Minería de datos: preprocesamiento y clasificación

3 de febrero de 2017

ugr | Universidad de Granada

Práctica para evaluación de asignatura: competición Kaggle

Máster Ciencia de datos e Ingeniería de Computadores, curso 2016-2017

Índice

---

1. Objetivos y evaluación

En esta práctica se pondrán en juego los métodos de preprocesamiento y aprendizaje vistos en la asignatura **Minería de datos: preprocesamiento y clasificación**. Para ello se hará uso de la plataforma **Kaggle**, que permite establecer una competición 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.

La evaluación de la práctica se hará 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.

2. Descripción del problema y reglas de la competición

La competición se centra en el uso de los métodos de clasificación (árboles de decisión, métodos **ensamble**, reglas, máquinas de soporte vectorial, etc) y de preprocesamiento (discretización, tratamiento de valores perdidos, eliminación de instancias con ruido, selección de características, etc) descritos en la asignatura, pero aplicados ahora a un problema real. El acceso a la plataforma de competición se hace mediante el enlace:

`https://inclass.kaggle.com/c/datcom-preprocesamiento-clasificacion`

Para poder participar cada estudiante debe registrarse en **Kaggle**, haciendo uso de la invitación recibida por el correo asignado por 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.

El conjunto de datos de entrenamiento se denomina *accidentes-kaggle.csv* (formato separado por comas). Contiene 30002 instancias y 30 variables. La variable clase se denomina *TIPO_ACCIDENTE* y sus valores caracterizan diferentes tipos de accidentes en carreteras españolas: *Atropello, Colision_Obstaculo, Colision_Vehiculos, Otro, Salida_Via, Vuelco*. Es sobre este conjunto sobre el que tendréis que aplicar las técnicas de preprocesamiento y aprendizaje que consideréis más oportunas con el objetivo de obtener los mejores resultados posibles en la clasificación de las instancias del conjunto de test.

El conjunto de test se denomina *accidentes-kaggle-test.csv* e incluye 19998 instancias que carecen de valor para la variable clase. Estos dos conjuntos de datos están disponibles en la plataforma de la competición.

El trabajo debe basarse en la repetición del siguiente procedimiento las veces que consideréis oportunas:

- preprocesar y aprender algún modelo sobre el conjunto de entrenamiento, usando validación cruzada
- usando este modelo, predecir la etiqueta de cada una de las instancias del conjunto de test
- enviar la predicción a **Kaggle**

La evaluación realizada por **Kaggle** se basa en el cálculo del porcentaje de bien clasificados de cada envío. Este cálculo puede realizarse ya que la plataforma contiene un archivo con las etiquetas correctas de todas las instancias del conjunto de test. De todas estas instancias, algunas están dedicadas a ofrecer la evaluación del modelo a lo largo del tiempo de vida de la competición. El resto de las instancias se reservan para producir el valor final de evaluación cuando la competición finaliza. De esta forma, se intenta poner a prueba la capacidad de generalización del modelo.

3. Archivo de predicción a subir a Kaggle

Una vez tratado el conjunto de entrenamiento de forma adecuada mediante las técnicas de preprocesamiento que consideréis oportunas y tras haber aprendido algún modelo, hay que realizar la predicción de las instancias del conjunto de test. A partir de la predicción se generará el archivo que hay que subir a **Kaggle**, con la siguiente estructura:

```
Id, Prediction
1, Colision_Vehiculos
2, Salida_Via
3, Atropello
.....
19998, Otro
```

Es decir, debe contar con una fila inicial con los identificadores **Id** y **Prediction** y tantas líneas como instancias hay en el conjunto de test (19998). Para cada instancia se muestra su identificador y, separado por una coma, la predicción que el modelo realiza para ella.

4. Entrega

Se trabajará en la competición de forma individual. La competición finaliza el día 31 de Marzo, a las 23:59 horas. Se dispondrá de dos semanas adicionales (hasta el 14 de Abril) para generar la documentación donde debéis explicar las técnicas de preprocesamiento y clasificación que se han ido empleando a lo largo del trabajo en la competición, justificando, en la medida posible, las razones por las que se han ido descartando o adaptando durante vuestra experimentación con los datos. Es decir, se trata de elaborar una especie de diario de trabajo que refleje los pasos seguidos desde la recepción de los datos hasta la entrega final.

El estudiante debe entregar dos archivos en la plataforma **PRADO**, en la parte correspondiente a la asignatura. El primero contendrá el código **R** necesario para generar la solución final entregada a la competición, suficientemente explicado como para poder ejecutarlo y obtener los resultados incluidos en la memoria (a partir de los datos de partida). El segundo documento (en formato **pdf** exclusivamente) describe el trabajo llevado a cabo, siguiendo las indicaciones dadas en el párrafo superior.

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