HENRY

Evaluación de modelos I

Data Science

Agenda

- Balanceo de dataset
- Evaluación de modelos de clasificación
- Entrenamiento y test
- Matriz de Confusión
- Curva ROC
- Teorema de Bayes y Naive Bayes
- Underfitting y overfitting
- Sesgo y varianza
- Parámetros/hiperparámetros

OBJETIVOS DE LA CLASE

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

- ➜ Comprender los conceptos de Balanceo de dataset, overfitting y underfitting
- ➜ Entender cómo es la Evaluación de modelos de clasificación
- ➜ Conocer el concepto Matriz de Confusión
- ➜ Aplicar el entrenamiento con conjuntos de validación y test
- ➜ Diferenciar los conceptos de Sesgo y Varianza
- ➜ Diferenciar los conceptos de Parámetros e Hiperparámetros

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

Balanceo de datasets

Datasets desbalanceados

- En determinadas ocasiones, nos enfrentaremos a datasets que están desbalanceados.
- Esto significa que habrá una prevalencia de una clase por sobre otra.
- La realidad es que un poco de desbalance de clases es de esperar y no afecta a nuestro análisis.

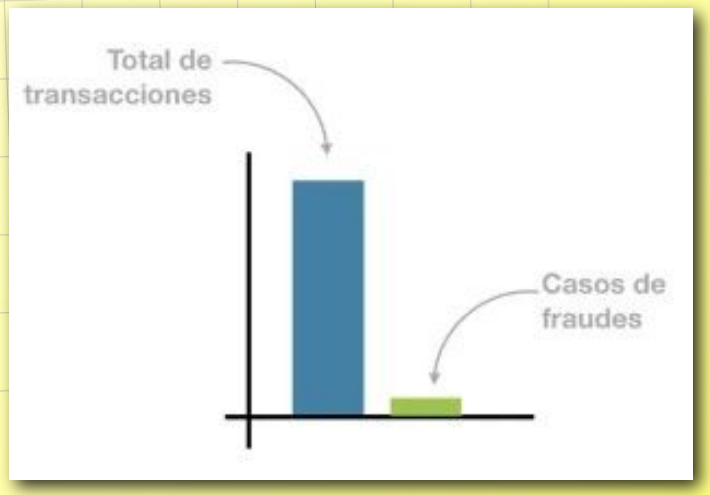
Datasets desbalanceados

En un dataset desbalanceado encontramos muchas instancias de una clase y muy pocas de la otra, dificultando el entrenamiento.

Por ejemplo, en el caso binario, 90:10, 99:1, y peor.

Bajo ciertas problemáticas, suelen haber datasets muy desbalanceados:

- ➜ Detección de fraudes
- ➜ Diagnóstico médico
- ➜ Falla en cadena de producción



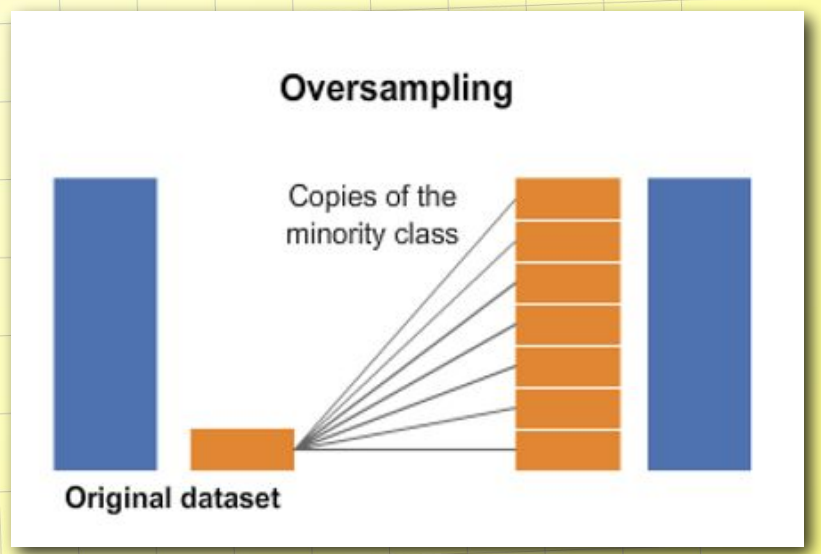
Balanceo de datasets

Ante un dataset desbalanceado, se pueden tomar diversas acciones:

- ➜ Revisar la posibilidad de conseguir nuevos datos.
- ➜ Utilizar otras métricas: exactitud, Matriz de Confusión, Precisión, Exhaustividad (lo desarrollaremos más adelante).
- ➜ Hacerle al dataset un remuestreo: Oversampling, Undersampling.
- ➜ Probar diferentes modelos (modelos de ensamble) y/o agregarle peso a la clase subrepresentada.

