Guía de estudio - Aprendizaje Supervisado: Regresión

¡Hola! Te damos la bienvenida a esta nueva guía de estudio.

¿En qué consiste esta guía?

En esta guía explicaremos una variedad de técnicas y modelos para abordar problemas de regresión, una de las áreas más fundamentales del aprendizaje automático.

La regresión es una técnica que nos permite predecir valores numéricos continuos en función de características o variables independientes. Este enfoque es ampliamente utilizado en aplicaciones del mundo real, como la predicción de precios de bienes raíces, estimación de ingresos, pronóstico de ventas y muchos otros escenarios donde la predicción de cantidades numéricas es esencial.

Comenzaremos nuestro viaje explorando la Regresión Lineal, uno de los modelos más simples pero poderosos en el ámbito de la regresión. Aprenderemos cómo funciona, cómo se ajusta a los datos y cómo interpretar los resultados obtenidos.
Una vez que entendamos la regresión lineal, nos sumergimos en el mundo de la regularización. Explicaremos las técnicas de Ridge y Lasso, que nos permiten mejorar la generalización del modelo y abordar el sobreajuste.

Continuaremos nuestro viaje con los Árboles de Regresión, una técnica no lineal que nos permite modelar relaciones complejas entre características y el objetivo. Aprenderemos cómo construir, interpretar y visualizar estos árboles.

Es esencial evaluar el rendimiento de nuestros modelos de regresión. En esta sección, veremos diversas métricas de evaluación, como el Error Cuadrático Medio (MSE), el Coeficiente de Determinación (R^2) y otros indicadores clave para medir la calidad de nuestras predicciones.
La práctica es fundamental para fortalecer los conceptos aprendidos. Por lo mismo, realizaremos ejercicios prácticos utilizando conjuntos de datos reales y simulados. Estos ejercicios nos permitirán aplicar los conocimientos adquiridos, ajustar los modelos y comprender cómo enfrentar desafíos comunes que pueden surgir en el mundo real.

¡Esperamos que disfrutes de este emocionante viaje en el mundo de los problemas de regresión y aprender herramientas valiosas para resolver problemas del mundo real con técnicas de Machine Learning! ¡Vamos a comenzar!

¡Vamos con todo!

Tabla de contenidos

Guía de estudio - Aprendizaje Supervisado: Regresión **1**
¿En qué consiste esta guía? 1
Tabla de contenidos 3
Regresión **5**
Aplicaciones de regresión en las industrias 6
Regresión lineal **7**
Modelo de regresión 7
Ventajas y Desventajas 8
Ventajas 8
Desventajas 9
Regularización paramétrica **9**
Ridge 10
Lasso 10
ElasticNet 11
Árboles de regresión **11**
Ventajas y Desventajas 12
Ventajas 13
Desventajas 13
Hiperparámetros 14
Profundidad Máxima (Max Depth) 14
Número Mínimo de Muestras en una Hoja (Min Samples Leaf) 14
Número Mínimo de Muestras para Dividir (Min Samples Split) 14
Número Máximo de Nodos (Max Nodes) 14
Función de Criterio (Criterion) 14
Métricas de Regresión **15**
¡Manos a la obra! - Aplicando lo aprendido 16
Preguntas de proceso 16
Referencias bibliográficas 17

¡Comencemos!

Regresión

La tarea de regresión en el aprendizaje automático consiste en predecir valores numéricos continuos a partir de un conjunto de características o variables independientes. En otras palabras, se busca establecer una relación funcional entre las características de entrada y un valor objetivo, que puede ser cualquier número real. Por ejemplo, predecir el precio de una casa en función de sus características como el tamaño, número de habitaciones, ubicación, etc.

La principal diferencia entre la tarea de regresión y la clasificación radica en el tipo de variable objetivo que se desea predecir. En la regresión, el objetivo es predecir un valor numérico continuo, mientras que en la clasificación, se busca asignar una etiqueta o clase discreta a cada instancia de entrada.

