Evaluación del modelo de ML

¿Cómo podemos evaluar si nuestro modelo está aprendiendo correctamente de nuestros datos?

Una respuesta posible sería, que para evaluar si nuestro modelo aprendió o no de nuestros datos, observemos su desempeño o performance frente a nuevas instancias, es decir, frente a datos que nunca vio.
A partir de este concepto, surgen las siguientes características en ML: “Entrenamiento” y “Validación” para luego hablar del “Sobreajuste” o “Sub-ajuste.

Aprendizaje o Entrenamiento: Proceso en el que se detectan los patrones de un conjunto de datos, es decir, es el corazón del Machine Learning. Cuando identificamos los patrones, se pueden hacer predicciones con nuevos datos que se incorporen al sistema.

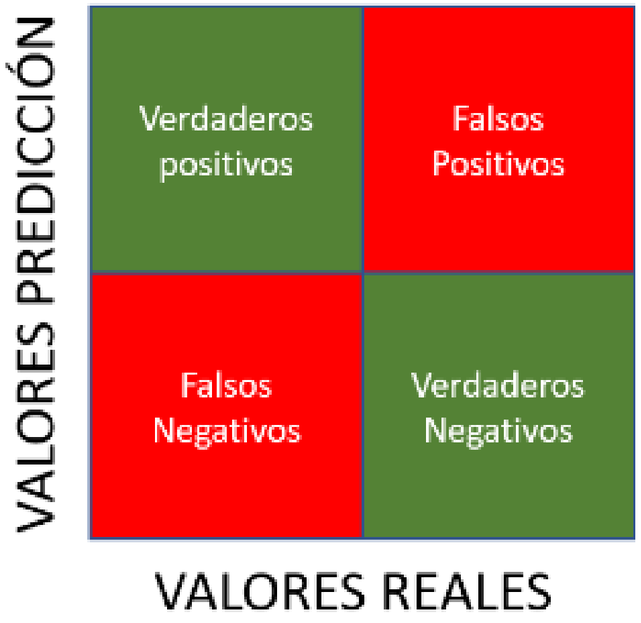
Validación: Proceso de evaluar un modelo entrenado sobre un conjunto de datos de prueba. Para poder evaluarlo correctamente, hay que realizar “split de datos” es decir, separar nuestro dataset original en “Datos de Entrenamiento”, que serán usados justamente para entrenar a nuestro modelo y en “Datos de Test o de Testing” que serán aquellos datos que utilizaremos para evaluar la performance de nuestro modelo.

¿Qué porcentaje se usa para train y test?: No existe una única respuesta, en términos generales se suele utilizar un 70 % de nuestros datos para el training y un 30 % para el testing.

by Noelia Ferrero

Métricas que existen dentro del Machine Learning para evaluar la performance de nuestro modelo.

- **Matriz de Confusión:** Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. En términos prácticos entonces, nos permite ver qué tipos de aciertos y errores está teniendo nuestro modelo.



Interpretación:

- Verdadero Positivo (TP): Predije que era positivo y lo era.
- Verdadero Negativo (TN): Predije que era falso y lo era.
- Falso Positivo (FP): Predije que era positivo, pero resultó ser negativo.
- Falso Negativo (FN): Predije que era negativo, pero resultó siendo positivo.

Los Verdaderos Positivos como Negativos son aciertos. Los Falsos Negativos como Positivos son errores.

by Noelia Ferrero

Terminología muy importante:

Ratio	Nombre
FP/N	False positive rate, Probabilidad de FALSA ALARMA
TN/N	True negative rate, **Especificidad**
TP/P	True positive rate, **Sensibilidad**
FN/P	False Negative rate, **MISS RATE**
TP/P*	Positive predictive rate, **Precision**
FP/P*	False Discovery rate
TN/N*	Negative predictive value
FN/N*	False omission rate
P/n	Prevalence
(TP/TN)/n	Accuracy

by Noelia Ferrero

Matriz de confusión y sus métricas

Exactitud: se refiere a lo cerca que está el resultado de una medición del valor verdadero. En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación. Se representa por la proporción entre los positivos reales predichos por el algoritmo y todos los casos positivos. En forma práctica la Exactitud es el % total de elementos clasificados correctamente. (VP+VN)/(VP+FP+FN+VN) * 100

