

Organización de Datos (75.06/95.58)

Primer Cuatrimestre 2021

Trabajo Práctico N°2

Curso 1: Argerich

Objetivo

El objetivo del segundo trabajo práctico es resolver un problema de machine learning con los datos del TP1. Puntualmente, **se debe predecir la variable `damage_grade`** que representa el nivel de daño recibido por la edificación. El error de la solución se calculará con la métrica F1 Score, particularmente con la variación no binaria micro averaged F1 score.

$$F_{micro} = \frac{2 \cdot P_{micro} \cdot R_{micro}}{P_{micro} + R_{micro}}$$

$$P_{micro} = \frac{\sum_{k=1}^3 TP_k}{\sum_{k=1}^3 (TP_k + FP_k)}, \quad R_{micro} = \frac{\sum_{k=1}^3 TP_k}{\sum_{k=1}^3 (TP_k + FN_k)}$$

Podríamos pensar que una opción alternativa de métrica para el error sería promediar los F1 Score individuales de cada clasificación. El promedio de eso nos daría una métrica llamada macro-F1-score, que daría igual relevancia a todas los F1-score por igual sin importar su peso en la estimación.

Ej: supongamos que F1-grade-1 = 50%; F1-grade-2 = 62% y F1-grade-3 = 34%. Macro-F1 = 48,66%

La alternativa utilizada para el TP, micro-F1-score, pone más énfasis en los labels más comunes logrando de esta manera que aquellos labels poco comunes no afecten a la métrica general. Se deja el [siguiente](#) paper para aquellos interesados en leer con más profundidad al respecto.

El trabajo práctico se desarrollará en el contexto de una competencia de Driven Data. Para esto deben tener dos conjuntos de datos:

- El de entrenamiento `Train Values` con sus respectivos labels `Train Labels`, sobre el cual ajustarán su/s modelo/s junto con sus hiperparámetros.

- El de test `Test Values`, sobre el cual realizarán predicciones y subirán los resultados a Driven Data. El score estará dado por la ecuación anterior.

También contarán con un modelo de submission que deberán respetar.

Trabajo esperado

Se espera que **armen un pipeline completo** de machine learning experimentando en cada etapa. Esto quiere decir:

- Realizar *feature engineering*, encontrando codificaciones que sirvan, generando nuevos atributos, analizando cuáles atributos aportan y cuáles no.
- Probar modelos predictivos que crean que se encuadran al problema. Esto puede incluir realización de ensambles, también.
- Búsqueda de hiperparámetros óptimos de alguna manera automatizada.

Como sugerencia, recomendamos realizar una primera versión de cada etapa y que luego vayan en profundidad en cada una.

Además de los aspectos funcionales del código, también esperamos que **trabajen de manera científica**. Esto implica:

- Plantear preguntas o hipótesis y responderlas con experimentos y resultados. Las cosas no se hacen por que sí; debe haber una motivación basada en el conocimiento.
- Trabajo en equipo. Entendemos que la virtualidad y la asincronía complican la comunicación, pero en cualquier trabajo científico, en particular en los trabajos de aprendizaje automático, es fundamental la difusión de conocimiento entre las personas. Esto evita la redundancia de experimentos, genera un trabajo más cohesivo y hace que cada integrante tenga una visión más completa del problema.
- Realizar un poco de investigación del problema. En los apuntes de la materia, en las bibliotecas del lenguaje que utilizan, en internet. Esto no implica probar cosas a mansalva y sin criterio -- como mencionamos antes, debe haber una motivación o justificación.
- Interpretación de los resultados. Una vez que se realiza un experimento hay que considerar si funcionó como fue esperado o no y por qué. Los resultados negativos también son resultados. Es parte de la interpretación entender el problema en específico y encuadrar bien la solución.
- Presentación. Cualquier trabajo en ingeniería o ciencia requiere de una presentación a los colegas. Como pares de ustedes, esperamos que la comunicación que hagan del proceso y de los resultados se sustente en argumentos, datos y gráficos.