La tarea de regresión es esencial para la toma de decisiones informada y la optimización de procesos en una amplia gama de industrias. La capacidad de predecir valores numéricos continuos permite a las organizaciones tomar decisiones basadas en datos, identificar patrones y tendencias, y mejorar la eficiencia operativa. En resumen, la regresión es una herramienta fundamental para comprender relaciones complejas y realizar predicciones precisas en numerosos escenarios del mundo real.

age	sex	bmi	bp	glucosa
0.038076	0.050680	0.061696	0.021872	-0.017646
-0.001882	-0.044642	-0.051474	-0.026328	-0.092204
0.085299	0.050680	0.044451	-0.005670	-0.025930
-0.089063	-0.044642	-0.011595	-0.036656	-0.009362
0.005383	-0.044642	-0.036385	0.021872	-0.046641
...	...	...	...	...
0.041708	0.050680	0.019662	0.059744	0.007207
-0.005515	0.050680	-0.015906	-0.067642	0.044485
0.041708	0.050680	-0.015906	0.017293	0.015491
-0.045472	-0.044642	0.039062	0.001215	-0.025930
-0.045472	-0.044642	-0.073030	-0.081413	0.003064

Imagen 1. Dataset de Regresión
Fuente: Desafío Latam

Aplicaciones de regresión en las industrias

La importancia de la tarea de regresión radica en su amplio rango de aplicaciones en el mundo real. La capacidad de predecir valores numéricos continuos es esencial en muchas áreas, como:

- **Predicción de Precios de Bienes Raíces:** Estimar el precio de casas, apartamentos u otros bienes inmuebles en función de características como tamaño, ubicación, número de habitaciones, entre otros.
- **Estimación de Ingresos y Gastos:** Predecir los ingresos futuros de una empresa en función de sus ventas, inversiones, gastos y otras variables financieras.
- **Pronóstico de Ventas:** Prever la demanda futura de productos o servicios en función de datos históricos de ventas y factores económicos.
- **Análisis de Datos Climáticos:** Modelar y predecir patrones climáticos y cambios en temperaturas utilizando datos históricos y variables meteorológicas.
- **Predicción de Demanda Energética:** Estimar la demanda futura de energía eléctrica o combustibles en función de factores como el clima, la población y la actividad económica.
- **Predicción de Rendimiento de Estudiantes:** Prever el rendimiento académico de los estudiantes en función de variables como las calificaciones pasadas, la asistencia y el nivel socioeconómico.
- **Estimación de Tiempo de Entrega:** Calcular el tiempo de entrega de productos o servicios en función de la distancia, el tráfico y otros factores logísticos.
- **Análisis de Datos de Salud:** Modelar y predecir la evolución de enfermedades y el riesgo de complicaciones en función de factores médicos y genéticos.
- **Predicción de Costos de Atención Médica:** Estimar los costos de atención médica futuros de los pacientes en función de su historial médico y el tipo de tratamiento requerido.
- **Predicción de Calidad de Productos:** Predecir la calidad de los productos manufacturados en función de variables de proceso y características del producto.
- **Análisis de Precios de Acciones:** Prever los precios futuros de acciones y otros instrumentos financieros utilizando datos históricos y variables económicas.
- **Predicción de Tiempo de Respuesta en Sistemas:** Estimar el tiempo de respuesta de sistemas informáticos o procesos industriales en función de variables de carga y rendimiento.

Estas son solo algunas de las muchas aplicaciones de problemas de regresión en diferentes industrias. La versatilidad y utilidad de la regresión hacen que sea una herramienta esencial para la toma de decisiones y la predicción en una amplia gama de campos.

Regresión lineal

La regresión lineal es una de las técnicas más fundamentales en el campo del aprendizaje automático y la estadística. Se utiliza para establecer una relación lineal entre una variable dependiente (o variable objetivo) y una o más variables independientes (o variables predictoras). La regresión lineal es ampliamente utilizada en una variedad de aplicaciones para predecir valores numéricos continuos y entender la relación entre las variables.