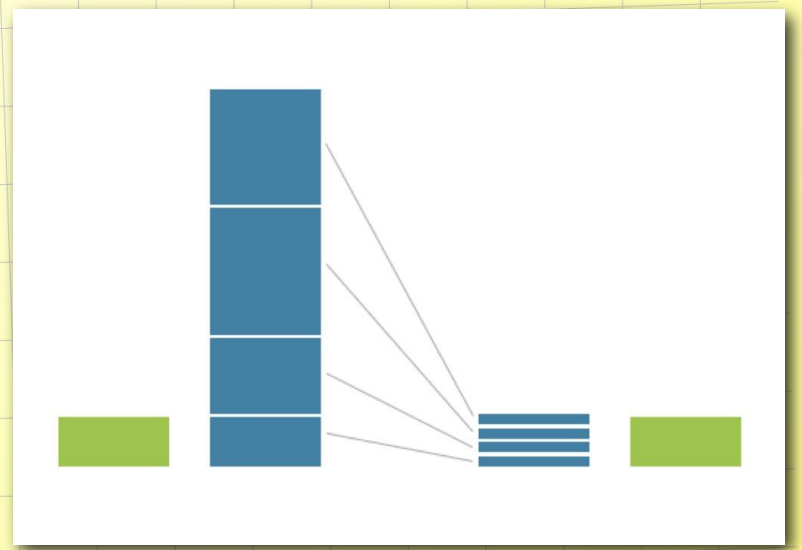
Oversampling

Se realiza un sobremuestreo de la clase minoritaria. En caso que no podamos obtenerlos de alguna fuente externa, se soluciona simplemente copiando registros que corresponden a esa categoría en nuestro propio dataset.



Undersampling

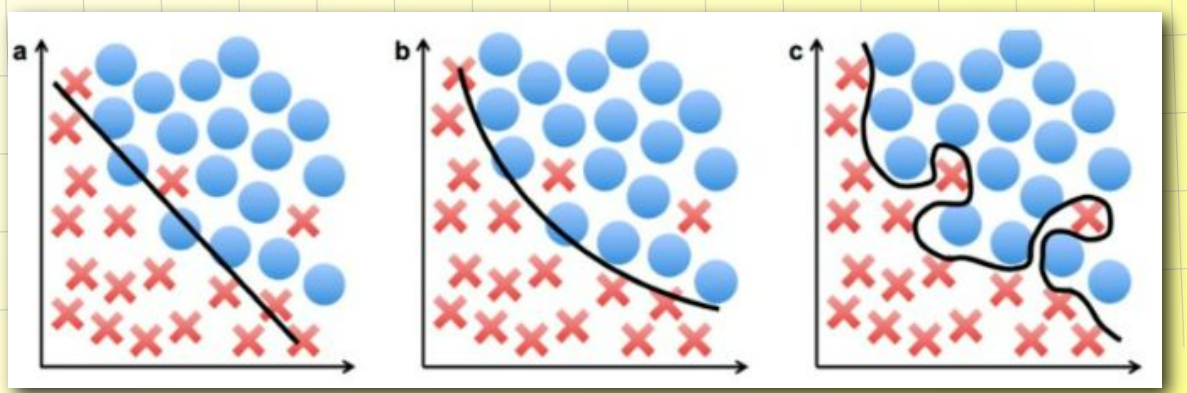
Se realiza un submuestreo de la clase mayoritaria. Es decir, eliminamos registros de nuestro dataset que contengan como etiqueta o variable de salida aquella clase hegemónica o predominante.



