Instituto Tecnológico de Buenos Aires
Maestría en Ciencia de Datos
Data Mining , 2023 B

Tarea para el Hogar UNO

Este documento queda definitivo el viernes 15 de septiembre a las 23:00

Las Tareas para el Hogar no contribuyen con nota a la asignatura, son simplemente muy importantes andamios de aprendizaje que lo ayudan a escalar con mayor facilidad los principales conceptos, a comprender los scripts oficiales, a que usted pueda decir “entiendo lo que estoy haciendo” y no que le suceda que diga al final de la asignatura “lo único que hice fue correr scripts”.

Las Tareas para el Hogar tienen actividades que debe realizar ya que serán discutidas en la siguiente clase presencial.

Iniciar la asignatura implica instalar todas las herramientas que utilizaremos en el laboratorio.
Lea estratégicamente las tareas, y a partir de su computer literacy, tiempo disponible y gustos encárelas en el orden que más le convenga.

En las Tareas para el Hogar hay tres secciones : Pasado, Deseable y Complementaria.

- *Sección Pasado*, tareas que usted ya debería haber hecho para estar al día, por si quizas faltó a alguna de las clases. Es de extrema utilidad para quienes no han podido asistir a alguna de las clases.
- *Sección Deseable*, ejercicios y problemas que TODOS los alumnos deberían hacer para estar al día, entender lo que se discute la siguiente clase y finalmente aprobar la asignatura.
- *Sección Complementaria*, que es para alumnos que dispongan principalmente de tiempo, motivación y una base de conocimientos adecuada.

Son ineludibles

- *10. Grid Search Optimización de Hiperparámetros*
- *12. Corrida Colaborativa Árboles Azarosos*

Sección Pasado

Estos son todos temas que ya debería haber realizado, están aquí simplemente como una recapitulación.

1. **Videos y Presentaciones Clases 01 y 02 (12 y 15 de septiembre)**
   (tiempo estimado 15 minutos a 1.5x , dificultad baja)
   - Hiperparámetros de un Árbol de Decisión
     - pres Hiperparámetros de un Árbol de Decisión
     - video Hiperparámetros de un Árbol de Decisión
   - Optimización de Hiperparámetros
     - pres Optimización Hiperparámetros
     - video Optimización Hiperparámetros

Existe una extensión para Google Chrome https://chrome.google.com/webstore/detail/video-speed-controller/nffaoalbilbmmfgbnbgppjihopabppdk?hl=en que permite ver los videos de la plataforma a distintas velocidades en caso que su navegador no le estuviera mostrando los controles

- Instale la extensión en el navegador Chrome/Brave
- Play al video
- Presionando la tecla "d" se aumenta la velocidad de play, con la tecla "s" se disminuye

2. **Arranque en Frío**
   Del Libro de la Asignatura si le han quedado pendientes, realice por única vez todas las altas en plataformas e instalaciones de aplicaciones solicitadas en el
   - capítulo 3 Arranque en Frío

   solicite ayuda en Zulip si se le presenta algún problema
   y realice el Check In a la Asignatura
   (tiempo promedio estimado 60 minutos, dificultad media)

3. **Temas Administrativos Asignatura**
   Del Libro de la Asignatura, si le han queado pendiente, leer por única vez
   (tiempo estimado 25 minutos, dificultad baja)
   - capítulo 1 La Asignatura
   - capítulo 2 Metología de Enseñanza

4. Operación de la Asignatura

Del Libro de la Asignatura , si le ha quedado pendiente, para aprender a manejarse eficientemente con Zulip leer:
(tiempo estimado 10 minutos, dificultad baja)

- capítulo 4 Herramientas, Conceptos, Operación y Buenas Prácticas

5. Diccionario de Datos

Leer detenidamente el archivo `DiccionarioDatos` que se encuentra en
https://storage.googleapis.com/open-courses/itba2023b-52fa/DiccionarioDatos_2023.ods
Le permitirá conocer los campos que posee el dataset y comenzar a ganar intuición sobre las variables más importantes que afectan la predicción y así poder crear variables derivadas, tarea llamada Feature Engineering en la Ciencia de Datos y que permite aumentar el poder predictivo de los modelos, en nuestro caso, aumentar la ganancia.
Plantear dudas y observaciones en Zulip.
(tiempo estimado 15 minutos, dificultad baja)