En el ámbito del aprendizaje automático, la regresión lineal se considera una técnica de "aprendizaje supervisado". En este enfoque, se tiene un conjunto de datos etiquetados, es decir, datos para los cuales ya conocemos los valores reales de la variable objetivo. El objetivo es entrenar un modelo de regresión lineal que pueda aprender a mapear las variables predictoras a la variable objetivo, de manera que pueda hacer predicciones precisas en nuevos datos no vistos.

Por otro lado, en el campo de la econometría, la regresión lineal se utiliza para estudiar las relaciones económicas y analizar cómo una variable dependiente puede ser influenciada por variables independientes. Aquí, el enfoque es más interpretativo, y se busca entender las relaciones causales y las implicaciones económicas de los coeficientes estimados en el modelo.

Modelo de regresión

El modelo de regresión lineal simple se define como:

$$y = \beta_0 + \beta_1 * x + \epsilon$$

Donde:

- y es la variable dependiente o variable objetivo.
- x es la variable independiente o variable predictora.
- β_0 y β_1 son los coeficientes que representan la intersección y la pendiente de la línea de regresión, respectivamente.
- ϵ es el término de error, que captura la diferencia entre el valor real y el valor predicho por el modelo.

Para estimar los parámetros β en un modelo de regresión lineal, existen diferentes métodos. El método más común es el Método de Mínimos Cuadrados Ordinarios (MCO).

El Método de Mínimos Cuadrados es el enfoque más utilizado para estimar los coeficientes β en un modelo de regresión lineal. El objetivo del MCO es encontrar los valores de β que minimizan la suma de los cuadrados de las diferencias entre los valores reales y_i y los valores predichos $\hat{y}_i$ por el modelo. La función objetivo a minimizar se define como:

$$MCO: \min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Para encontrar los coeficientes β óptimos, se calcula el gradiente de la función objetivo con respecto a β y se resuelve el sistema de ecuaciones resultante utilizando métodos numéricos, como la eliminación de Gauss o el descenso de gradiente.

La solución de forma matricial para MCO es:

$$\beta = (X^T X)^{-1} X^T Y$$

Esta fórmula nos da los valores óptimos de los coeficientes β_0 y β_1 que minimizan la función objetivo del MCO y nos permiten ajustar la línea de regresión lineal a los datos de entrenamiento de manera eficiente utilizando operaciones matriciales.

Ventajas y Desventajas

El modelo de regresión lineal tiene sus ventajas basadas en su simplicidad e interpretabilidad y sus desventajas debido a lo mismo, a continuación se enumeran las principales ventajas y desventajas del algoritmo.

Ventajas

- **Interpretabilidad:** la regresión lineal es fácil de interpretar, ya que la relación entre las variables se representa mediante una línea recta, lo que facilita la comprensión y explicación del modelo.
- **Eficiencia y Velocidad:** la regresión lineal es computacionalmente eficiente, especialmente para conjuntos de datos grandes, lo que la hace adecuada para problemas con muchas características.
- **Implementación Sencilla:** es relativamente fácil de implementar y entender, lo que lo hace un buen punto de partida para abordar problemas de regresión.

Desventajas

- **Sensibilidad a Outliers:** la regresión lineal es sensible a los valores atípicos, lo que puede afectar significativamente la precisión del modelo.
- **Limitaciones en Datos No Lineales:** si la relación entre las variables es no lineal, la regresión lineal puede no ser el modelo más adecuado para el problema.
- **Multicolinealidad:** si las características están altamente correlacionadas entre sí, la regresión lineal puede producir estimaciones inestables.

Regularización paramétrica

La regresión lineal es una herramienta poderosa para modelar la relación entre variables, pero puede sufrir de problemas como el sobreajuste (overfitting) cuando se enfrenta a conjuntos de datos complejos o con alta dimensionalidad. El sobreajuste ocurre cuando el modelo se ajusta demasiado a los datos de entrenamiento y no generaliza bien a nuevos datos, lo que resulta en predicciones poco confiables.

Para abordar este problema, surge la necesidad de la regularización en la regresión lineal. La regularización es una técnica que introduce una penalización a los coeficientes del modelo, evitando que tomen valores extremadamente grandes y, por lo tanto, reduciendo la complejidad del modelo. Esto ayuda a prevenir el sobreajuste y mejora la capacidad de generalización del modelo a nuevos datos.

