Curso de Especialización de
Inteligencia Artificial y Big Data (IABD)

Sistemas de aprendizaje Automático

UD05. Técnicas avanzadas y evaluación del modelo. Resumen.

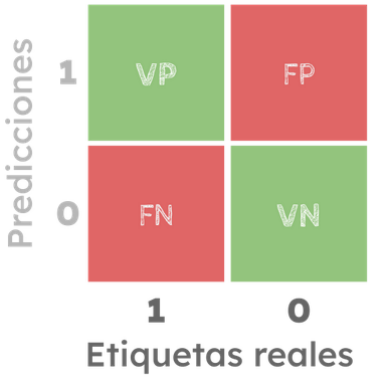
JUAN ANTONIO GARCIA MUELAS

Dentro del aprendizaje automático supervisado, es decir, los modelos en los que tenemos los casos "etiquetados" o con su variable de salida correcta correspondiente, hay dos tipos de problemas: clasificación y regresión.

Evaluación en problemas de clasificación

Matriz de confusión

Ampliamente utilizada, permite inspeccionar y evaluar visualmente las predicciones de nuestro modelo.



1	2	2
0	1	15
	1	0

Las **métricas cuyos valores nos indican la calidad del** modelo son:

- ✓ **Exactitud o accuracy**: la **fracción de predicciones que** el modelo realizó **correcta**mente. Se representa como un porcentaje o un **valor entre 0 y 1**. Es una **buena métrica** cuando tenemos un **conjunto de datos balanceado**, esto es, cuando el número de etiquetas de cada clase es similar. **Dividiendo el número de casos correctamente clasificados de todas las clases por el número total de casos.**
- ✓ **Recall o sensibilidad**: indica la **proporción de ejemplos positivos** que están **identificados correctamente** por el modelo **entre todos los positivos reales**. Es decir, VP / (VP + FN). En nuestro ejemplo, el valor de sensibilidad sería 2 / (2 + 1) = 0,67. **Dividiendo el número de casos de una clase identificados correctamente por todos los casos reales de dicha clase.**
- ✓ **Precisión**: esta métrica está determinada por la **fracción** de elementos **clasificados correctamente como positivo entre todos los** que el modelo ha clasificado como **positivos.** La fórmula es VP / (VP + FP). El modelo de ejemplo tendría una precisión de 2 / (2 + 2) = 0.5. **Dividiendo el número de casos de una clase identificados correctamente por todos los casos de esa clase identificados por el modelo.**
- ✓ **F1 score**: **combina** las métricas **Precision y Recall** para dar un único resultado. Esta métrica es la más **apropiada cuando tenemos conjuntos de datos no balanceados**. Se calcula como la **media armónica de Precisión y Recall**. La fórmula es F1 = (2 * precision * recall) / (precision + recall).

Evaluación en problemas de regresión.

En los modelos de regresión es casi imposible predecir el valor exacto, más bien se **busca estar lo más cerca posible del valor real**, por lo que la mayoría de las métricas, con sutiles diferencias entre ellas, van a centrarse en medir eso: lo cerca (o lejos) que están las predicciones de los valores reales.

Algunas de las **métricas de evaluación más comunes** para los **modelos de regresión** son:

- ✓ **Error medio absoluto**: Es la **media de las diferencias absolutas entre el valor objetivo y el predicho**. Al no elevar al cuadrado, **no penaliza los errores grandes**, lo que la hace **no muy sensible a valores anómalos**, por lo que no es una métrica recomendable en modelos en los que se deba prestar atención a éstos. Esta métrica también representa el error en la misma escala que los valores reales. Lo más **deseable** es que su valor sea **cercano a cero**.
- ✓ **Media de los errores al cuadrado (error cuadrático medio)**: Una de las medidas más utilizadas en tareas de regresión. Es simplemente la **media de las diferencias entre el valor objetivo y el predicho al cuadrado.** Al elevar al cuadrado los errores, **magnifica los errores grandes**, por lo que hay que utilizarla con cuidado cuando tenemos valores anómalos en nuestro conjunto de datos. Puede tomar **valores entre 0 e infinito.** Cuanto más cerca de cero esté la métrica, mejor.
- ✓ **Raíz cuadrada de la media del error al cuadrado**: Es igual a la raíz cuadrada de la métrica anterior. La ventaja de esta métrica es que **presenta el error en las mismas unidades que la variable objetivo**, lo que la hace más **fácil de entender**.
- ✓ **R cuadrado**: también llamado **coeficiente de determinación.** Esta métrica difiere de las anteriores, ya que **compara nuestro modelo con un modelo básico que** siempre **devuelve** como predicción **la media de los valores objetivo de entrenamiento**. La **comparación** entre estos dos modelos se realiza **en base** a la media de los errores **al cuadrado de cada modelo**. Los **valores** que puede tomar esta métrica van **desde menos infinito a 1**. **Cuanto más cercano a 1** sea el valor de esta métrica, **mejor** será nuestro modelo.
- ✓ **R cuadrado ajustado:** es una mejora de R cuadrado. El problema de la métrica anterior es que cada vez que se añaden más variables independientes (o variables predictoras) al modelo, R cuadrado se queda igual o mejora, pero nunca empeora, lo que puede llegar a confundirnos, ya que, porque un modelo utilice más variables predictoras que otro, no quiere decir que sea mejor. **R cuadrado ajustado compensa la adición de variables independientes. El valor** de R cuadrado ajustado **siempre va a ser menor o igual al de R cuadrado,** pero esta métrica mostrará mejoría cuando el modelo sea realmente mejor.

A la hora de trabajar con algoritmos de aprendizaje supervisado es muy importante la elección de una métrica de evaluación correcta para nuestro modelo. Para los **modelos de clasificación** es muy importante prestar **atención al conjunto de datos y** comprobar **si es balanceado o no.** En los **modelos de regresión** hay que **considerar los valores anómalos y si queremos penalizar errores grandes o no.**

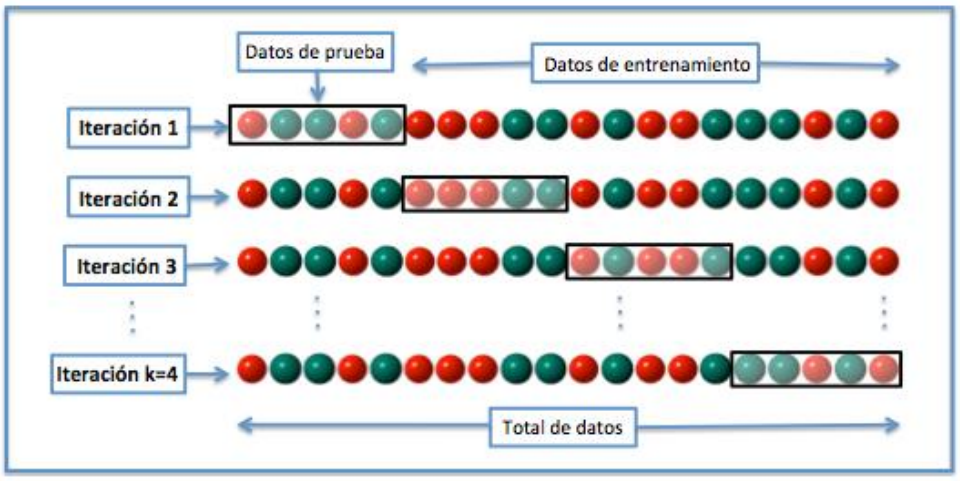
Validación cruzada

Consiste en **repetir y calcular la media aritmética** obtenida de las medidas de evaluación sobre diferentes particiones del conjunto de datos.

La principal desventaja de la validación cruzada dejando uno fuera es que es un proceso muy costoso computacionalmente.