Evaluación de modelos

Evaluación de modelos

- La intención es evaluar si el modelo está aprendiendo o no de los datos.
- Una manera de hacerlo es verificar su desempeño frente a nuevas instancias.
- Pero, ¿por qué necesitamos nuevas instancias y no usamos, simplemente, las instancias que usamos para entrenar?



Entrenamiento y test

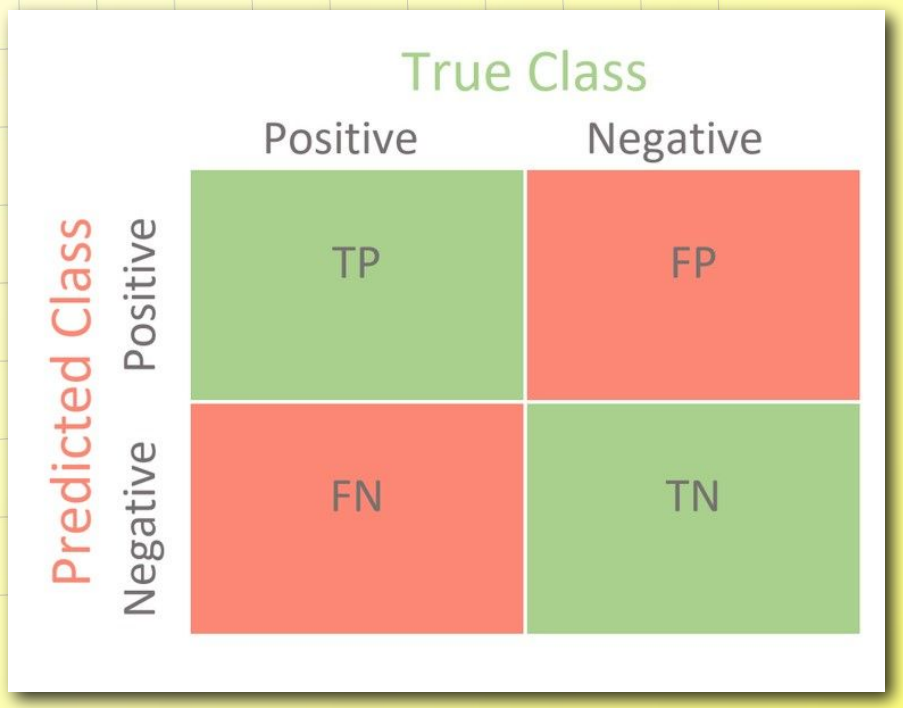
- Se separa una porción de los datos: En ocasiones esta separación no es al azar, sino que tiene un cierto criterio, esto depende del problema.
- Se evalúa el desempeño del modelo sobre los datos de entrenamiento.
- Luego, se evalúa sobre los datos que restan, que van a oficiar de esos datos que el modelo "nunca vio" y son nuevos.
- En todos los entornos de desarrollo de Machine Learning existe una función que hace la tarea de separación de los datos. En Scikit-Learn, la función se llama train_test_split().

Matriz de confusión

Esta matriz es para una clasificación binaria.

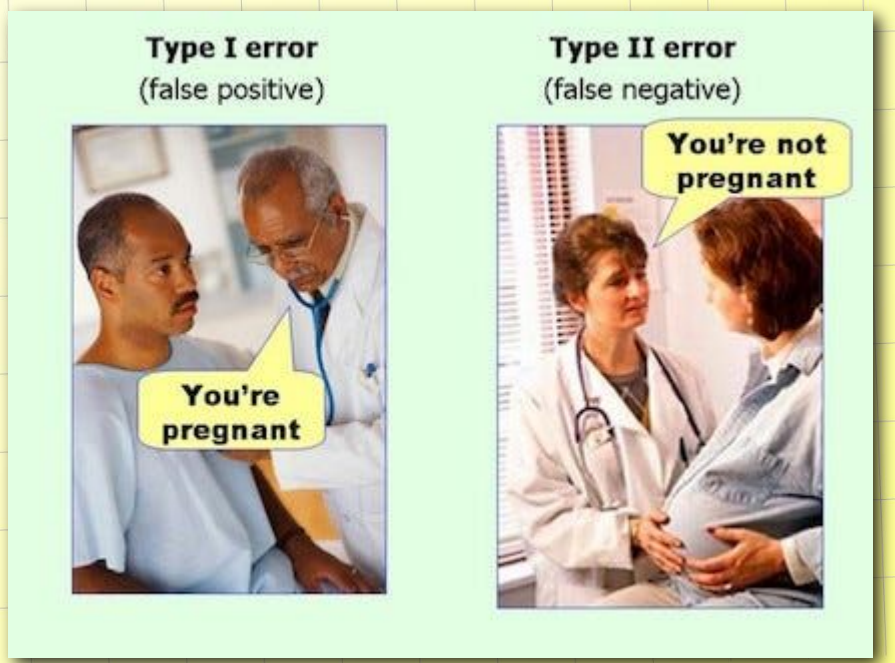
En el eje y tenemos las etiquetas predichas, mientras que en el eje x las etiquetas reales.