Competencia

Para el transcurso de la competencia se utilizará la plataforma provista por Driven Data (<https://www.drivendata.org/competitions/57/nepal-earthquake/>). La idea es que a medida que vayan avanzando con el trabajo y experimentando distintas cosas puedan ir subiendo sus predicciones a la plataforma. Esto les va a permitir ver su avance y lo hace más divertido.

También nos permite entender cuánto trabajaron. Si hicieron pocos envíos y sólo durante la semana anterior a la entrega, no es un buen indicio.

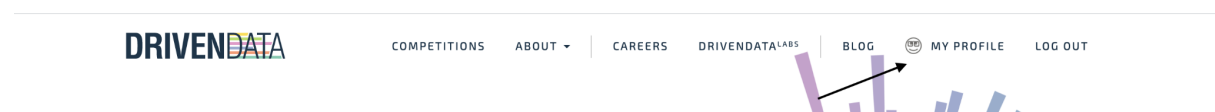
Driven Data cuenta con un límite de 3 submits diarios por equipo, tener en cuenta esto y planificar acordemente las subidas.

Se deja un [tutorial](#) provisto por la plataforma que puede servir como base para comenzar.

Pasos para darse de alta en el board privado

1- Haber seguido el [tutorial](#) provisto con anterioridad.

2- Acceder al perfil de mi usuario



3- Ingresar el código `UBA2021` y unirse

MY COMMUNITIES

Competition communities allow groups of participants to compete with each other and see a leaderboard filtered to their community. If this feature is available to you for a competition, you will have been provided a dedicated community code to enter below.

Community code

Join

Name	Code	Rank	Participants	Leaderboard
University of Buenos Aires	UBA2021	1	4	View ✖

Formato de entrega

La entrega se realiza en la última semana de julio. Tradicionalmente, se entrega un informe y luego tenemos una entrevista con cada grupo para charlar sobre el proceso, los resultados y el trabajo. Este cuatrimestre vamos a reemplazar la entrevista por un video corto de 10-15 minutos, aparte del informe.

Parte de la entrega también son los submits a Driven Data durante el trabajo. El puntaje en la competencia de Driven Data influye en la nota.

Informe

El informe se espera que sea detallado yendo a los experimentos individuales, resultados tabulados, visualizaciones y en general contando cómo fue el proceso de prueba e investigación que realizaron. Recuerden las cosas que esperamos del informe en la sección de Trabajo Esperado.

El informe debe contener un link al repositorio público para que podamos ver el trabajo que hicieron así como las contribuciones individuales.

Video

El video debe ser relativamente corto, de diez a quince minutos. No es necesaria una gran producción audiovisual; pueden grabarse presentando unas 3-4 transparencias por meet. Es importante que todos participen. Recuerden que cuando se presenta un trabajo grupal no hay un *yo*, sino un *nosotros*.

El objetivo del video es comunicarle a un/a colega suyo el trabajo que realizaron. Esto implica contar las cosas más importantes que hicieron a nivel *feature engineering*, construcción de modelos y/o ensambles y otras decisiones. Generalmente esto implica contar el mejor modelo, pero también puede haber otras conclusiones interesantes a contar aunque no fueron las que arrojaron el mejor resultado.

A continuación se dejan ejemplos de cuatrimestres pasados, la idea es que sirvan de inspiración y no como guía.

- <https://www.youtube.com/watch?v=2-FaOnBVDHc>
- <https://www.youtube.com/watch?v=-sxTeAaTgJ8>
- <https://www.youtube.com/watch?v=9H-dcwVNTcU>
- https://www.youtube.com/watch?v=olIZ_9prXD5
- https://drive.google.com/file/d/1Vw_h8fNTV6qsoyCbMFwzI8a18yp1GVSw/view