HENRY

Evaluación de modelos II

Data Science

Agenda

- Validación Cruzada
- K-fold Cross Validation
- Validación Cruzada Aleatoria
- Curvas de Validación
- Optimización de Hiperparámetros
- Random Search
- Descenso de Gradiente

OBJETIVOS DE LA CLASE

Al finalizar esta lecture estarás en la capacidad de...

- ➜ Comprender técnicas de Optimización de Entrenamiento de Modelos

Al **finalizar** cada uno de los temas, tendremos un **espacio de consultas**.

Hay un **mentor** asignado para responder el **Q&A**.

¡Pregunta, pregunta, pregunta! :D

Validación **cruzada**

¿Cómo?

¿Cómo podemos evaluar si el modelo está aprendiendo o no de nuestros datos?

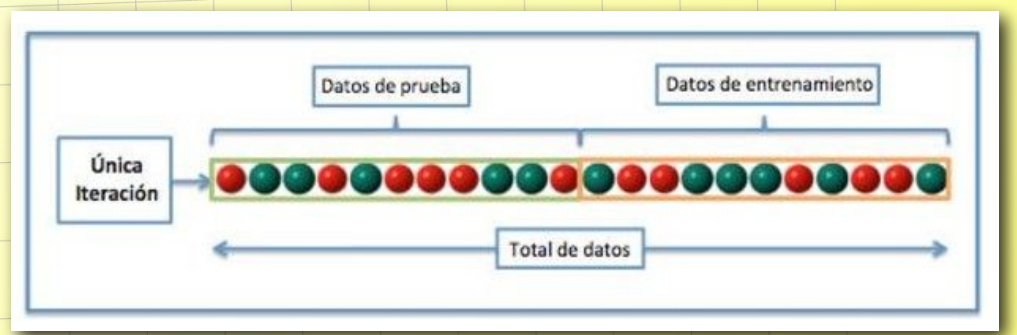
Una forma práctica de hacerlo es observar su desempeño frente a nuevas instancias.

¿Cómo?

En nuestro flujo de trabajo, tendremos que emular una situación donde el modelo es entrenado con ciertos datos y luego es evaluado con datos nuevos.

Train Test Split:

- ➜ Separo los datos en dos conjuntos, Train y Test.
- ➜ Entreno con los datos de Train
- ➜ Evalúo el desempeño con los datos de Test.



Beneficios

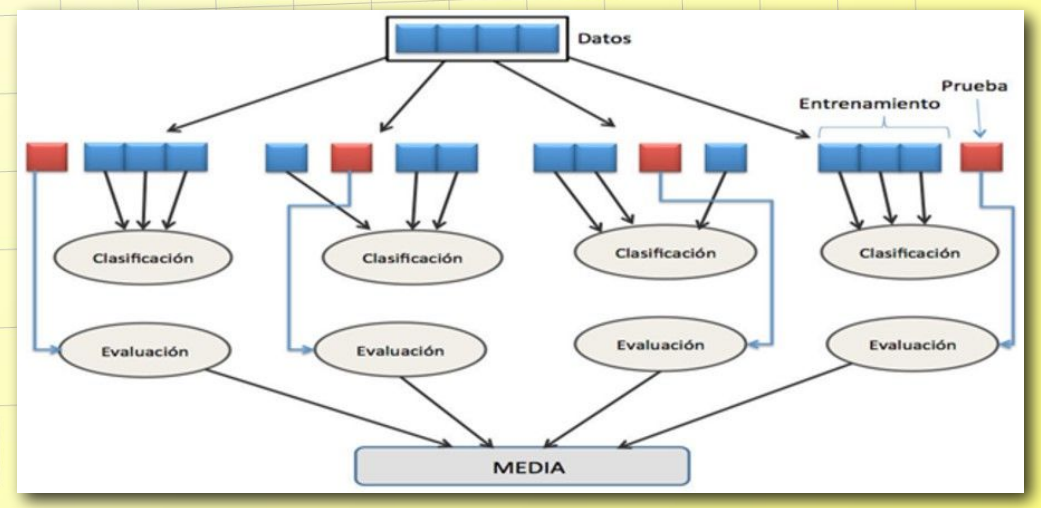
Evaluar el desempeño del sobreajuste de Test tiene varios usos:

- ➜ Obtenemos una evaluación realista del desempeño de nuestros modelos.
- ➜ Nos permite seleccionar el modelo que mejor desempeña sobre nuestros datos.