Precisión: (Positive Predictive rate) Se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión. Es una proporción entre el número de predicciones correctas (tanto positivas como negativas) y el total de predicciones. En forma práctica, es el porcentaje de casos positivos detectados y nos sirve para medir la calidad del modelo de ML en tareas de clasificación. Se calcula como: VP/(VP+FP)

by Noelia Ferrero

Sensibilidad o Tasa de Verdaderos Positivos: Es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo. En términos prácticos sería la capacidad de una prueba para identificar correctamente a las personas con la característica (ej. enfermedad). Se calcula: VP/(VP+FN) o lo que sería igual en términos de salud: Verdaderos positivos.

Especificidad - Tasa de Verdaderos Negativos: Se trata de los casos negativos que el algoritmo ha clasificado correctamente. Expresa cuán bien puede el modelo detectar esa clase. En términos prácticos es la capacidad de la prueba para identificar correctamente a las personas sin la característica (e.g enfermedad).Se calcula: VN/(VN+FP) o en términos de salud: Verdaderos Negativos

F1 - Score: Esta es otra métrica muy empleada porque nos resume la Precisión (Precisión) y Sensibilidad (Recall) en una sola métrica. Es una medida general del desempeño de un modelo combinando Precisión y Sensibilidad. Un valor alto indica pocos Falsos Positivos y pocos Falsos Negativos. Los valores típicos están entre 0 y 1.
Se calcula: 2 * (Recall * Precisión) / (Recall + Precisión)

by Noelia Ferrero

En resumen

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$

$$\text{Especificity} = \frac{\text{True Negatives}}{\text{True Negatives + False Positives}}$$

by Noelia Ferrero

- **Error Cuadrático Medio** (RMSE, por sus siglas en inglés, Root Mean Squared Error): Es la métrica más comúnmente utilizada para las tareas de regresión y representa a la raíz cuadrada de la distancia cuadrada promedio entre el valor real y el valor pronosticado. Indica el ajuste absoluto del modelo a los datos, cuán cerca están los puntos de datos observados de los valores predichos del modelo.
- **Error Absoluto Medio** (MAE, Mean Absolute Error): Es la diferencia absoluta entre el valor objetivo y el valor predicho por el modelo. Es más robusto para los valores atípicos, sin embargo, este tipo de métrica, no es adecuada para aplicaciones en las que desea prestar más atención a los valores atípicos.
- **R-Cuadrado:** indica la bondad o la aptitud del modelo, a menudo se utiliza con fines descriptivos y muestra que tan bien las variables independientes seleccionadas explican la variabilidad en sus variables dependientes. R-cuadrado tiene la propiedad útil de que su escala es intuitiva, va de 0 a 1, con 0 indicando que el modelo propuesto no mejora la predicción sobre el modelo medido y 1 indica una predicción perfecta.

Existen varias métricas más, por ejemplo, el R cuadrado ajustado (R^2), MSPE – Error de porcentaje cuadrático medio, entre otras.

by Noelia Ferrero

Overfitting y Underfitting

- Las principales causantes de obtener malos resultados en Machine Learning son el **Overfitting** o el **Underfitting** de los datos. Dado que cuando entrenamos nuestro modelo intentamos “hacer encajar” - fit - los datos de entrada entre ellos y con la salida.
- Tanto el Over como el Under – Fitting, se relacionan al fallo de nuestro modelo al generalizar -encajar- el conocimiento que pretendíamos que adquieran.

¿Cómo prevenir el Overfitting? Sucede cuando nuestro modelo aprende los datos de train perfectamente, por lo que no es capaz de generalizar y cuando le lleguen nuevos datos obtiene pésimos resultados.