6. Variabilidad de la Ganancia

Esto corresponde a una actividad realizada en la clase del viernes 15 de septiembre.
Lea en detalle el script `src/rpart/z211_traintest_estratificado.r` , ejecútelo línea a línea, entienda en profundidad lo que hace.
Pruebe correrlo con cada una de sus cinco semillas aleatorias, cambiándolas en la linea que dice

```
 particionar( dataset, division=c(7,3), agrupa="clase_ternaria", seed= 102191 )
#Cambiar por la primer semilla de cada uno !
```

¿Se esperaba la variabilidad que observó?
Para la corrida con su primer semilla, cargue los valores que le devolvió en la Planilla Colaborativa en la hoja `z211`
(tiempo estimado 10 minutos, dificultad baja)

7. Estimación Montecarlo

La primera parte corresponde a una actividad realizada en la clase del viernes 15 de septiembre.
Lea con detalle el script src/rpart/z222_traintest_montecarlo.r , ejecútelo línea a línea, entienda en profundidad lo que hace.
Pruebe correrlo utilizando sus cinco semillas aleatorias, cambiándolas en la línea

```
linea ksemillas  <- c(102191, 200177, 410551, 552581, 892237 ) #reemplazar por las propias semillas
```

Cargue los valores que le devolvió en scripten la Planilla Colaborativa en la hoja `Montecarlo05`

Realice una nueva corrida, pero ahora en lugar de 5 semillas, utilice 20 (procúrelas en alguna tabla de números primos), cargue los valores en la Planilla Colaborativa en la hoja `Montecarlo20`

¿Como comparan los valores al pasar de 5 semillas a 20 semillas?

(tiempo estimado 10 minutos, dificultad baja)

8. Instalación Google Cloud

En esta asignatura entrenaremos complejos modelos predictivos que demandarán decenas de horas de procesamiento en grandes servidores en la nube.
Primero debe tener una cuenta de gmail, y luego crear una cuenta de Google Cloud en donde le van a regalar USD 300 por 3 meses https://cloud.google.com/
Allí es donde Google le va a pedir una tarjeta de crédito, le va a debitar un dólar y se los vuelverá a acreditar, este paso lo realiza Google para evitar abusos.

Seguir este instructivo https://storage.googleapis.com/open-courses/itba2023b-52fa/GoogleCloud_ITBA2023b.pdf
El instructivo es hiperdetallado y tiene la gran ventaja que instala sobre un ambiente virgen en Google Cloud, no como el Arranque en Frío que fue sobre cada laptop distinta.

La instalación le demandará unas dos horas, de las cuales 65 minutos serán de forma desatendida.

Sección Deseable

9. Fascículos Coleccionables zero2hero

Este punto está orientado a alumnos que cumplan alguna de estas condiciones (conector lógico OR) :

- sin experiencia en programación
- menos de 300 horas de experiencia en lenguage R
- sin experiencia en la librería `data.table`
- tiene curiosidad por hackear Kaggle

Se han lanzado los primeros fascículos coleccionables llamados "from Zero to Hero" que muy detalladamente, paso a paso enseñan todo lo necesario de R para entender los scripts oficiales de la asignatura. Están en formato Jupyter Notebook, usar Jupyter Lab o directamente VSCode.
Podrá encontrar los Jupyter Notebooks zero2hero en el repositorio GitHub de la asignatura carpeta `./src/zero2hero`

Utilice Zulip para preguntar/comentar en el stream # zero2hero
(tiempo estimado 60 minutos, , dificultad media)

10. Grid Search - Optimización Hiperparámetros

Modifique el script `src/rpart/z241_gridsearch_esqueleto.r` para que recorra TODOS los hiperparámetros de rpart, es decir, debe agregar loops, y luego póngalo a correr.
En la línea

```
ksemillas  <- c(102191, 200177, 410551, 552581, 892237) #reemplazar por las propias semillas
```

reemplece por sus propias semillas

La corrida, bien hecha, con los cuatro loops anidados que corresponde, y una rica granularidad, le llevará alrededor de 4 horas.

