Universidad Austral
Maestría en Ciencia de Datos
Laboratorio de Implementación I , 2024 virtual1

Fecha de liberación oficial: miércoles 03 de abril 23:01

Tarea para el Hogar TRES

Hágase amigo de los scripts, ejecútelos línea a línea, consulte en ChatGPT que hace cada línea, vaya siguiendo que va quedando en cada variable. Haga una copia y pruebe cambiar partes y observe como cambia el resultado, experimente, juegue!
No corra los scripts como un autómata obediente carente de emocinoes, hagalos suyos.
Realmente no le queremos engañar de la cátedra : entender los scripts le va a demandar bastante tiempo el cual va a depender principalmente de su expreciencia previa en programación. Pero es mejor hacerlo ahora que tiene más de 70 compañeros y dos profesores a quien preguntar ya que jamás en su vida volverá a estar en un proyecto donde hay tantas personas haciendo exactamente las mismas tareas que usted, aproveche esta oportunidad única, le está costando dinero y principalmente *su invaluable tiempo.*
Eleve su espíritu por sobre la robotizada literalidad inicial, resuene con el código, capte el espíritu de los scripts, los diálogos entre los distintos bloques de código, las tensiones en las optimizaciones, perciba las emociones plasmadas en ellos.

Sección Pasado (ya lo debería haber hecho)

1. Videos de Ensembles de Árboles de Decisión
   De la clase 3 del campus
   vea los videos
   - Ensembles de Arboles de Decisión
   - Algoritmo Random Forest

   (tiempo estimado 30 minutos a 1.5x, dificultad media)

2. zero2hero
   Ver 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)

3. Script almacenamiento Mis Semiilas

   Si no pudo hacerlo durante la Clase 03 hágalo ahora, ya que es requisito fundamental para correr todos los scripts que siguen en la asignatura.

   En el script `src/ambiente/z301_GrabarMisSemillas.r` reemplace en la linea 10 por sus cinco semillas y ejecutelo.
   Esto hará que se almacenen en SU bucket sus cinco semillas, de forma que todos los siguientes scripts irán a buscarlas allí.
   (tiempo estimado 2 minutos, , dificultad muy baja)

4. Script Optimización Bayesiana

   Usted debería haber corrido durante la Clase 03 el siguiente script
   Lea en detalle el script `src/rpart/z321_rpart_BO.r` y entienda lo que hace, corralo línea a línea si hace falta.
   La salida de la optimización 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 a la solapa “bayesian” de la planilla colaborativa.
   ¿ Mejoró o emperó respecto al Grid Search en el Public Leaderboard ?

   El tiempo de corrida desatendida del script será de alrededor de 30 minutos, fue pensado para que lanzado al comienzo de la clase, finalizara durante la clase con tiempo más que suficiente para analizar los resultados.

Sección Deseable

5. El overfitting en TODA la ciencia
   Lea el siguiente artículo https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124&xid=17259,15700019,15700186,15700190,15700248
   (20 min)

6. Ensembles de Modelos
   Vea el siguiente hermoso video
   - *https://www.youtube.com/watch?v=iOucwX7Z1HU* (5 min)

   Lea los siguientes artículos
   - https://www.all-about-psychology.com/the-wisdom-of-crowds.html (10 min)
   - https://machinelearningmastery.com/what-is-ensemble-learning/ (10 min)

7. *Análisis Curvas de Ganancia*
   Usted deberá jugar/experimentar con el script `src/rpart/z365_curva_ganancia.r`
   Lea en detalle el script, comprendiendo cada sección
   Experimente cambiando los hiperparámetros del arbol, a partir de valores que ha visto en su corrida de Grid Search
   ¿Qué sucede con arboles de profundidad 20 y hojas pequeñas?
   ¿Qué sucede con arboles de profundidad 3 y hojas grandes?
   ¿Le sorprende la enorme diferencia que hay entre las ganancias de training y testing?
   ¿Qué conclusión original obtiene?
   Escriba en Zulip stream `#v-Tarea Hogar 03`, topic : `Curvas de Ganancia` las conclusiones que obtiene

