Problema Cazatalentos, buscando a “El verdadero mejor”

La lectura de este documento es obligatoria, analiza de forma didáctica la razón por la que se produce el overfitting.

La participación en los desafíos Cazatalentos 15k y 14k es opcional, son dos desafíos distintos, y terminan cuando alguien encuentre una solución *correcta*, situación que será informada en la cartelera.
Probailísticamente es de esperar que para la noche del sábado 02-octubre ya estén resueltos ambos desafíos.

En estos desafíos suelen participar entre el 10% y el 20% de los alumnos, y dado lo muy simple de los scripts, se observan alumnos de las más diversas profesiones (en una edición casi ganó un médico psiquiatra alumno de la maestría y en otra un abogado-escribano)
La dificultad de los desafíos no está en la programación en R, sino en pensar una estrategia de torneos, de cierta forma, pensar como evitar el overfitting y encontrar el verdadero mejor modelo.

Motivación

El siguente desafío busca intruducir el problema fundamental de la Ciencia de Datos que es el overfitting desde un ángulo distinto, presentando los fenómenos del efecto del tamaño de la muestra, el conocido problema de las múltiples comparaciones y “la maldición del ganador”.
Aunque está basado en algo deportivo, se extiende fácilmente a la efectividad de accciones de marketing, a la comparación de modelos predictivos, a la elección del mejor corte en un nodo de un árbol de decisión, etc

Generalidades

Varios cazatalentos de la gran ciudad van recorriendo pueblos en búsqueda de los mejores jugadores de basket para tiros libres.
El procedimiento es el siguiente, un cazatalentos llega a un pueblo y va a la única cancha de basket del lugar donde lo están esperando los jóvenes interesados, los hace tirar tiros libres y elige al mejor jugador.
Una vez elegido el mejor jugador, se lo lleva a la gran ciudad en donde está ansioso el entrenador queriendo comprobar qué tan bueno es el jugador que le trae el cazatalentos.

A los jugadores los podemos equiparar con modelos predictivos.

Definición del Problema

¿Que algoritmo debe seguir el cazatalentos para encontrar en la menor cantidad de tiros libres totales al *verdadero mejor* , de tal forma que el 99% de las veces que se aplica el algoritmo, el único elegido por el método sea el *verdadero mejor ?*
Se le permite al algoritmo un margen de falla del 1%, que se llama False Discovery Rate
Escribir código en R que realicen estas simulaciones y permitan probar fácilmente distintas algoritmos de elección del mejor.

Cada jugador posee un "índice de enceste", que solo van a ser conocidos por nosotros, el cazatalentos no tiene forma de conocerlos; lo único que puede hacer el cazatalentos es hacer que los jugadores haga tiros libres, y medir los aciertos y no aciertos.

Primer caso, el caso trivial

El cazatalentos llega a una pequeña localidad y lo están esperando tan solo dos esperanzados adolescentes.
Uno es supertalentoso se llama Michael Jordan, nosotros sabemos que tiene un "índice de enceste" del 0.85, es decir cada vez que hace un tiro libre su probabilidad de encestar es del 85%. Si hace 100 tiros libres, en promedio encesta 85 veces, a veces puede ser más, a veces menos, pero en promedio encesta 85 de 100.
El otro es un auténtico desastre, se llama Gustavo, y su índice de enceste es del 0.10. Esto quiere decir que si hace 100 tiros libres, en promedio encesta 10.

El cazatalentos NO tiene forma de conocer los valores de 0.85 y 0.10, solo puede hacerlos tirar y contar.

El cazatalentos tiene el siguiente algoritmo, llega al pueblo, hace que Michael Jordan y Gustavo hagan 10 tiros libres cada uno, y se lleva al ganador de esos 10 tiros libres a la gran ciudad.

¿Garantiza este método que por lo menos el 99% de las veces elige a Michael Jordan?

¿El desastroso Gustavo puede llegar a encestar los 10 tiros libres? Efectivamente sí existe esa posibilidad, pero con una muy baja probabilidad, la probabilidad es de 0.10 ^10 = 1e-10
El simbolo de ^ significa "elevado a la potencia de"
¿El adolescente Michael Jordan puede llegar a encestar los 10 tiros libres? Efectivamente, su probablidad es 0.85 ^10 = 0.1968744

En el lenguaje R la instrucción `runif(10)` genera un vector con 10 números aleatorios con distribución uniforme en el intervalo [0,1], por ejemplo

```
[1] 0.2448951 0.1048925 0.1667085 0.2220606 0.4176893 0.2039536 0.4750847
[8] 0.3709089 0.9333959 0.3848157
```

la instrucción `sum( runif(10) < 0.85 )` calcula cuantos de los 10 valores son menores a 0.85, es decir, cuantos encestes hizo Michael Jordan en esos 10 tiros libres.

