**PREWORK**

**SESIÓN 07**

**Objetivo:**

En esta sesión realizaremos una arena de clasificadores: Colocaremos varios clasificadores a competir para obtener el que mejores resultados nos da.

**Pre-requisitos (técnicos):**

Necesitaremos que tengas presente los programas que utilizamos en la sesión 02, ya que los utilizaremos mucho durante la sesión para evaluar cada uno de nuestros clasificadores.

**Pre-requisito 1: Entendiendo las métricas y rubricas de evaluación.**

Durante la sesión 02 hablamos acerca de las métricas que podemos usar para evaluar a nuestros clasificadores. Sin embargo solamente los utilizamos para evaluar un clasificador. En esta ocasión evaluaremos múltiples clasificadores, por lo cual deberás de tener en mente algunas cuantas cosas:

La sensibilidad, como recuerdas, está dada por la fórmula:

¿Pero que significa exactamente? Lo que nos dice la sensibilidad es que tan bueno es un clasificador para encontrar valores positivos (aun cuando se equivoque en ello). Y puede llegar a ser sumamente importante tener una alta sensibilidad. Por ejemplo, en casos con pruebas de COVID: Es muy importante que una prueba de COVID tenga una alta sensibilidad para ser un factor de prevención.

Imagina el caso contrario: ¿Qué pasaría si tuviera una mala sensibilidad? pacientes con el virus podrían creer que se encuentran sanos, y el contagio potencial podría ser catastrófico. Es por ello que es de suma importancia que no solamente la precisión se tome en cuenta, sino también la sensibilidad.

Por el otro lado, la especificidad se da por la fórmula:

Y nos indica que tan bueno es un clasificador para obtener los negativos. Supongamos que te encargan un clasificador que trata de hallar defectos en la línea de producción de una fábrica. Los resultados positivos son aquellos que han pasado la prueba de calidad, y los negativos son aquellos que son defectuosos.

Es realmente muy importante que encuentres los productos defectuosos. Inclusive si en el proceso se van algunos cuantos que no lo estaban, ningún producto defectuoso debe de pasar por alto. En este escenario es sumamente importante la especificidad, más que la sensibilidad.

Como puedes notar, si bien es cierto que la precisión es importante, debes de pensar muy bien cuál de las otras dos tendrá más importancia en tu solución: ¿Es importante capturar la mayor cantidad de positivos aunque algunos negativos se vayan? o bien ¿No puedes dejar escapar negativos, aunque haya un ligero desperdicio?

**Pre-requisito 2: Competencia justa.**

Recordando la sesión 02, cuando realizas pruebas debes evitar tener datos “favoritos” o fácilmente clasificables, debido a que esto acarrea la falacia de “cherry picking”. Sin embargo, también debes de ser cuidadoso cuando pruebas diferentes clasificadores, de tal manera que no favorezcas más a un clasificador que a otro, y que las pruebas que hagas salgan con algún sesgo.

Para evitar que esto suceda, lo más recomendable es que utilices los mismos datos para probar en todos los clasificadores. Esto quiere decir que:

* Separes tus datos en entrenamiento y prueba
* Entrenes todos los clasificadores con los mismos datos.
* Pruebes todos los clasificadores con los mismos datos.

Esto te permitirá generar métricas que sean fidedignas, así utilices una separación K-Fold, una separación LOOCV, o buen separes de manera aleatoria con las funciones que hemos visto en la sesión 02, de tal manera que la competencia en clasificadores siempre sea justa.

**¿Porque esto es importante?**

Supongamos que te encargan un clasificador que estará embebido en una pieza de maquinaria, la cual no puede dejar de trabajar: esto es muy común por ejemplo en la industria del vidrio, donde detener un horno significa que el horno entero podría necesitar ser reemplazado. Si tu clasificador está mal ajustado o no hiciste las pruebas correctas, puedes acarrear problemas que ameriten detener la producción de una planta.

En otro espectro, imagina que tu algoritmo estará embebido en un robot como el Curiosity: Una vez que despega el cohete rumbo a marte, no hay manera de cambiarlo. Es por eso que las pruebas son de vital importancia en el desarrollo de Machine Learning.

**Pre-requisito 3: Clasificación, Machine Learning y heurística.**

En algún punto podrías llegar a preguntarte: ¿Porque la clasificación es tan importante para Machine Learning? Y la respuesta estriba en que una clasificación es una decisión compleja. La capacidad de clasificar entre experiencias positivas y experiencias negativas son lo que te ayuda a tomar una decisión cuando una experiencia se te presenta: No sabes si será positiva o negativa, pero puedes intuir hacia qué lado se decantará dependiendo de los patrones que has aprendido a lo largo de tu vida.

De igual manera, la clasificación tiene algo que ver con un concepto llamado “heurística”, que lo puedes encontrar en muchas áreas del conocimiento humano, sin embargo, donde más nos interesa es en el área de las ingenierías: La heurística es –palabras más, palabras menos- la condición que te exige experimentar algo para saber el resultado.

Esto es sumamente importante para Machine Learning, puesto que son procesos 100% heurísticos. Para muestra un ejemplo: Tú no necesitas desglosar una multiplicación en sumas para saber que 4 x 5 = 20. Puedes predecir sin mayor problema que si sumas el número 4 cinco veces, obtendrás 20. Este es un ejemplo de un proceso no-heurístico.

Sin embargo, no hay manera de saber que se siente subirse a una montaña rusa, hasta que lo experimentas. Esto es un proceso heurístico: Debes vivirlo para conocer el resultado. Y los algoritmos de Machine Learning operan igual: Deben llevar a cabo el entrenamiento para saber si son o no útiles para resolver el problema. Es por ello que se hace tanto hincapié en el proceso de entrenamiento, validación y prueba.

**Quiz:**

**Estas construyendo un sistema que debe detectar si una aeronave tiene una falla. Si la primera característica importante es la precisión, ¿cuál es la segunda más importante?**

1. Sensibilidad
2. Falsos Positivos
3. **Especificidad**
4. Falsos Negativos

**¿Qué es lo que necesitas asegurarte de que ocurra para que una competencia sea justa en clasificadores?**

1. Que todos los clasificadores tengan los mismos parámetros
2. **Que todos los clasificadores sean entrenados y probados igualmente**
3. Que todos los clasificadores tengan el mismo tiempo de ejecución
4. Que todos los clasificadores usen la misma cantidad de memoria.

**¿Qué significa heurística?**

1. Que un sistema debe someterse a pruebas para saber si funciona bien
2. Que un sistema debe aprender de sus errores.
3. Que necesitas experimentar con varias ideas.
4. **Que necesitas llevar a cabo una idea para saber si funciona o no.**

**¿Qué pasa si un algoritmo de ML no está bien probado?**

1. **Puede fallar en escenarios reales.**
2. Tiene una baja precisión
3. Tiene un tiempo de vida muy corto
4. Puede acarrear problemas financieros.