Una vez que termine, cargue la salida en un Excel, ordene en forma descendente por ganancia.
Llamaremos posición 1 del ranking a la combinación de parámetros que obtuvo la mayor ganancia en los datos de testing, posición 2 a la segunda mejor ganancia, y así sucesivamente.
Para cada una de las combinaciones de hiperparámetros de las posiciones del ranking { 1, 2, 3, 10, 50, 100 } , vaya al script `src/rpart/z101_PrimerModelo.R` reemplace por esos hiperparámetros, y suba el archivo a Kaggle, y obtenga la ganancia en el Public Leaderboard.
Cargue los correspondientes valores de los hiperparámetros, la ganancia en testing y la ganancia del Public en la Planilla Colaborativa, hoja *GridSearch*

(tiempo promedio estimado 30 minutos, dificultad media-alta)

11. Videos Ensembles

(tiempo estimado 22 minutos a 1.5x, dificultad media)

1. Cambiando la clase para mejorar el modelo
   - pres Cambiando la Clase
   - video Cambiando la Clase
2. Ensembles de arboles de decision
   - pres Ensembles de Árboles de Decisión
   - video Ensembles de ÁRboles de Decisión

12. *Corrida Colaborativa* Arboles Azarosos

Lea en detalle el script `src/ArbolesAzarosos/z321_arboles_azarosos.r` y entienda lo que hace, ejecútelo línea a línea si hace falta.
Vea en la Google Sheet Colaborativa, hoja `ArbolesAzarosos` las corridas que le corresponde hacer a usted
Haga sus corridas con el script, suba a Kaggle las salidas, y anote en la hoja ArbolesAzarozos de la Google Sheet compartida sus resultados, tenga muy presente la limitación que solamente se pueden hacer hasta 20 submits diarios a Kaggle; le va a llevar al menos tres dias hacer todos los submits.
El resultado va en la columna “Public Leaderboard” . En la columna semilla, reemplace “sem1” por la semilla que utilizó, que debe ser su primer semilla.

hint: puede abrir varias sesiones simultaneas de R/RStudio para correr en paralelo sus scripts.
Por favor, sus corridas deben estar finalizadas para ANTES de las 17:00 del viernes 12 de mayo, ya que en clase la analizaremos y miraremos a la cara al overfitting.

(tiempo estimado 60 minutos, dificultad baja)

Sección Complementaria

13.zero2hero Optimización Bayesiana

Prerequisito: *NA*
Se han lanzado los fascículos 0112 al 0202 de "from Zero to Hero"
Podrá encontrar los Jupyter Notebooks zero2hero en el repositorio GitHub de la asignatura carpeta `./labo/src/zero2hero`

Utilice Zulip para preguntar/comentar en el stream # zero2hero
(tiempo estimado 20 minutos, , dificultad media)

14.Script Optimización Bayesiana

Prerequisito: punto 13. *zero2hero Optimización Bayesiana*
Lea en detalle el script `./labo/src/rpart/z321_rpart_BO.r` y entienda lo que hace, ejecútelo línea a línea si hace falta.
Corra el script, cambiando por sus semillas
La salida de la optimizacion bayesiana queda en `./exp/HT3210/HT321.txt`
cargue la salida en una planilla, ordénelo en forma descendente por el campo ganancia, y analice cuales son los mejores hiperparametros.
Compárelos con los que obtuvo en el Grid Search.
A los mejores hiperparámetros, carguelos en el script `./src/rpart/z101_PrimerModelo.R`
genere la salida para Kaggle, y súbala.
¿ Mejoró o emperó respecto al Grid Search en el Public Leaderboard ?

El tiempo de corrida desatendida del script será de unas 2 horas.
(tiempo estimado 30 minutos, , dificultad alta)

15.data.table vs data.frame

En el lenguaje R tradicional se utiliza el objeto `dataframe` que posee severas deficiencias en cuanto a performance y sintaxis complicada. En la materia utilizamos la librería `data.table` que es ampliamente superadora y que permite manejar grandes volúmenes de datos, tiene una sintaxis más simple PERO muy distinta a la de dataframes, por lo que es necesario aprenderla.
Leer los siguientes artículos
https://towardsdatascience.com/data-table-rs-best-data-object-c95b7d5f0104
https://towardsdatascience.com/blazing-fast-data-wrangling-with-r-data-table-de5045cc4b4d

Si realiza alguna prueba por su cuenta, haga los comentarios en Zulip, de igual forma si es un férreo defensor de tidyverse , dplyr, p Pandas en Python ¡ No le tenemos miedo !

(la correctitud de estos artículos ligeros ha sido verificada por el profesor)
(tiempo estimado 15 minutos, dificultad media)