- ➜ Verde: etiquetas predichas correctamente (valor real y predicho coinciden).
- ➜ Rojo: etiquetas que el modelo clasificó erróneamente.

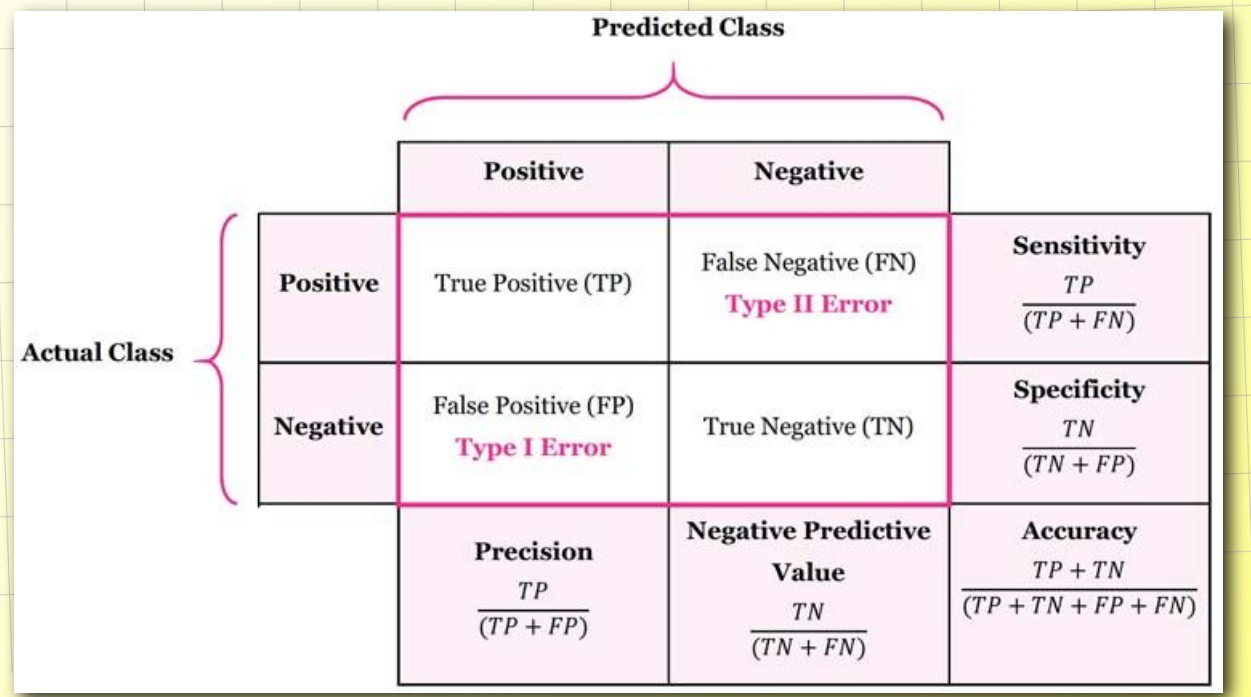


Matriz de confusión

¿Falso positivo
o
falso negativo?



Matriz de confusión: métricas



Matriz de confusión: métricas

- Precisión: resultados correctos sobre el total de muestras seleccionadas (verdaderos positivos sobre los verdaderos positivos + los falsos positivos). Indica cuánto acertó el modelo dentro de todo el universo.
- Exhaustividad (sensibilidad o recall): resultados correctos por sobre todos los resultados que buscamos identificar (verdaderos positivos sobre los verdaderos positivos + falsos negativos).
- Especificidad: indica los verdaderos negativos.