Existen diferentes formas de prevenir el Overfitting:

- Dividir nuestros datos en training, validación y testing.
- Obtener un mayor número de datos.
- Ajustar los parámetros de nuestros modelos.
- Utilizar modelos más simples en caso de ser posible

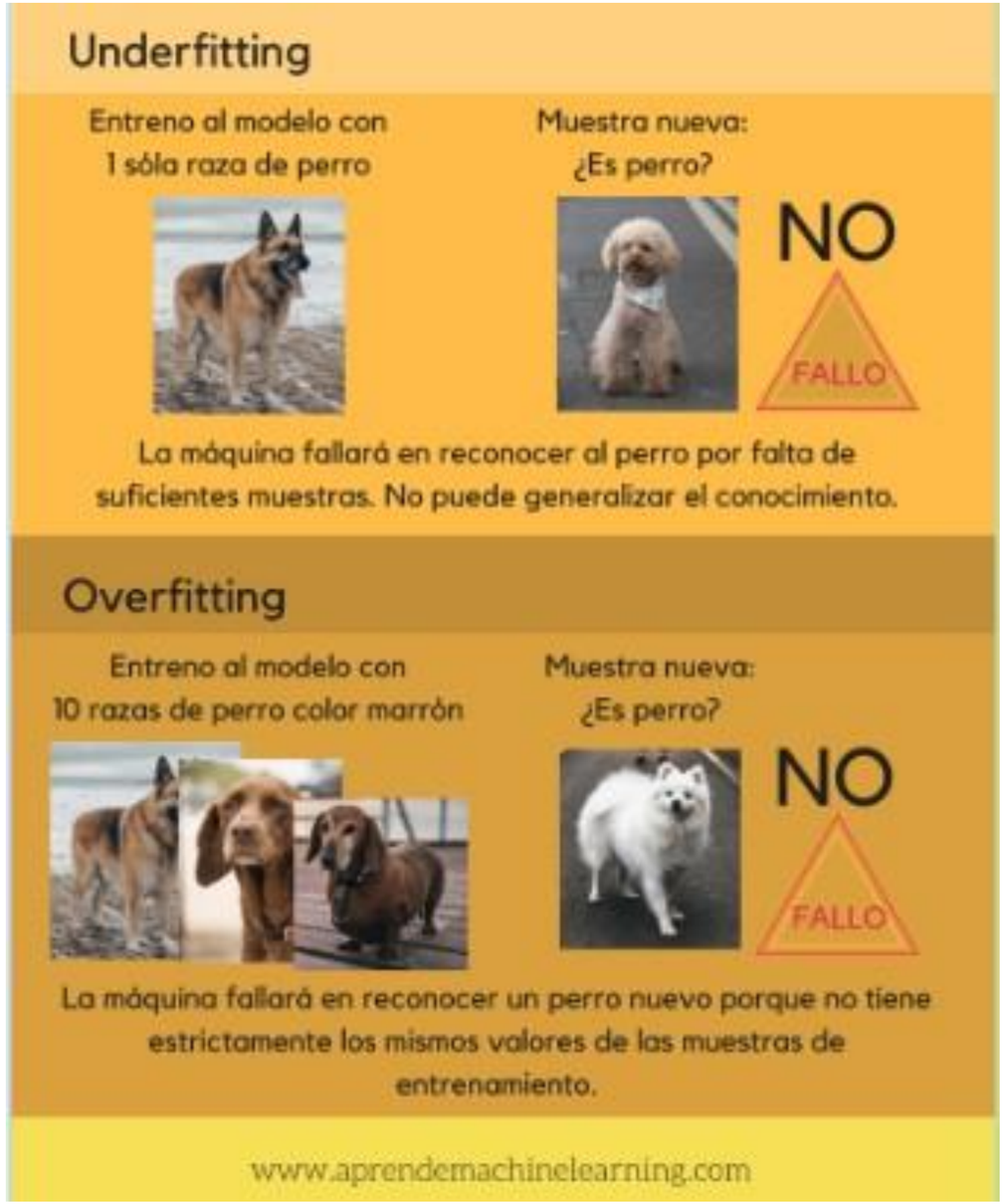
¿Cómo prevenir el Underfitting? Sucede cuando nuestro modelo no es capaz de identificar patrones. Por lo que obtendrá siempre pésimos resultados.

Existen diferentes formas de prevenir el Underfitting:

- Tratar los datos correctamente, eliminando outliers y variables innecesarias.
- Utilizar modelos más complejos.
- Ajustar los parámetros de nuestros modelos.

by Noelia Ferrero

Ejemplo



by Noelia Ferrero