Objetivo

El objetivo de la validación cruzada es obtener una evaluación de performance de nuestro modelo que sea independiente de la partición en entrenamiento y prueba de los datos.

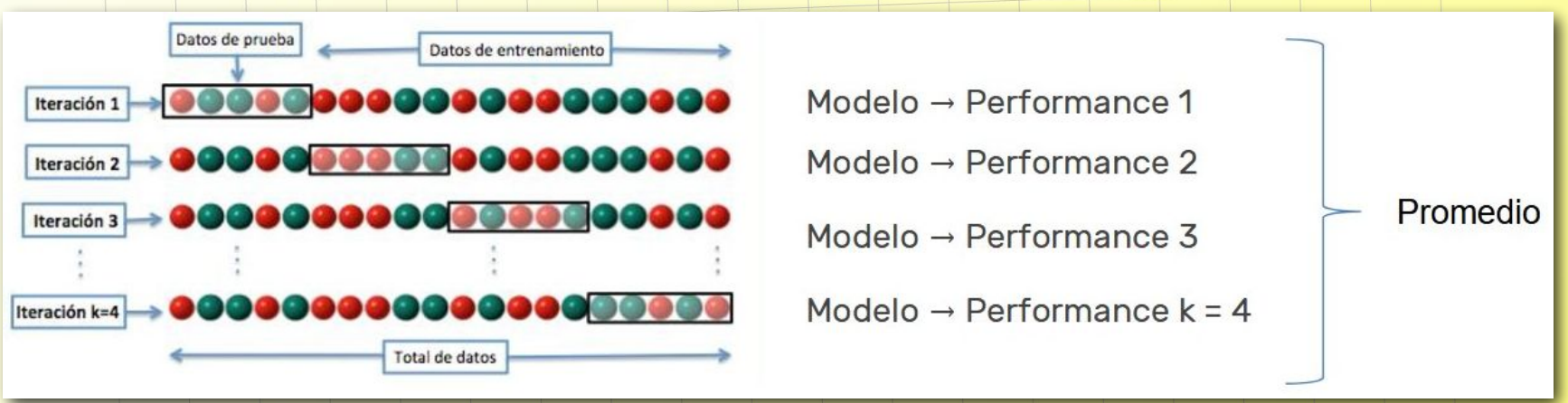
Haciendo muchas particiones esperamos que la medida de performance sea independiente de la partición de los datos.



K-fold Cross Validation

¿Cómo?

Es importante notar, que cada dato aparece una sola vez en los datos de prueba y k-1 en los datos de entrenamiento.