Matriz de confusión: métricas

- F1 score: combina precisión y exhaustividad de forma tal de mantener una relación entre las dos (no aumenta mucho una en detrimento de la otra).
  Si la precisión o el recall son bajos, también lo será F-Score. Mientras que, si las dos son altas -cercanas a 1-, también lo será F-Score.

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

Matriz de confusión: ejemplo

Un examen médico tiene una probabilidad de detección de 0.99 y una probabilidad de Falso Positivo de 0.01.
El objetivo del Test es detectar una enfermedad de relativa baja prevalencia, que solo la tiene una persona en mil.

Luego de hacer el examen a 100000 personas y completar la matriz de confusión, ¿Cuál es la probabilidad de que una persona tenga la enfermedad si el examen dio positivo?

Predicha / Verdadera	Positivos	Negativos
Positivos	99	999
Negativos	1	98901

Matriz de confusión: ejemplo

Predicha / Verdadera	Positivos	Negativos
Positivos	Verdaderos Positivos (TP)	Falsos Positivos (FP)
Negativos	Falsos Negativos (FN)	Verdaderos Negativos (TN)

- **Exactitud** = TP + TN / (TP + TN + FP + FN)
- **Precisión** = TP / (TP + FP)
- **Exhaustividad** = TP / (TP + FN)

- **F1-Score** = 2 * Precisión * Exhaustividad / (Precisión + Exhaustividad)

Matriz de confusión: ejemplo

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

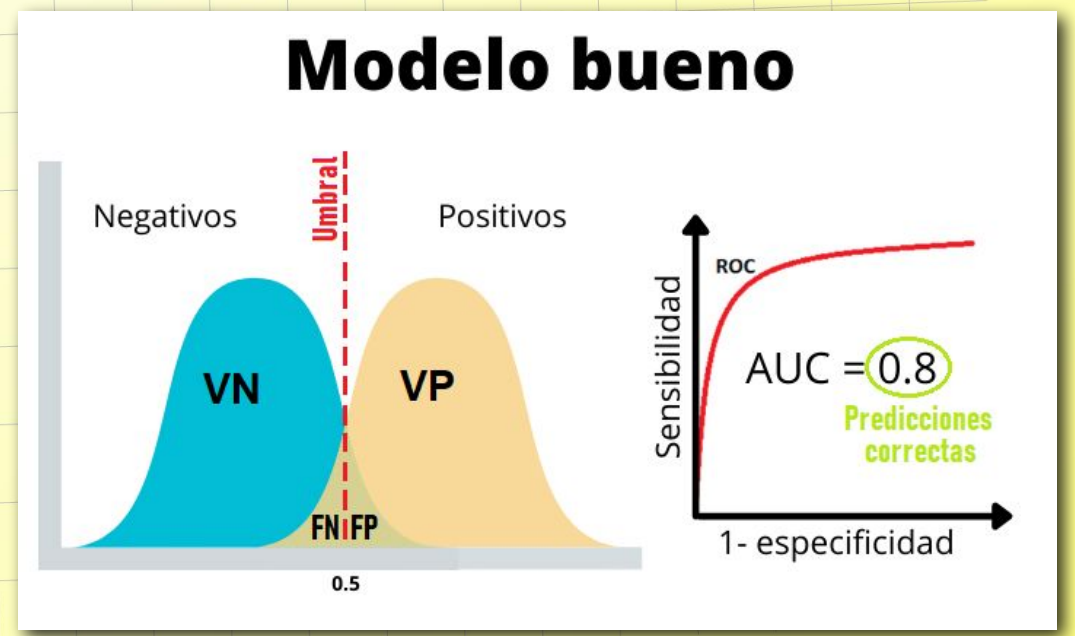
Predicha / Verdadera	Positivos	Negativos
Positivos	99	999
Negativos	1	98901

De 1098 predichos, 99 eran verdaderos positivos…
Es decir, ~9% (o probabilidad = 0.0902).
Esa es la **precisión** de la clase

De 100 realmente positivos, 99 fueron predichos… Es decir 99%
Esa es la **exhaustividad** de la clase

Curva ROC

La curva ROC es una representación gráfica de la relación entre las tasas de falso positivo (FPR) y las tasas de verdadero positivo (TPR).



