Minería de datos: preprocesamiento y clasificación

25 de noviembre de 2021

UNIVERSIDAD DE GRANADA

Práctica para la evaluación de asignatura: competición DrivenData

Máster en Ciencia de Datos e Ingeniería de Computadores, Curso 2021-2022

Índice

---

1. Objetivos y evaluación

En esta práctica se deberán utilizar los métodos de preprocesamiento y aprendizaje vistos en la asignatura **Minería de datos: preprocesamiento y clasificación**. Para ello se hará uso de la plataforma **DrivenData**, 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 varios criterios, como el desarrollo del trabajo realizado, la coordinación del equipo, la posición que ocupen los resultados ofrecidos por los modelos de cada equipo en la competición, la exposición del trabajo realizado, la diversidad de técnicas utilizadas, ... La documentación que debe entregar cada equipo, consistente en una presentación que debe cumplir con un mínimo de especificaciones indicadas en la Sección 5, debe detallar de forma clara el proceso de trabajo seguido desde el inicio hasta la entrega de los resultados finales.

2. Composición de los equipos

Los equipos se componen de 5 miembros. Los equipos deben formarse a lo largo del curso, de forma que durante las últimas sesiones del curso informen a los profesores de los componentes del equipo, junto con el nombre del mismo, así como una previsión de la distribución del trabajo que se realizará.

3. Descripción del problema y reglas de la competición

La competición se centra en el uso de cinco métodos de clasificación (kNN, árbol de clasificación C4.5/J4.8, árbol de clasificación (CART/RPART), algoritmo de reglas (RIPPER/JRip) y Naïve Bayes), junto con las técnicas de preprocesamiento (discretización, tratamiento de valores perdidos, eliminación de instancias con ruido, selección de características...) que se consideren oportunas para cada una de las técnicas de aprendizaje indicadas. Quedan incluidas todo tipo de técnicas basadas en multiclasificación o

ensembles, siempre y cuando se implementen manualmente utilizando como clasificador débil cada uno de los cinco métodos mencionados, desarrollando el método de división y el método de agregación, entre otras componentes que se estimen oportunas. También se permite el uso de técnicas de descomposición de problemas (OVO, OVA) en caso necesario. No estará permitido usar la implementación disponible de algoritmos como RandomForest, Adaboost, XGBoost, etc.

La manera de trabajar entre los componentes del equipo consiste en que cada uno de ellos se ocupe de una técnica de aprendizaje, del procesamiento necesario y el desarrollo (opcional) del modelo de ensemble asociados a la técnica de aprendizaje; poniendo luego en común entre todos los resultados obtenidos y ofreciendo posibles soluciones o ideas al resto de componentes del equipo.

Se recomienda el trabajo en local (en vuestras máquinas) con el conjunto de datos de entrenamiento que se os ofrece. También se dispondrá de un conjunto de datos de test en el que se ha eliminado la columna correspondiente a la variable clase. El objetivo es clasificar estas instancias de test y enviar el resultado a la plataforma DrivenData, en el formato que se indica más adelante. Cabe destacar que la competición es pública y ya está activa, por lo que el profesorado no la gestiona y se compartirá la tabla de puntuación (leaderboard) con otros usuarios en todo el mundo.

El nombre del equipo asociado a la actividad no se debe utilizar en la competición de **DrivenData**. La idea es que cada usuario se registre por separado y trabaje independientemente con su algoritmo de aprendizaje asignado. De esta manera, la limitación de pruebas (se detallará más adelante) afecta a cada algoritmo y no al grupo entero. La asignación de cada miembro al equipo se indicará en la documentación de la práctica. Si alguien tiene problemas para entrar que se ponga en contacto con los profesores de la asignatura para solventarlo lo antes posible.

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

`https://www.drivendata.org/competitions/66/flu-shot-learning/page/211/`

El conjunto de datos de entrenamiento se denomina *training_set_features.csv* y *training_set_labels.csv* (formato separado por comas). Contiene 26707 instancias y 35 variables predictivas (la lista está disponible en la web de la competición). La variable clase están en un fichero .csv diferente y sus valores posibles son 0 y 1. El conjunto es multi-etiqueta, tiene dos variables de salida binaria. Se puede tratar como un conjunto de cuatro clases, considerando la combinación de los valores de este modo: $\{0,0\},\{0,1\},\{1,0\},\{1,1\}$. 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 incluye 26708 instancias en las que solo aparecen las variables predictivas. 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/s etiqueta/s de cada una de las instancias del conjunto de test,

- enviar la predicción a **DrivenData**.

La evaluación realizada por **DrivenData** se basa en el cálculo de la tasa de los ejemplos correctamente 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. En nuestro caso, la evaluación de esta actividad termina antes que la finalización de la competición, por lo que nos basaremos exclusivamente en la tasa que nos ofrece **DrivenData** durante el transcurso de la práctica.

Para la realización de esta actividad estará permitido el uso de dos lenguajes de programación y sus bibliotecas o paquetes: **R** y **Python**. **R** y cualquier paquete (visto en clase o no) estará permitido para cualquier tipo de tarea. **Python** y sus bibliotecas (por ejemplo *scikit-learn*, *NumPy*) ó *Pandas* solo estará permitido para realizar tareas de preprocesamiento de datos, debido a que el tamaño del conjunto de datos puede ser demasiado grande para algunas implementaciones integradas en **R**. Se recomienda la utilización de **Python** para preprocesamientos simples y formateado de datos, y **R** para preprocesamientos de datos más complejos. Por tanto, los clasificadores que se deben usar son exclusivamente de **R**.