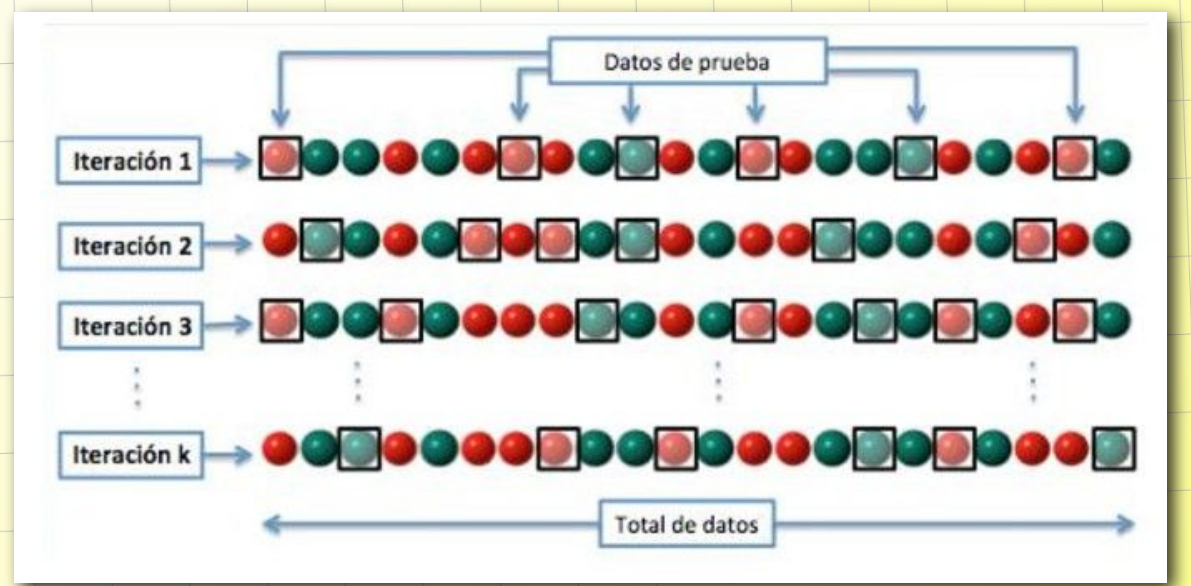
Conclusiones

- La validación cruzada es un procedimiento de remuestreo que se utiliza para evaluar modelos de aprendizaje automático en una muestra de datos limitada.
- El hiperparámetro más importante es k que se refiere al número de grupos en que se dividirá una muestra de datos dada.
- La validación cruzada está íntimamente relacionada con la optimización de hiperparámetros.

validación cruzada **aleatoria**

Evaluación de modelos

En este caso, cada dato puede aparecer más de una vez en el conjunto de prueba.

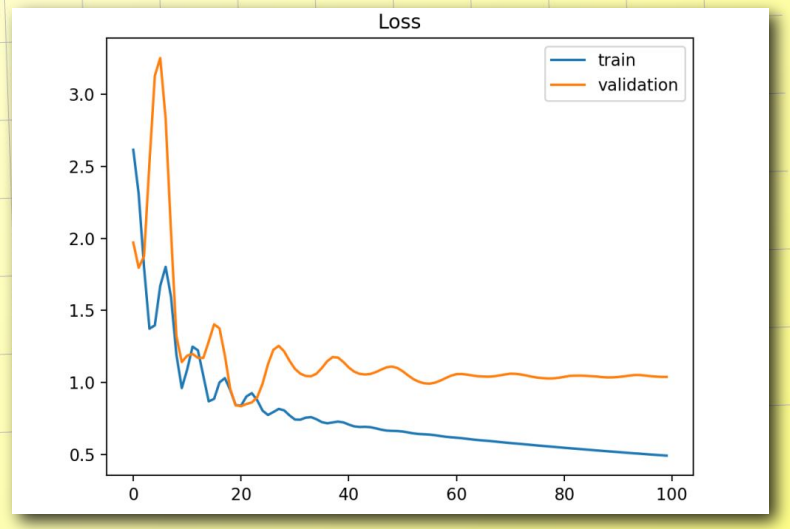


Curvas de validación

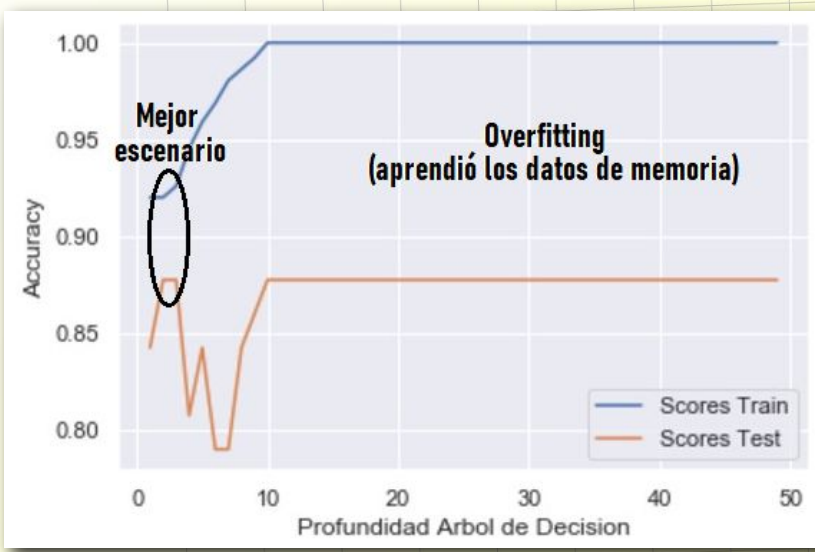
¿Para qué?