Curva ROC

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

True Positive Rate → Es la **Exhaustividad**

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

False Positive Rate → Es la **Especificidad**

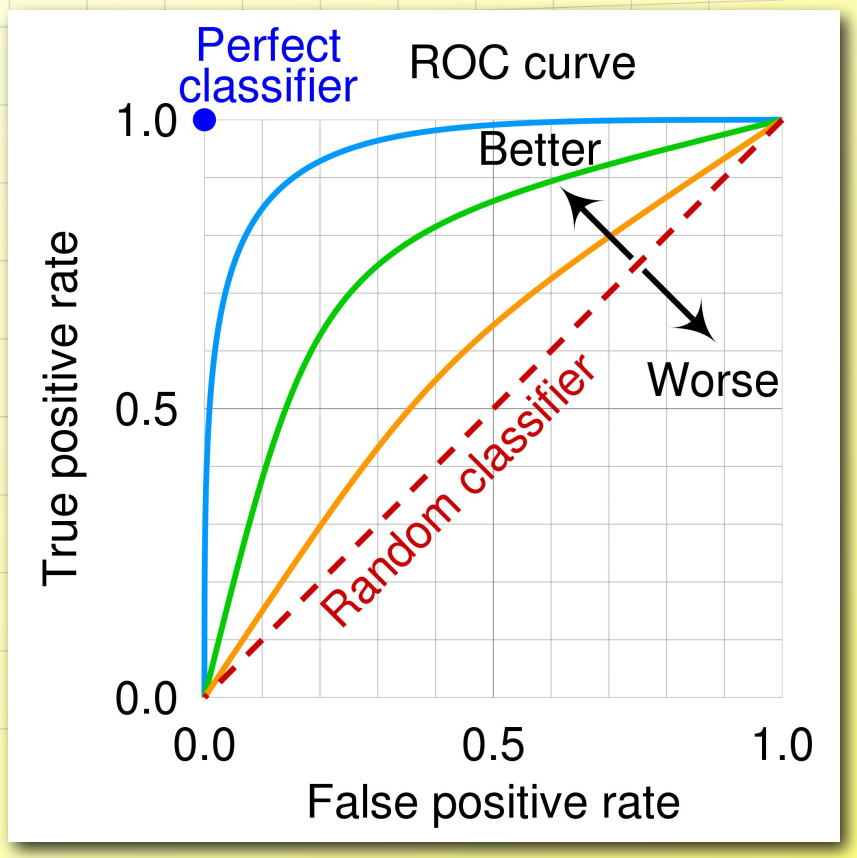
TPR : describe qué tan bueno es el modelo para predecir la clase positiva cuando el resultado real es positivo.

FPR : (tasa de falsas alarmas) resume la frecuencia con la que se predice una clase positiva cuando el resultado real es negativo.

Curva ROC

El área bajo la curva indica la probabilidad de que el modelo sea capaz de distinguir entre una clase y la otra.

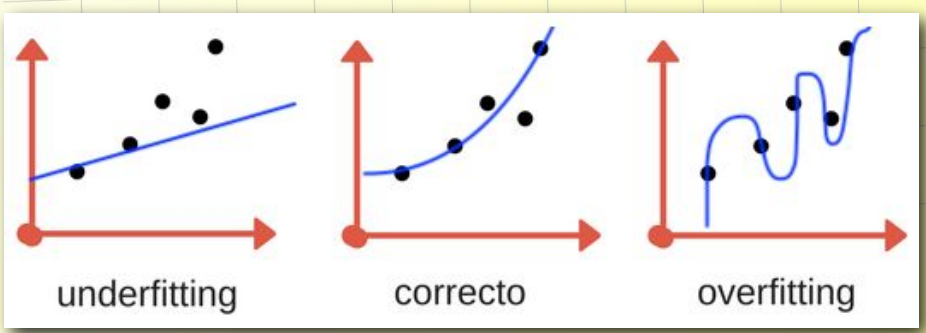
- ➜ Un modelo excelente tendrá un AUC = 1.
- ➜ El el azar tendrá un AUC = 0.5
- ➜ Con un AUC = 0 tendremos un modelo que clasifica todas las etiquetas al revés.



