[AFO02848T] IFCD093PO MACHINE LEARNING APLICADO USANDO PYTHON [MOD02447L] IFCD093PO MACHINE LEARNING APLICADO USANDO PYTHON [UDI1502C6] Glosario

Glosario

Α

**AUC-Curva ROC:** AUC — ROC: es una de las métricas de rendimiento más importantes utilizadas para comprobar el rendimiento del modelo. AUC — ROC se utiliza para la clasificación binaria y también multiclase, pero se utiliza principalmente en problemas de clasificación binaria. En este blog, consideraremos una clasificación de clase binaria.

В

**Bias:** El sesgo en el aprendizaje automático se define como el fenómeno de observar resultados que se ven perjudicados sistemáticamente debido a suposiciones erróneas. Sin embargo, sin suposiciones, un algoritmo no tendría mejor desempeño en una tarea que si el resultado se eligiera al azar, un principio que fue formalizado por Wolpert en 1996 en lo que llamamos el teorema de No Free Lunch.

**BI:** business intelligence (BI, inteligencia empresarial). Técnicas que utilizan técnicas de visualización de datos (como paneles de control o dashboards) para explorar los datos almacenados en bases de datos estructuradas. Con este tipo de herramientas, crean paneles visuales (o cuadros de mando) para hacer que la información sea accesible para quienes no son especialistas en datos. Los paneles ayudan a analizar y comprender el desempeño pasado y se utilizan para adaptar la estrategia futura para mejorar los KPI (indicadores clave de negocios).

Ε

Environment (entorno): [Power BI Desktop, Power BI Mobile, etc.] utilizan el entorno de Power BI. En clientes con varios inquilinos, use este término para hacer referencia a la elección del inquilino al que se va a conectar. En aquella documentación en que pueda servir a los analistas de empresas familiarizados con el término "inquilino", se puede utilizar "Entorno Power BI (inquilino)" para ayudarles a reconocer que se trata del mismo concepto.

F

Falso negativo (FN): Representa que la etiqueta predicha es negativa, pero la etiqueta real es

[AFO02848T] IFCD093PO MACHINE LEARNING APLICADO USANDO PYTHON [MOD02447L] IFCD093PO MACHINE LEARNING APLICADO USANDO PYTHON [UDI1502C6] Glosario

positiva, predicha erróneamente. Predijimos que el correo recibido es 'spam' (negativo), pero el correo real recibido es 'normal (positivo).

**Falso positivo (PF):** Representa que la etiqueta predicha es positiva, pero la etiqueta real es negativa, predicha erróneamente. Predijimos que el correo recibido es 'normal (positivo), pero el correo real recibido es 'spam' (negativo).

K

**k-folds cross-validation:** Validación de modelos en la que los datos se dividen en k subconjuntos de igual tamaño, que también se denominan "folds". Uno de los k-folds actuará como el conjunto de prueba, también conocido como el conjunto de holdout o conjunto de validación, y los folds restantes entrenarán el modelo. Este proceso se repite hasta que cada uno de los folds ha actuado como datos de prueba. Después de cada evaluación, se conserva una puntuación y, cuando se han completado todas las iteraciones, las puntuaciones se promedian para evaluar el rendimiento del modelo en general.

L

**LLF (Log-likelihood function):** La función de verosimilitud describe la probabilidad conjunta de los datos observados en función de los parámetros del modelo estadístico elegido.

M

Machine Learning (Aprendizaje automático): es una técnica que puede detectar patrones subyacentes en miles de datos individuales. El desarrollo de aplicaciones predictivas es una de las fortalezas más importantes, ya que facilitan la automatización de procesos, la toma de decisiones y el aprendizaje continúo basado en datos. Además, son sistemas que aprenden automáticamente con el tiempo, se integran en los desarrollos de la empresa y se adaptan a entornos cambiantes cuando se alimentan constantemente de nuevos datos.

**Matriz de confusión:** una matriz de confusión es la matriz de dimensión N \* N en la que un eje representa la etiqueta 'Real' mientras que el otro eje representa la etiqueta 'Predicho'.

0

[AFO02848T] IFCD093PO MACHINE LEARNING APLICADO USANDO PYTHON [MOD02447L] IFCD093PO MACHINE LEARNING APLICADO USANDO PYTHON [UDI1502C6] Glosario

**Overfitting:** el sobreajuste ocurre cuando el modelo ha sido sobreentrenado o cuando contiene demasiada complejidad, lo que provoca altas tasas de error en los datos de prueba. El sobreajuste de un modelo es más común que el subajuste (underfitting). El subajuste generalmente ocurre debido al esfuerzo por evitar el sobreajuste a través de un proceso llamado "detención temprana" (early stopping).

Si el entrenamiento es insuficiente o la falta de complejidad resultan en un ajuste insuficiente, entonces una estrategia de prevención lógica sería aumentar la duración del entrenamiento o agregar entradas más relevantes. Sin embargo, si se entrena demasiado el modelo o se le agregan demasiadas características, también se puede sobreajustar el modelo, lo que resulta en un sesgo (bias) bajo pero una varianza alta (es decir, la compensación sesgo-varianza). En este caso, el modelo estadístico se ajusta demasiado a los datos de entrenamiento, lo que lo hace incapaz de generalizar bien a nuevos puntos de datos. Es importante tener en cuenta que algunos tipos de modelos pueden ser más propensos al sobreajuste que otros, como los árboles de decisión o KNN.

P

PCM (Principal Component Analysis): El análisis de componentes principales (PCA) es una técnica utilizada para identificar el número menor de variables no correlacionadas conocidas como componentes principales a partir de un conjunto más grande de datos. La técnica es ampliamente utilizada para enfatizar la variación y capturar patrones fuertes en un conjunto de datos. Inventado por Karl Pearson en 1901, el análisis de componentes principales es una herramienta utilizada en modelos predictivos y análisis de datos exploratorios. El análisis de componentes principales se considera un método estadístico útil y se utiliza en campos como la compresión de imágenes, el reconocimiento facial, la neurociencia y los gráficos por computadora.

S

**SVD (Singular Value Decomposition):** La descomposición de valores singulares es un tipo particular de factorización de una matriz en un producto de tres matrices, de las cuales la segunda es una matriz diagonal que tiene como entradas en su diagonal los valores singulares de la matriz original.

U

Underfitting: Es un escenario en el que un modelo de datos no puede capturar la relación entre las variables de entrada y salida con precisión, generando una alta tasa de error tanto en el conjunto de entrenamiento como en los datos invisibles. Ocurre cuando un modelo es demasiado simple, lo que puede ser el resultado de que un modelo necesita más tiempo de entrenamiento, más funciones de entrada o menos regularización. Al igual que el sobreajuste, cuando un modelo está infra ajustado, no puede establecer la tendencia dominante dentro de los datos, lo que resulta en errores de entrenamiento y un rendimiento deficiente del modelo. Si un modelo no puede generalizar bien a nuevos datos, entonces no se puede aprovechar para tareas de clasificación o predicción. La generalización de un modelo a nuevos datos es, en última instancia, lo que nos permite utilizar algoritmos de aprendizaje automático para hacer predicciones y clasificar datos.

V

**Verdadero Positivo (TP):** Representa que la etiqueta predicha es positiva y también la etiqueta real es positiva, predicha correctamente. Predijimos que el correo recibido es 'normal (positivo) y el correo real recibido también es 'normal (positivo).

**Verdadero negativo (TN):** Representa que la etiqueta predicha es negativa y también la etiqueta real es negativa, predicha correctamente. Predijimos que el correo recibido es 'spam' (negativo) y el correo real recibido también es 'spam' (negativo).