Ahora simulamos 10000 veces la estrategia de hacer tirar 10 tiros libres y quedarse con el ganador.

```
gustavo_ganador <- 0
for( i  in 1:10000 )
{
  aciertos_michael <-  sum( runif(10) < 0.85 )
  aciertos_gustavo <-  sum( runif(10) < 0.10 )

  if( aciertos_gustavo > aciertos_michael )  gustavo_ganador <-
gustavo_ganador +1
}

print( gustavo_ganador )
```

Lo que da la cantidad de 0, o sea que en 10000 veces, siempre ganó Michael Jordan, con lo cual la estrategia de quedarse con el mejor de 10 tiros ha funcionado, para este caso donde hay apenas dos jugadores y hay una diferencia abismal entre ellos.

En este caso NO aparece el overfitting por ningún lado

Segundo Caso, aparece el overfitting

El cazatalentos llega ahora a un pueblo donde hay 100 jugadores los que son mucho más parejos entre si desde el punto de vista del ratio de enceste.

Hay un jugador, que llamaremos jugador 1 que tiene un “índice de enceste” de 0.70
Los 99 jugadores restantes que llamaremos “el pelotón” tienen los indices de enceste { 0.501, 0.0502, 0.503, …, 0.599 }
Es decir el jugador 1 tiene 0.70 y el mejor jugador del pelotón tiene 0.599, hay un poco más de 0.10 de diferencia. La diferencia de 0.10 es significativa en basket.

¿Qué sucede con el algoritmo del cazatalentos de hacer tirar 10 tiros libres a cada uno de los jugadores y elegir al que más encestes logró ?
El resultado se puede ver en el script `basket_02.r`

Aquí ya pasa algo asombroso, el overfitting en todo su esplendor.
Si hago tirar 10 tiros libres a cada uno de los 100 jugadores, apenas 1450 veces de las 10000 este método devuelve al verdadero mejor, que es el jugador con un índice de enceste de 0.70

Lo que está sucediendo es que la inmensa mayoría de las veces, uno de los 99 jugadores del

pelotón tiene mucha suerte y supera a jugador de 0.70, con lo cual se elije a uno del pelotón y no al verdadero mejor ! El tener 99 jugadores en el pelotón hace que al ser tantos hay varios que tienen mala suerte, pero también hay varios con muy buena suerte, y el que tuvo más suerte supera al verdadero mejor !

Ahora pasemos a ver aún algo más notable.
Tenemos a nuestro jugador de 0.70, y a los 99 jugadores del pelotón. Los hacemos tirar a cada uno 10 tiros libres, elegimos al ganador registrando cuantos aciertos tuvo, y solo a ese ganador lo hacemos tirar una nueva ronda de 10 tiros libres, finalmente comparamos estos nuevos aciertos con los originales.
Este puede verse en el script `basket_03.r`

Aciertos ganador	Nueva ronda ganador
10	5
8	7
9	5
9	7
9	5
9	7
10	3
9	5
10	7
8	2

Lo que se observa en este caso es “La maldición del ganador”, la performance que el jugador que logró más aciertos en la competencia general luego NO LA PUEDE MANTENER, en todos los casos vemos que en la nueva ronda de diez tiene menos aciertos que en la primera ronda de diez.
La primera ronda el que resultó ganador fue debido a la suerte, y esa suerte ya no lo acompaña para la segunda ronda de 10 tiros libres.

Tercer Caso, el overfitting en su plenitud

¿Cuándo es más caso extremo el efecto de “La maldición del ganador” ?
Supongamos por un momento que ahora tenemos 100 jugadores, todos con un “índice de enceste” de 0.70 ; recordar que ese valor jamás es conocido por el cazatalentos.

Ahora nos ponemos más estrictos, y los hacemos tirar 100 veces a cada uno, elegimos al ganador, y solo a ese ganador lo hacemos tirar 100 nuevos tiros libres. Podemos pensarlo como que el cazatalentos se lleva al mejor jugador a la gran ciudad, le habla maravillas de él al entrenador, y el entrenador dice “probémoslo a ver si es tan bueno como decís, que haga 100 tiros libres”
`basket_04.r`

100 jugadores de 0.70	
Aciertos ganador	Nueva ronda ganador
82	68
82	78
83	78
79	64
82	69
81	73
79	71
80	69
82	61
80	73

Nuevamente vemos, que el puntaje alcanzado por el mejor jugador cuando compitió contra los otros 99, NO ES VUELTO A ALCANZAR en la nueva ronda de 100 tiros libres que hace ese jugador. Este efecto es exactamente el mismo que vemos cuando alguien se empecina en lograr en el Leaderboard Público el mayor puntaje de la clase, le va muy bien en el Público, pero se derrumba en el Privado. Decimos que “overfitió el Leaderboard Publico”.