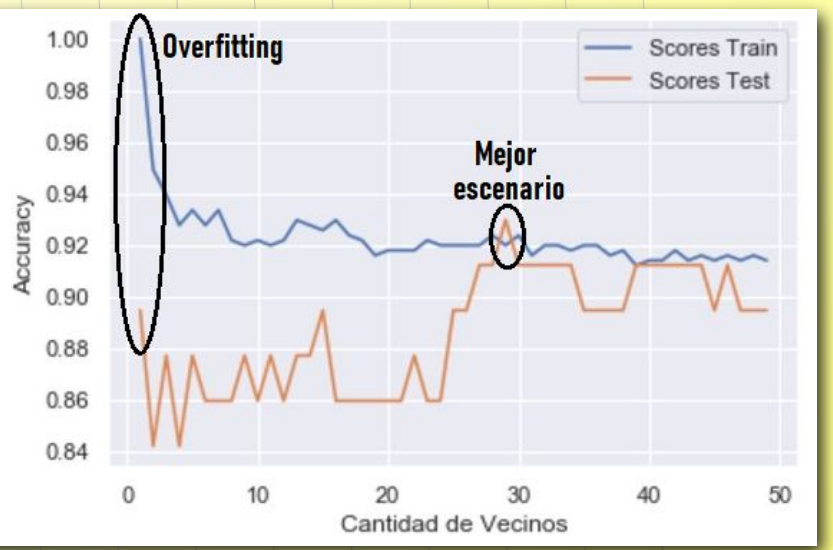
En general, el desempeño de un modelo depende de muchos hiperparámetros. Pero a veces hay uno que es el más importante, el que predomina sobre el resto.

Para elegir el valor de ese hiperparámetro - y también caracterizar mejor el desempeño de nuestro modelo -, es útil obtener las curvas de validación.



Ejemplos


Mejor
escenario
Overfitting
(aprendió los datos de memoria)
Accuracy
1.00
0.95
0.90
0.85
0.80
0
10
20
30
40
50
Profundidad Arbol de Decision
Scores Train
Scores Test


Overfitting
Mejor
escenario
Accuracy
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0
10
20
30
40
50
Cantidad de Vecinos
Scores Train
Scores Test

