ugr | Universidad de Granada

Competición Kaggle

Minería de Datos, Aspectos Avanzados

Dpto. Ciencias de la Computación e Inteligencia Artificial
E.T.S. de Ingenierías Informática y de Telecomunicación
Universidad de Granada

DECSAI

Índice

1. Objetivos

En esta práctica se pondrán en juego los conocimientos y métodos de aprendizaje vistos durante el desarrollo de la asignatura **Minería de datos: Aspectos Avanzados**. Para ello se hará uso de la plataforma *Kaggle*, que permite establecer una competición de clasificación avanzada entre todos los alumnos.

El estudiante debe adquirir destrezas para mejorar los datos con el objetivo de obtener los mejores resultados posibles explorando diferentes algoritmos de aprendizaje (y sus posibles parametrizaciones) y familiarizarse con una de las plataformas más usuales en el ámbito de **Ciencia de datos**. En esta plataforma existen muchos conjuntos de datos disponibles puestos a disposición del público general, con el objetivo de conseguir resultados de su análisis. Algunas de las organizaciones que intervienen ofrecen premios a los participantes que obtengan mejores resultados.

Para llevar a cabo este trabajo, se ha dispuesto de dos competiciones en el marco de Kaggle (http://www.kaggle.com) y Kaggle InClass (http://inclass.kaggle.com). La primera de ellas está versada sobre la clasificación no balanceada, mientras que la segunda será la clasificación de dígitos basada en el conjunto de datos MNIST. **IMPORTANTE**: *Solo deberá realizar una de las dos competiciones programadas.*

La competición se realizará a nivel *individual* por lo que cada estudiante deberá crearse una cuenta propia en Kaggle para poder participar. La cuenta deberá estar asociada a la dirección de correo institucional (@ugr.es o @correo.ugr.es). No se permitirán entregas a través de una dirección diferente.

En el resto de este documento, se especifican con más detalle los dos problemas propuestos (de los que se resolverá únicamente uno de ellos), el material a entregar (la documentación asociada) y el sistema de evaluación propuesto. En particular, la práctica se puntuará en función de la posición que ocupen los resultados ofrecidos por los modelos de cada participante y de la documentación aportada. Esta documentación debe detallar de forma clara el proceso de trabajo seguido desde el inicio hasta la entrega de los resultados finales.

El *plazo límite de la competición* será el día **7/04/2017 a las 23.59**. No se permitirá el envío de ningún resultado adicional después de dicha fecha.

La *fecha de entrega de la documentación* en Prado será el día **21/04/2017 a las 23.55**. IMPORTANTE: No se aceptará ningún envío a través del correo electrónico ni ninguna otra vía.

2. Clasificación no Balanceada

El primero de los dos posibles problemas a resolver está enmarcado en el contexto de la clasificación con conjuntos de datos no balanceados. Para ello, deberá utilizar diversos métodos de clasificación (árboles de decisión, métodos ensamble, reglas, máquinas de soporte vectorial, etc.) con las variantes para abordar la clasificación con distribución de clases no equilibradas descritas en la asignatura, pero aplicados ahora a un problema real. Las herramientas, metodología, número y tipo de pruebas a realizar, etc. serán decisión unilateral de cada estudiante.

El acceso a la plataforma de competición se hace mediante el enlace:

https://inclass.kaggle.com/c/imbclass

Para poder participar cada estudiante debe registrarse en Kaggle, haciendo uso del correo institucional de la Universidad de Granada. Si alguien tiene problemas para entrar que se ponga en contacto con los profesores de la asignatura para solventarlo lo antes posible.

Se trabajará con un problema de carácter matemático en formato CSV (archivo separado por comas) con un total de 6.400 instancias, que se han dividido al 50% entre entrenamiento y test. Este conjunto de datos está representado por un total de 22 atributos con valores numéricos (enteros y reales) y dos clases, con un ratio de desbalanceo (IR) de aproximadamente 2. El conjunto de entrenamiento contiene una columna adicional con las etiquetas de clase {0, 1}. Es sobre este fichero sobre el que tendréis que aplicar las técnicas de que consideréis más oportunas con el objetivo de obtener los mejores resultados posibles en la clasificación de las instancias del conjunto de test.