Ahora nos preguntamos, en este caso, en promedio, cuanto más aciertos tiene el ganador (debido a la suerte) en la primera ronda con respecto a la segunda ?
Corriendo `basket_05.r` vemos que la diferencia promedio es de `11.06131`

Exactamente este mismo efecto se da en el marketing digital. Si en un A/B testing se prueban cien alternativas y se elige a la que tiene más efectividad, se observará que en un nuevo experimento disminuirá el rendimiento de esa alternativa.
También sucede si se prueban muchos medicamentos al mismo tiempo, el que resulte más efectivo disminuirá su efectividad en la nueva prueba.

Este problema es llamado “El problema de las múltiples comparaciones”

Cuarto caso, ¿cuándo se atenúa/desaparece el overfitting?

Supongamos 99 jugadores con un índice de enceste de 0.60 y ahora sumamos al adolescente Michael Jordan con su superlativo índice de enceste del 0.85
Los hacemos tirar a todos 100 tiros libres, llevamos al ganador a la gran ciudad, y allí le contamos al entrenador de nuestro ganador Jordan. `basket_06.r`

99 jugadores de 0.60 1 jugador de 0.85	
Aciertos ganador torneo	Nueva ronda solo ganador
88	91
85	86
79	91
86	81
87	87
84	84
86	83
87	86
85	77
92	83

En este caso observamos dos cosas, en primer lugar SIEMPRE el ganador del torneo fue Michael Jordan, y en segundo cuando tuvo que mostrar en la ciudad su performance, fue la misma que en el primer torneo. Es tan bueno Jordan que le ganó al que tuvo más suerte del pelotón, el efecto suerte no está afectando a Jordan porque está solo, y gana por su superioridad, no por la suerte.

Quinto caso, tamaño de la ronda

El cazatalentos llega ahora a un pueblo donde hay 100 jugadores los que son mucho más parejos entre si desde el punto de vista del ratio de enceste.

Hay un jugador, que llamaremos jugador 1 que tiene un “índice de enceste” de 0.70
Los 99 jugadores restantes que llamaremos “el pelotón” tienen los índices de enceste { 0.501, 0.0502, 0.503, …, 0.599 }
Es decir el jugador 1 tiene 0.70 y el mejor jugador del pelotón tiene 0.599, o sea hay un poco más de 0.10 de diferencia. La diferencia de 0.10 es significativa en basket.

¿Cuántos tiros libres debe el cazatalentos pedirles que haga cada jugador para que si elige el mejor jugador tiene la certeza de llevarse el “verdadero mejor” el 99% de los casos ?

script `basket_07.r`

Tiros libres	Probabilidad de elegir correctamente al verdadero mejor jugador
10	0.0327
20	0.0899
50	0.2756
100	0.5463
200	0.8618
300	0.9578
400	0.9879
415	0.9901
500	0.9970
600	0.9994
700	1
1000	1

Haciendo tirar 415 tiros libres a cada uno de los 100 jugadores, y quedándose con el ganador de ese torneo, en el 99.10% (0.9910) de los casos estoy seguro que ese es el "verdadero mejor".

Conclusiones

El overfitting aparece cuando se comparan muchos jugadores (modelos predictivos) muy parecidos entre sí. En el modelado predictivo es la situación más común que se tengan modelos con similar poder predictivo ya que generalmente solo difieren apenas en valores de los hiperparámetros o algunas columnas nuevas con feature engineering.

Los jugadores no son deterministicos, sino probabilísticos, con lo cual al comparar muchos parecidos el ganador es quien tuvo más suerte en ese torneo. El jugador ganador ganó por mera suerte, y NO puede sostener esa suerte en un nuevo torneo, por lo que el puntaje del primer torneo es mentiroso, no se sostiene en el nuevo torneo.

El overfitting se atenua o puede llegar a desaparecer del todo, si en el torneo hay un jugador ampliamente superior al resto. Ese jugador SI mantiene su performance de un torneo al otro, ya que su puntaje en el primer torneo fue alto porque es realmente bueno, *porque no es el máximo de muchos parecidos a él*.