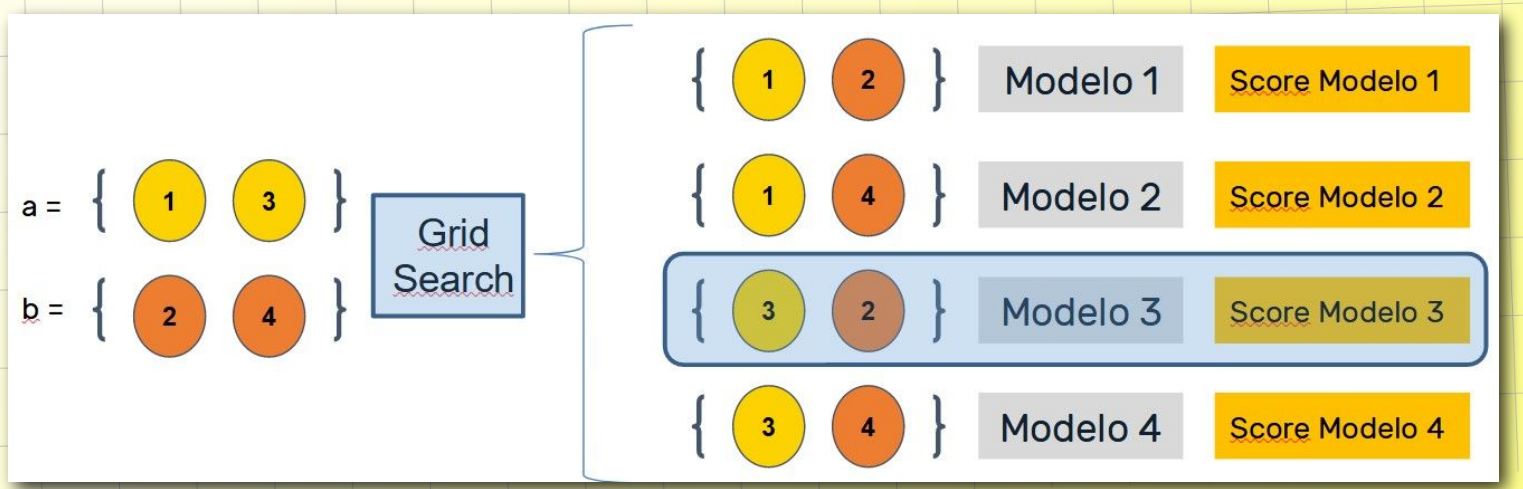
optimización de hiperparámetros

¿Cómo?

- ¿Cómo elegimos los mejores hiperparámetros para nuestro problema?
- ¿Qué es mejor, exactitud, precisión o exhaustividad? ¿Área bajo la curva ROC?
- Primero, se debe definir una métrica a optimizar. Una vez que se sabe cuál métrica optimizar, hay que probar los distintos valores de hiperparámetros.
- Se debe hacer una búsqueda exhaustiva. Es decir probando con todos los valores de los hiperparámetros que podamos y eligiendo la mejor combinación. Este método se llama Grid Search ("búsqueda de cuadrícula").

Ejemplo Grid Search

Si tenemos dos hiperparámetros, a y b, que pueden tomar valores a = {1,3} y b = {2,4}



Ejemplo Grid Search

Pasos:

1. Elegir los valores que puede tomar cada hiperparámetro.
2. Armar las combinaciones “todos con todos” → Armar la grilla.
3. Recorrer la grilla entrenando el modelo para cada combinación y evaluarlo.
4. Elegir los hiperparámetros que definen el mejor modelo.

Random Search

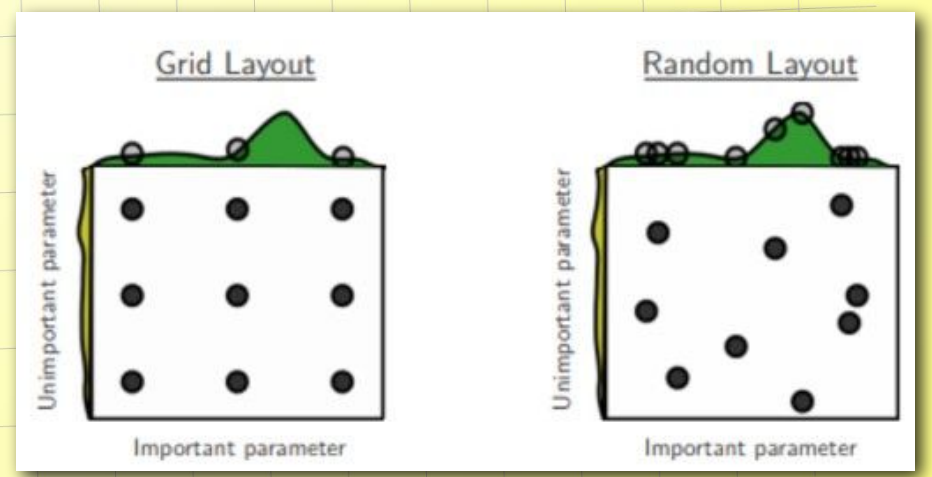
Si por ejemplo, se tienen cinco hiperparámetros y cinco valores para probar por hiperparámetro, el tamaño de la grilla comienza a crecer.

Además, para cada modelo se debe hacer la Validación Cruzada. Este proceso puede ser computacionalmente muy demandante.