Los ejemplos de test están además numerados con un identificador que deberá incluirse en los ficheros de salida que se sometan al sistema. Para facilitar el formato de entrega en *Kaggle*, se ha incluido también un fichero "sample.csv" con un ejemplo de entrada de datos de test. Todos estos ficheros están disponibles en la Web de la competición de *Kaggle* anteriormente indicada.

El objetivo será alcanzar la máxima precisión en términos de la medida AUC para el conjunto de test. Solamente se podrán enviar hasta 3 ficheros de prueba por día para comprobar los resultados obtenidos sobre el conjunto de test etiquetado (interno al sistema).

Los resultados que se pueden observar en el ranking público son calculados con sólo la mitad del conjunto de test, mientras que la otra mitad del conjunto se utiliza para realizar un ranking privado, que será el que se utilice en la evaluación final (ver Sección 5). Por tanto, al final de la competición el estudiante podrá elegir como máximo dos de los envíos para su entrega. En caso de no elegir ninguno, se asignarán por defecto los dos primeros (con el mejor resultado en el conjunto "público" de test).

Será con el conjunto privado de test con el que se asigne finalmente el ranking de cara a la evaluación final de este apartado.

3. Clasificación de Imágenes

En este segundo problema se pretende tomar una imagen de un dígito escrito a mano y determinar de manera automática cuál es. El conjunto de datos seleccionado para esta tarea será el de MNIST, es decir, el mismo que se utilizó en clase de prácticas.

La Web de esta competición se encuentra en:

https://www.kaggle.com/c/digit-recognizer

desde donde el estudiante podrá acceder a mucha más información relativa a este problema.

Aunque en clase de prácticas se explicó con detalle el formato de los ficheros de entrenamiento y test, es necesario que revise tranquilamente la documentación de la Web de *Kaggle* al respecto. Del mismo modo, deberá fijarse en la estructura del fichero solución para el envío y posterior puntuación.

Se podrán enviar hasta 5 ficheros de prueba por día para comprobar el ranking obtenido mediante la clasificación realizada en el conjunto de test.

Las herramientas, metodología, número y tipo de pruebas a realizar, etc. serán decisión unilateral de cada estudiante.

4. Documentación a Entregar

Cada estudiante deberá enviar, vía Prado, un documento en PDF con la siguiente información:

- Nombre y apellidos del estudiante.
- **Nombre del usuario de Kaggle seleccionado**.
- "Bitácora" del proceso de aprendizaje y clasificación
- Resumen de todas las pruebas realizadas y resultados obtenidos

Es **muy importante** definir con claridad el nombre del usuario en Kaggle dado que éste será el utilizado para chequear el ranking obtenido por el estudiante para asignar una parte de la nota final (ver Sección 5).

La "bitácora" se refiere a una redacción completa sobre todos los pasos seguidos hasta llegar a la metodología que permitió alcanzar los mejores resultados. De este manera, el estudiante incidirá en cuáles fueron las principales ideas e hipótesis que se pusieron en práctica, los fallos, los éxitos, y en definitiva la sucesión de decisiones que se tomaron hasta llegar al objetivo final, y las razones detrás de las mismas.

Se requiere la inclusión de una TABLA que detalle la secuencia/modificación de cada uno de los envíos a la plataforma Kaggle. Esta tabla deberá constar de una línea por cada envío realizado (no hay límites en el número de intentos) con el valor de rendimiento obtenido y una descripción breve del algoritmo / parámetros / modificación / secuencia que consigue dicho resultado.

Por último, el resumen de las pruebas realizadas deberá incluir UN gráfico con los diferentes valores de rendimiento alcanzados para cada envío de Kaggle, indicando qué metodología fue utilizada en los principales hitos del proceso.

5. Sistema de Evaluación

IMPORTANTE: *Solo deberá realizar una de las dos competiciones programadas.*

La nota final será la suma ponderada de dos apartados:

1. Posición obtenida en el ranking de la competición (No balanceada o MNIST): 50% del total.
2. Documentación del proceso realizado: 50% del total.

Recordamos que el *plazo límite de la competición* será el día 0**7/04/2017 a las 23.59**. No se permitirá el envío de ningún resultado adicional después de dicha fecha.

La *fecha de entrega de la documentación* en PRADO será el día **21/04/2017 a las 23.55**. IMPORTANTE: No se aceptará ningún envío a través del correo electrónico ni ninguna otra vía.

RECORDATORIO: la documentación debe especificar claramente el usuario que habéis utilizado para la competición. En caso de no indicarse NO se evaluará la práctica.