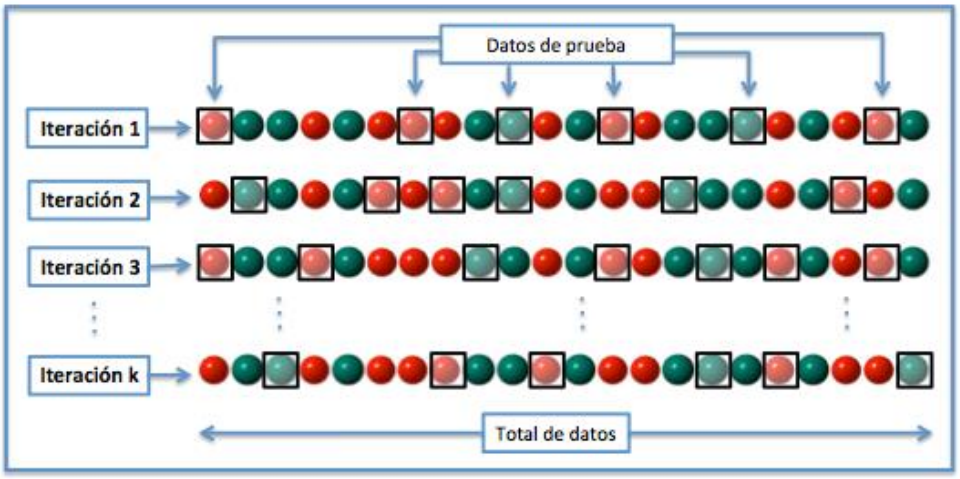
Validación cruzada de K iteraciones

En la validación cruzada de K iteraciones o K-fold cross-validation **los datos** con los que contamos **se dividen en K subconjuntos**. **Lo más común es utilizar** la validación cruzada de **10 iteraciones** (10-fold cross-validation). El inconveniente es que este proceso **es lento**.



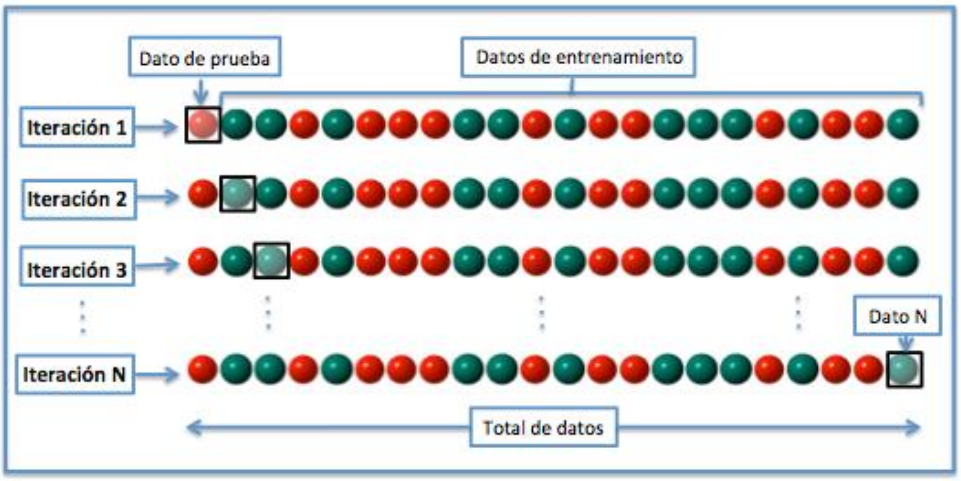
Validación cruzada aleatoria

En este segundo enfoque, el **bloque de validación se escoge aleatoriamente,** repitiéndose el proceso k veces (siendo tanto el tamaño de cada bloque como el número k cifras arbitrarias).



Validación cruzada dejando uno fuera

En este tercer tipo de validación cruzada, **cada registro** u observación **será la muestra para validar o para test, dejando** todo el **resto** de muestras (N-1) **para el entrenamiento**.



Optimización de hiperparámetros.

La optimización de hiper-parámetros se realiza normalmente mediante la utilización de un **proceso de búsqueda** cuyo **objetivo** consiste en **encontrar la mejor selección de valores para un conjunto finito de hiper-parámetros** con el objetivo de **generar el mejor modelo posible.**

Se suele utilizar la métrica de **accuracy para guiar el proceso.**

Son los que debemos elegir y fijar para que la arquitectura esté definida (Ej: el valor K en KNN), y su valor se utiliza para controlar el proceso de aprendizaje.