Underfitting Vs. Overfitting

Underfitting:

- El modelo establece una línea divisoria demasiado generalista.
- Tendrá un bajo desempeño para hacer una predicción tanto con los datos de muestreo como los poblacionales.



Overfitting:

- El modelo está demasiado ajustado a los datos.
- El modelo se aprendió los datos de memoria, pero no aprendió a generalizar.
- Cuando se da el sobreajuste, el modelo va a dar buenos resultados con los datos de entrenamiento, pero no va a funcionar tan bien para los datos nuevos que no haya visto.

Sesgo y varianza

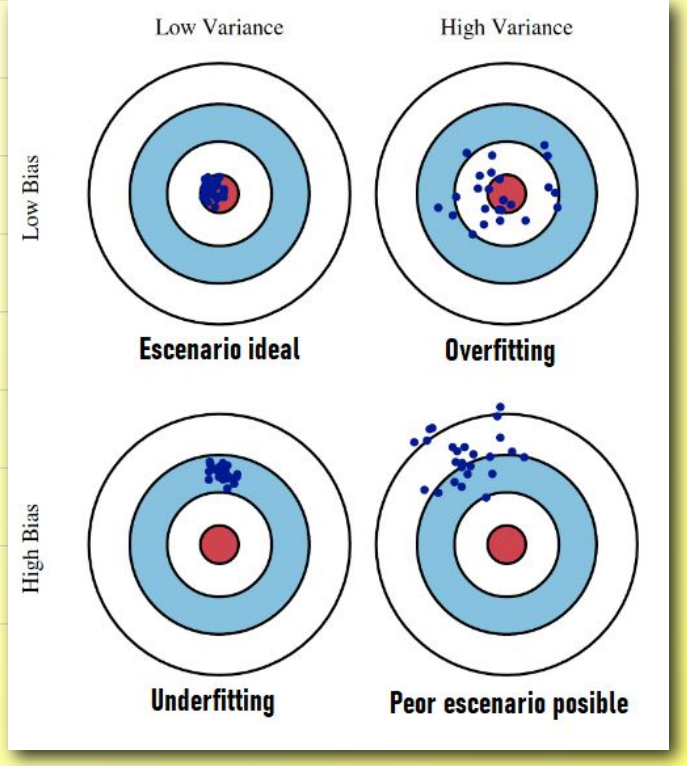
Sesgo:

- Lo introducimos al intentar explicar un problema al que le correspondería un modelo complejo con uno simple.
- Asociado al underfitting.

Varianza:

- Es la cantidad en la que cambiaría la predicción si hubiera entrenado al modelo con otros datos.
- Asociado al overfitting.

4 escenarios posibles



Parámetros Vs. Hiperparámetros

Parámetros:

- ➜ son las variables que se estiman en el proceso de entrenamiento.
- ➜ Definen cómo usar los datos de entrada para obtener la salida deseada.
- ➜ Se aprenden en el momento que se realiza el entrenamiento.
- ➜ Los valores de los parámetros no los indica manualmente el data scientist, sino que son obtenidos. Es decir, los ajusta el modelo por sí solo.

Hiperparámetros:

- ➜ Es el parámetro cuyo valor se define antes de que el modelo comience a entrenarse.
- ➜ Se definen para controlar el proceso de aprendizaje.
- ➜ Se utilizan para mejorar el aprendizaje del modelo.

Parámetros Vs. Hiperparámetros: Ejemplos

Parámetros:

- ➜ Ordenada al origen y pendiente en regresión lineal.
- ➜ Valor del índice de Gini en una hoja en árbol de decisión.

Hiperparámetros:

- ➜ Cantidad de vecinos en KNN.
- ➜ Profundidad del árbol en árbol de decisión.

¿PREGUNTAS?

¿Alguien dijo Homework?

HENRY

Próxima lecture

Evaluación de modelos II

¡Feedback!

Click on me

Dispones de un formulario en:

- Homeworks
- Guías de clase
- Slack


HENRY