El conjunto de datos propuesto contiene exclusivamente valores categóricos (nominales) en sus atributos. Cuatro de los algoritmos de clasificación propuestos trabajan con valores categóricos de forma natural, mientras que el algoritmo kNN requiere del uso de métricas de distancia concretas para la gestión de valores categóricos o técnicas de codificación. En clase de teoría se explicarán algunas de estas métricas/técnicas y el estudiante que escoja este método, deberá utilizar/implementar al menos tres métricas/técnicas distintas vistas en clase (o no).

Se deberá conservar dos registros de resultados para cada algoritmo: el resultado sin usar multi-clasificación (ensemble) y el resultado usando multi-clasificación[1]. De esta manera, se apreciará mejor el margen de mejora de los métodos de preprocesamiento sobre el algoritmo original y el alcance de la multiclasificación en cada método. Nótese que el/los ensemble/s implementado/s se podrá/n realizar por todo el equipo y compartir desarrollos entre cada miembro solo cambiando el clasificador débil en cada caso.

4. Archivo de predicción a subir a DrivenData

Una vez tratado el conjunto de entrenamiento de forma adecuada mediante las técnicas de preprocesamiento que consideréis oportunas y tras haber aprendido algún modelo usando los parámetros que mejor resultado hayan dado en la experimentación, 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 **DrivenData**, con la siguiente estructura:

[1]Se recomienda hacer la versión multi-clasificador pero es opcional, aunque su no realización se contemplará en la valoración del trabajo realizado por el estudiante

```
respondent_id, h1n1_vaccine, seasonal_vaccine
26707, 0.5, 0.7
26708, 0.5, 0.7
26709, 0.5, 0.7
26710, 0.5, 0.7
26711, 0.5, 0.7
.....
53414, 0.5, 0.7
```

Es decir, debe contar con una fila inicial con los identificadores **respondent_id**, **h1n1_vaccine** y **seasonal_vaccine** y tantas líneas como instancias haya en el conjunto de test (26708). Para cada instancia se muestra su identificador y, separados por comas, la probabilidad de tener valor 1 en la primeta etiqueta (probabilidad de cada persona en recibir la vacuna correspondiente) y la probabilidad de tener valor 1 en la segunda etiqueta. Estas probabilidades las proporciona cada algoritmo utilizado y se utilizan para calcular la medida AUC de las curvas ROC en entornos multi-etiqueta utilizando una macro-average.

La competición tiene una limitación estricta de 3 subidas de resultados diarias, que serán aplicada a cada algoritmo de aprendizaje por separado, pues cada usuario tiene un algoritmo distinto asignado.

5. Entrega

La competición finaliza el día 14 de Enero del 2022, a las 23:55 horas. Nótese que la competición continúa hasta Marzo de 2022, pero no se considerarán envíos posteriores a esa fecha en la evaluación de esta actividad. Se dispondrá de un día adicional (hasta el 15 de Enero, a las 23:55) para generar la documentación (presentación) donde debéis explicar el trabajo realizado por cada miembro del equipo y las colaboraciones entre vosotros. Este trabajo será expuesto en público en la fecha que se indicará para la evaluación de la asignatura.

Cada equipo debe entregar dos archivos en la plataforma **PRADO**, en la parte correspondiente a la asignatura. El primero contendrá (en un comprimido) el proyecto software (todos los scripts de **R** y **Python**) generado durante el trabajo en la práctica, suficientemente explicado como para poder ejecutarlo y obtener los resultados incluidos en la memoria (a partir de los datos de partida). El segundo documento, que corresponde a las diapositivas con la presentación del trabajo (en formato **pdf** exclusivamente), describe el trabajo llevado a cabo, indicando de forma clara el trabajo realizado por cada componente del equipo y el trabajo común realizado (preprocesamiento necesario para todas las técnicas, aportación de ideas, etc).

La documentación constará de una presentación de unas 8 diapositivas por algoritmo de clasificación (suponiendo un máximo de 40 diapositivas por equipo) más una diapositiva que sirva de portada con el nombre del los componentes del equipo, nombre del equipo y asignación de clasificadores a cada miembro. Será obligatorio incluir dentro de esas 8 diapositivas por clasificador:

- Una diapositiva final con la evolución del ranking y de la tasa de acierto asociado a cada entrega en DrivenData, en formato gráfico. El valor de ranking y tasa de acierto final conseguido, junto a número de subidas final que se han realizado a DrivenData.

- Una diapositiva con la lista completa de algoritmos de preprocesamiento utilizados y sus diferentes configuraciones a lo largo de la práctica (qué se ha probado en el transcurso de la misma con intervalos de parámetros). Además, también se considerarán las diferentes configuraciones (parámetros) del algoritmo de aprendizaje.
- Una diapositiva con el detalle de la estrategia que ha obtenido el mejor ranking (*pipeline* de técnicas) y su justificación.

El resto de diapositivas se puede utilizar para describir el proceso y detallar mejor los resultados, motivaciones, justificaciones, decisiones tomadas, etc. No se debe incluir código en la presentación ni descripciones teóricas de los algoritmos explicados en clase. Sí que será necesario describir aquellas técnicas utilizadas que no se hayan visto en clase (breve descripción), incluir referencias, etc...

NOTA: la documentación debe especificar claramente el usuario asignado a cada equipo para la competición.