8. *Corrida Colaborativa* Árboles Azarosos
   Haremos una gigantesca corrida entre todo el curso para encontrar los hiperparámetros óptimos del algoritmo Arboles Azarosos.
   Lo conceptual de este script se explicará en la Clase 04, y luego de la explicación abordaremos los resultados de las corridas que ustedes habrán corrido colaborativamente durante la semana. A mayor cantidad de alumnos que colaboren en la corrida, más probabilidad. En el siguiente punto de esta Tarea para el Hogar se le pide que usted mismo saque conclusiones de los resultados,

ANTES de la clase.
Lea en detalle el script `src/ArbolesAzarosos/z351_arboles_azarosos.r` y entienda lo que hace, córralo linea a linea si hace falta.
Vea en la Google Sheet Colaborativa, solapa `ArbolesAzarosos`, el trabajo que le corresponde hacer a usted
Haga sus corridas con el script el que tomará automaticamente sus semillas, suba a Kaggle las salidas, y anote en la 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” .

hint: puede abrir varias sesiones simultaneas de R/RStudio/VSCode para correr en paralelo sus scripts.
Por favor, sus corridas deben estar finalizadas para ANTES de la Clase 04 del martes 09 de abril

Si no encuentra su nombre en la solapa Árboles Azarosos, escríba a @profesores por Zulip que le generarán su corrida.

(tiempo estimado 60 minutos, dificultad baja)

9. Análisis Arboles Azarosos

Analice los resultados del experimento Arboles Azarosos, conteste las siguientes preguntas en Zulip, stream #Tarea Hogar 03, topic Arboles Azarosos

- ¿A mayor cantidad de arbolitos, mejora el modelo? Elévese por sobre la literalidad de esta pregunta y los casos particulares.
- ¿La combinación de hiperparámetros que genera el mejor árbol unitario (arbolitos==1), es la misma que genera el mejor ensemble de 500 arboles?
- Cada set de hiperparámetros se ha corrido por DOS alumnos, cada uno con su semilla. ¿Que sucede con la variabilidad de los resultados a medida que aumenta la cantidad de arbolitos?
- ¿Qué otra conclusión puede obtener de los datos?

(tiempo estimado 30 minutos, dificultad alta si se hace bien)

10. Videos Prioritarios clase 04

ver los videos que aún no ha visto, clase 04 del campus virtual

1. Data Drifting

tiempo humano estimado: 10 **minutos** (a 1.5x)

Seccion Complementaria

11. Script Optimización Bayesiana 10-repeated 5-fold cross validation

Se ejecutará una 10-repeated 5-fold cross validation
Lea en detalle el script `src/rpart/z333_rpart_repe_BO.r` y entienda lo que hace, corralo línea a línea si hace falta.
La salida de la optimización bayesiana queda en `./exp/HT3330/HT333.txt`

- cargue la salida en una planilla, ordénelo en forma descendente por el campo ganancia, y analice cuales son los mejores hiperparametros.
- A los mejores hiperparámetros, carguelos en el script `./src/rpart/z101_PrimerModelo.R` genere la salida para Kaggle, y súbala a la solapa "bayesian" de la planilla colaborativa.
- ¿ Mejoró o emperó respecto al Grid Search en el Public Leaderboard ?
- Analice en profundidad la diferencia con `z321_rpart_BO.r` en cuanto a los resultados obtenidos, la variabilidad de las ganancias, y que tan bueno es el resultado en Kaggle

El tiempo de corrida desatendida del script será de unas 5 horas.
(tiempo estimado 30 minutos, , dificultad alta)

12. Videos complementarios clase 04

ver los videos que aún no ha visto, clase 04 del campus virtual

1. Workflow de trabajo
2. Catastrophe Analysis
3. Feature Engineering Intra Mes
4. Feature Engineering Histórico

tiempo humano estimado: 35 **minutos** (a 1.5x)

13. Análisis Variables Rotas

Analizará un problema que el sector de DataWarehouseing ha tenido en la generación de los datos, tal cual se mostró en el video Catastrophe Analysis

Desde la máquina virtual `desktop` corra el script
`src/CatastropheAnalysis/z505_graficar_zero_rate.r`

El proceso demorará alrededor de 20 minutos

La salida del script queda en su bucket, en la carpeta `~/buckets/b1/exp/CA5050`
Analice los archivos, en particular `zeroes_ratio.pdf`

tiempo computacional: **20 minutos**
tiempo humano estimado : **15 minutos**
dificultad : **baja**
creatividad requerida : 1**0%**