La forma de combatir el overfitting es hacer torneos de muchos tiros libres.
Para estimar que tan bueno es el jugador ganador de un torneo, hay que hacerlo tirar nuevamente tiros libres. El secreto está en hacer más de un torneo.

Desafíos

En el Quinto Caso vimos que con 100 jugadores, haciendoles tirar 415 tiros libres a cada uno y quedándonos con el ganador, estamos realmente para este caso eligiendo al “verdadero mejor, el que tiene una probabilidad de enceste del 0.70” en el 99.01% de los casos.
Pero este método demanda 100* 415 = 41500 tiros libres, la friolera de cuarenta y un mil quinientos mil tiros libres.

Desafío 15k "el fácil"

La gran pregunta es : ¿Es posible encontrar el ganador correcto en el 99% de las veces, y que el algoritmo demande menos de 15000 tiros libres?

Dicho de otra forma, se debe hacer un script en lenguaje R : el algoritmo que debe seguir el cazatalentos para encontrar al verdadero mejor, que para el ejemplo del Quinto Caso siempre produzca lo siguiente

- Probabilidad de que el ganador elegido sea "el verdadero mejor" > 0.99 de las veces
- Cantidad Total de Tiros Libres < 15000

por favor no se asuste porque 15k es mucho menor de 41.5k , es alcanzable, textualmente "En el medio de todo el proceso de investigación pasé por varias ideas, sistemas, intentos de "optimización" a más no poder de la configuración fases-tiros... idea que se me apareció como una epifanía a las 23:30 de ayer, mientras miraba el techo de mi pieza."

El desafío 15k contribuye con 1 punto a la nota final de la materia.
Solo una persona puede ser la ganadora del desafío 15k

Desafío 14k "dificil pero siempe alguien lo resuelve"

El desafío 14k es más complicado que el desafío 15k
¿Es posible encontrar el ganador correcto en el 99% de las veces, y que el algoritmo demande menos de 14000 tiros libres?

Dicho de otra forma, se debe hacer un script en lenguaje R : el algoritmo que debe seguir el cazatalentos para encontrar al verdadero mejor, que para el ejemplo del Quinto Caso siempre produzca lo siguiente

- Probabilidad de que el ganador elegido sea "el verdadero mejor" > 0.99 de las veces
- Cantidad Total de Tiros Libres < 14000

El desafío 14k contribuye con 2 puntos a la nota final de la materia, que es mucho !
Solo una persona puede ser la ganadora del desafío 14k

Aclaraciones

- Si usted es familiar de un alumno que ya participó en este desafio, está excluido de participar de la misma.
- Todas las veces que se lanzó estos desafios a alumnos, antes de los tres días ambos estuvieron resueltos.
- Estos desafíos son un problema que necesita de una muy buena idea; no es un ejercicio donde ya se sabe el camino a seguir.
- Lo más probable es que primero alguien resuelva correctamente el desafío 15k, esa solución se publique en la cartelera y recién luego inspirándose en esa solución alguien resuelva el desafío 14k.
- Si alguien resuelve el desafío 14k de entrada sin que nadie haya resuelto aún el deafío 15k, se terminan las dos competencias y la persona ganadora se lleva DOS puntos en la nota (aclaro que **no** se lleva tres puntos)
- Hay veces que los scripts enviados son muy confusos, y lleva un tiempo determinar si es correcto. En ese caso, se considera ganador a la persona que primero presentó el script
- Los scripts deben ser enviados por mensaje *privado* de Zulip a los profesores Alejandro y Gustavo.
- Esta es una tarea individual, son libres de comentar en Zulip lo que deseen, compartir, pero las entregas se consideran individuales, y el premio de la nota va solo a la persona que gana el desafío.
- Tener ganado ya un punto o dos de nota, alivia enormemente la presión del resbaloso y traicionero Leaderboard Privado; teansitan la materia desde un lugar completamente diferente.
- Se han subido los scripts `intento_A_01.r` `intento_A_02.r` e `intento_B.r` a simplemente modo de orientación, estos scripts utilizan el concepto de rondas. Se aclara que tambien existen otras soluciones alternativas a las ensayadas aquí.

Bibliografía

https://www.youtube.com/watch?v=FpCrY7x5nEE
https://www.youtube.com/watch?v=42QuXLucH3Q
https://www.youtube.com/watch?v=6ZxIzVjV1DE

https://en.wikipedia.org/wiki/Replication_crisis

https://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/308269/

https://www.nature.com/news/over-half-of-psychology-studies-fail-reproducibility-test-1.18248