo y los 0,5 puntos del último. Para ser evaluado no bastará con subir los resultados a DrivenData, se deberá también adjuntar un documento que describa el proceso seguido por el estudiante para resolver la práctica y demostrar mediante la actividad registrada en DrivenData que ha habido un esfuerzo por mejorar los resultados. En otro caso, el alumno no obtendrá ninguna puntuación en esta práctica.

Sobre la puntuación obtenida en base a la posición, se aplicará un factor corrector $[0{,}5, 1{,}5]$ (es decir, se podrá reducir o aumentar hasta un 50 %) en función de la calidad de la documentación presentada y las soluciones abordadas.

2. Descripción del Problema y Tareas

La competición será la *Richter's Predictor: Modeling Earthquake Damage* disponible en `https://www.drivendata.org/competitions/57/nepal-earthquake/`. Basado en aspectos

de la ubicación y construcción de edificios, el objetivo es predecir el nivel de daño a los edificios causado por el terremoto de Gorkha en 2015 en Nepal. Esta es una competencia de práctica de nivel intermedio. Los datos se recopilaron mediante encuestas realizadas por la Oficina Central de Estadística, que depende de la Secretaría de la Comisión Nacional de Planificación de Nepal. Esta encuesta es uno de los conjuntos de datos posteriores al desastre más grandes jamás recopilados, que contiene información valiosa sobre los impactos de los terremotos, las condiciones de los hogares y las estadísticas socioeconómico-demográficas.

El conjunto de entrenamiento consta de 260.601 instancias y 39 atributos (de los cuales, `building_id` toma valores únicos y solo sirve para identificar cada ejemplo) categóricos, enteros y binarios. Se trata de predecir la variable ordinal `damage_grade`, que representa un nivel de daño al edificio que fue golpeado por el terremoto. Hay 3 grados de daño: 1, representa un daño bajo; 2, representa una cantidad media de daño; y 3, representa la destrucción casi completa.

Al ser la variable a predecir ordinal, el problema puede verse como uno de clasificación o de regresión ordinal. La regresión ordinal a veces se describe como un problema intermedio entre la clasificación y la regresión. Para medir el rendimiento de nuestros algoritmos, utilizaremos la puntuación de F1 que equilibra la precisión y el *recall* de un clasificador. Tradicionalmente, la puntuación F1 se utiliza para evaluar el rendimiento de un clasificador binario, pero como tenemos tres posibles etiquetas, utilizaremos una variante llamada puntuación F1 micropromediada. En Python se puede calcular fácilmente usando `sklearn.metrics.f1_score` con el argumento `average='micro'`.

En esta competición se permite el uso de cualquier *software*, algoritmo o lenguaje que el alumno considere útil. Está terminantemente prohibido usar la clase en los datos de *test*, en caso de conocerse, para entrenar, configurar o mejorar el modelo predictivo. Cualquier indicio de esta conducta supondrá la anulación de la práctica.

La siguiente web sirve de ayuda para resolver esta competición: `https://www.kaggle.com/mullerismail/richters-predictor-modeling-earthquake-damage/kernels`

3. Documentación

La documentación explicará las estrategias seguidas y el progreso que se ha ido desarrollando durante la competición. Deberán razonarse brevemente los diferentes pasos tomados apoyándose en visualización de datos u otras técnicas de análisis para comprender las características del problema. Se recomienda añadir también extractos de los *scripts* para explicar el trabajo realizado. Será obligatorio incluir una tabla que contenga tantas filas como soluciones se han subido a DrivenData incluyendo columnas que resuman cada experimento conteniendo, al menos:

- la fecha y hora de subida a DrivenData,
- la posición que ocupó en ese momento,
- el *score* sobre el conjunto de datos de entrenamiento,

- el *score* obtenido en DrivenData al subir la predicción en *test*,
- breve descripción del preprocesado realizado,
- breve descripción de el/los algoritmo(s) de clasificación/regresión empleado(s) y
- configuración de parámetros de esos algoritmos.

La ausencia de esta tabla o una descripción incompleta de la misma supondrá la anulación de la práctica.

Adicionalmente, la segunda página de la documentación (después de la portada y antes del índice) contendrá una **captura de pantalla de la tabla** *Submissions* (que contiene las columnas *Score*, *Submitted by* y *Timestamp*) disponible en la web de DrivenData para cada usuario.

De cada subida realizada a DrivenData se conservará el fichero `.csv` y el *script* en Python o similar usado para ese experimento. Se nombrarán de forma clara y enumerada para poder identificar con facilidad a qué experimento de la tabla corresponde. Este material se entregará junto a la documentación.

El alumno deberá definir como usuario en DrivenData su nombre de pila y primer apellido terminando con `_UGR`. Por ejemplo: `Jorge_Casillas_UGR`.

4. Entrega

La competición en DrivenData finaliza el martes **31 de diciembre de 2019** a las 19:13. Tras acabar la competición, el estudiante deberá también entregar antes del martes **7 de enero de 2020** a las 23:59 una documentación que explique las tareas realizadas y todas las soluciones `.csv` subidas a DrivenData junto con los *scripts* utilizados.

Este material se entregará a través de la web de la asignatura en `https://decsai.ugr.es` en un único archivo `zip`. Por ejemplo, la estudiante "María Teresa del Castillo Gómez" subirá el archivo `P3-delCastillo-Gómez-MaríaTeresa.zip`. La documentación, contenida en ese mismo archivo `zip`, tendrá el mismo nombre pero con extensión `pdf`.