Random Search

Random Search explora opciones y combinaciones al azar, de manera menos “ordenada”.

En muchas circunstancias, esto es más eficiente, tanto desde el punto de vista de performance del modelo como de desempeño computacional.



Conclusiones

- Es necesario definir una métrica a optimizar (exactitud, precisión, RMSE, ROC AUC, etc.).
- Un modelo (regresor o clasificador).
- Una grilla de hiperparámetros. Depende del tipo de modelo utilizado.
- Un método para buscar o muestrear los candidatos:
  - Grid Search: Plantea opciones y explora todas las combinaciones.
  - Random Search: explora opciones y combinaciones al azar.
- Crear un modelo lo antes posible, en cualquier caso, un modelo fallido muchas veces da tanta información sobre el proceso real como uno válido

¿PREGUNTAS?

Teorema de Bayes

Dados dos eventos A y B:

- $P(A)$ es la probabilidad del evento A y $P(B)$ es la probabilidad del evento B
- $P(A|B)$ es la probabilidad condicional del evento A dado que ocurrió B y $P(B|A)$ es la probabilidad condicional del evento B dado que ocurrió A (No implican causalidad)
- Si $P(A|B) = P(A)$ y $P(B|A) = P(B)$, los eventos son independientes.

En general, $P(A|B) \neq P(B|A)$. Para obtener uno dado el otro, necesitamos el Teorema de Bayes:

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$$

Teorema de Bayes: ejemplo

¿Cuál es la probabilidad de que una persona tenga una enfermedad si el examen dio positivo?

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$$

P(A|B): posterior o probabilidad a posteriori
P(E+|T+): probabilidad de estar enfermo dado que el test dio positivo.
0.99*0.001/0.01098=

0.09

P(B|A): verosimilitud
P(T+|E+): probabilidad de que el test de positivo dado que la persona está enferma. ¡Es la probabilidad de detección! **= 0.99**

P(A): prior o probabilidad a priori de A
P(E+): El prior es la prevalencia de la enfermedad en la población **= 1/1000**

P(B): probabilidad marginal.
P(T+). La probabilidad de que el test dé positivo. Esto puede ocurrir si una persona está enferma pero también si no lo está. **= P(T+|E+)*P(E+) + P(T+|E-)*P(E-) = 99/100*0.001 + 999/99900*0.999 = 0.01098**

- A: estar enfermo **E+**
- B: dio positivo **T+**

Naive Bayes

- Este modelo está basado en el Teorema de Bayes con un supuesto de independencia entre los predictores.
- Naive Bayes supone que la presencia de una característica particular en una clase no está relacionada con la presencia de ninguna otra característica.
- El modelo Naive Bayes es fácil de construir y particularmente útil para conjuntos de datos muy grandes.

Naive Bayes: ejemplo

Una fruta puede considerarse una manzana si es roja, redonda y de aprox. 3" de diámetro.

Incluso si estas características dependen unas de otras o de la existencia de otras características, todas estas propiedades contribuyen independientemente a la probabilidad de que esta fruta sea una manzana y por eso se conoce como "ingenua".

Naive Bayes: ejemplo

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Tomando como ejemplo un conjunto de datos sobre el clima y la variable "Jugar". Debemos clasificar si los jugadores jugarán o no según las condiciones climáticas.

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Naive Bayes

Ventajas	Desventajas
Es fácil de entender y rápido.	Suposición de variables independientes. En la vida real es casi imposible que obtengamos un conjunto de variables completamente independientes.
Funciona bien en la predicción de clases múltiples.	Para las variables numéricas, supone una distribución Normal o Gaussiana, lo que implica un supuesto fuerte.
Cuando se asume la independencia, un clasificador Naive Bayes funciona mejor en comparación con otros modelos como la regresión logística y necesita menos datos de entrenamiento.	
Funciona bien en el caso de variables de entrada categóricas.	

¿Alguien dijo Homework?

HENRY

Próxima lecture

Series de tiempo

¡Feedback!

Click on me

Dispones de un formulario en:

- Homeworks
- Guías de clase
- Slack


HENRY