Búsqueda aleatoria (Random Search)

Es un proceso de **búsqueda** de tipo **aleatorio sobre** un **espacio** de búsqueda **finito**. **Se modifican aleatoriamente las soluciones previamente generadas en el espacio de búsqueda.**

Búsqueda en cuadrícula (Grid Search)

La búsqueda en cuadrícula (Grid Search) es un proceso de búsqueda donde los diferentes valores de **hiper-parámetros se combinan** para crear una maya (grid) donde se **incluyen todas las posibles combinaciones** de parámetros distribuidos de manera uniforme.

En redes neuronales, no es posible aplicar este tipo de optimización.

Dataset no balanceados.

¿En qué sectores encontramos problemas de clasificación muy desbalanceados?

- ✓ En el estudio de un funnel de marketing.
- ✓ En el área de la salud.
- ✓ En el ámbito del fraude.

Podemos tener casos donde alguna de las categorías sean **clases "minoritarias"** con pocas muestras. Esto, **provoca un desbalanceo** en los datos de entrenamiento.

Cuando tenemos un dataset con desequilibrio, suele ocurrir que **obtenemos** un **alto valor** de **precisión en** la **clase Mayoritaria y** un **bajo recall en** la clase **Minoritaria**.

En el aprendizaje automático, el **submuestreo y el sobremuestreo son dos técnicas que se ocupan de los desequilibrios** en un conjunto de entrenamiento (la parte de los datos utilizada para ajustar un modelo). Se puede **submuestrear la clase mayoritaria** (lo más eficaz como norma general), **sobremuestrear la clase minoritaria** o combinar las dos técnicas.

Submuestreo (Under-sampling)

El submuestreo implica la **selección** (incluso de forma aleatoria) de **ejemplos** de la **clase mayoritaria para eliminarlos del conjunto de** datos de entrenamiento.

Sobremuestreo (Over-sampling)

Al igual que el submuestreo, también se puede optar por un sobremuestreo aleatorio. En este caso, **se aumentan los casos** de la clase minoritaria, **de forma sintética**, **hasta alcanzar una escala similar** al volumen de la clase mayoritaria.

La ventaja es que no hay riesgo de perder información, pero la desventaja es que este método, **con volúmenes de datos muy grandes, penaliza el entrenamiento y aumenta el riesgo de overfitting**.

Para este método, se **suele utilizar el algoritmo SMOTE**.

Otras estrategias

- ✓ Algunos modelos **admiten parámetros** para contrarrestar el efecto de este desequilibrio, como el peso en árboles de decisión o el parámetro **class_weight** en la regresión logística.
- ✓ Aplicar **técnicas "Ensemble" como el algoritmo Random Forest**, que entrena un cierto número de modelos y el resultado final se "vota". O también, por ejemplo, en un XGBoost, aumentando el número de árboles, podemos ir corrigiendo los errores de los árboles anteriores.
- ✓ Usar **algoritmos de Stacking y algoritmos de aprendizaje por refuerzo:** del mismo modo que los Boosting, estos algoritmos **permiten ir mejorando los aciertos** de la clase minoritaria.

Detección de anomalías.

La detección de anomalías (o detección atípica) es la **identificación de elementos raros**, eventos u observaciones que generan sospechas al diferenciarse significativamente de la mayoría de los datos. Normalmente, los datos anómalos **se pueden conectar a** algún tipo de problema o evento raro como, por ejemplo, **fraude bancario, problemas médicos, defectos estructurales, equipo defectuoso, etc.**

Reducción de dimensionalidad con PCA y Autoencoders

La **detección de anomalías (outliers) con PCA y Autoencoders es una estrategia no supervisada** para identificar anomalías cuando los **datos no** están **etiquetados**, es decir, no se conoce la clasificación real (anomalía - no anomalía) de las observaciones.

Isolation Forest

Otro algoritmo típico para la detección de anomalías es el Isolation Forest. La lógica que sigue es diferente a otros métodos conocidos y gira en torno a la idea de que los **puntos anormales** dentro de los conjuntos de datos **son más fáciles de separar (isolate) que los puntos normales**.