En el contexto de la regularización en modelos de aprendizaje automático, las normas de penalización son herramientas fundamentales para controlar la complejidad y el sobreajuste de los modelos. Estas normas se aplican como términos de penalización en la función objetivo durante el proceso de entrenamiento. Su objetivo principal es restringir el tamaño de los coeficientes del modelo, evitando así la amplificación de pequeños ruidos en los datos y mejorando la generalización del modelo a nuevos ejemplos.

Existen dos normas de penalización comúnmente utilizadas en la regularización de modelos: la norma L1 (norma del valor absoluto) y la norma L2 (norma euclidiana al cuadrado). Estas normas, implementadas en técnicas como Ridge, Lasso y Elastic Net, aportan características distintivas que influencian la forma en que se realizan las estimaciones de los coeficientes y se logra la selección de características.

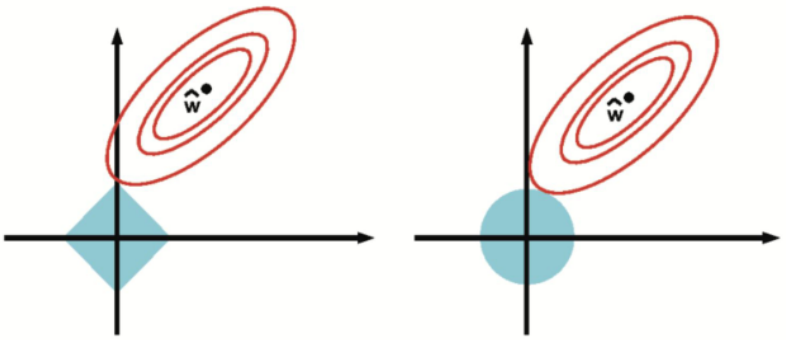

Imagen 2. Normas de penalización
Fuente: Desafío Latam

Ridge

La norma L2 penaliza los coeficientes utilizando el cuadrado de sus valores. En el caso de Ridge, esta penalización tiende a reducir la magnitud de los coeficientes, pero no los anula por completo. Ridge es eficaz para evitar coeficientes muy grandes que podrían llevar al sobreajuste. Aunque no realiza una selección rigurosa de características como Lasso, Ridge puede ser más adecuado cuando se cree que todas las características son relevantes.

La regresión Ridge (o regresión de contracción) agrega un término de penalización proporcional al cuadrado de los coeficientes (β) en la función objetivo de la regresión lineal. La función objetivo se convierte en:

$$\beta_{ridge} = argmin \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Donde λ es el hiper parámetro de regulación que controla la intensidad de la penalización. Ridge tiende a empujar los coeficientes hacia cero, pero no los anula completamente, lo que hace que todos los predictores contribuyan de alguna manera a las predicciones.

Lasso

La norma L1 impone una penalización proporcional al valor absoluto de los coeficientes del modelo. En términos prácticos, Lasso tiende a reducir algunos coeficientes a cero, llevando a una selección automática de características. Esta característica es especialmente útil cuando se enfrenta un conjunto de datos con muchas características, ya que Lasso puede ayudar a simplificar el modelo y eliminar variables irrelevantes.

La regresión Lasso (Least Absolute Shrinkage and Selection Operator) agrega un término de penalización proporcional al valor absoluto de los coeficientes (β) en la función objetivo de la regresión lineal. La función objetivo se convierte en:

$$\beta_{lasso} = argmin \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Lasso tiene la propiedad de reducir coeficientes a cero, lo que implica selección automática de características y ayuda en la simplificación del modelo.

ElasticNet

Elastic Net combina las normas L1 y L2 en una única técnica de regularización. Esto significa que posee características de ambas Lasso y Ridge. Elastic Net busca un equilibrio entre la selección de características y la reducción de coeficientes, lo que puede ser beneficioso en situaciones donde se requiere tanto control de la dimensionalidad como mitigación del sobreajuste.

Elastic Net es una combinación de Ridge y Lasso. Agrega ambos términos de penalización (el cuadrado de los coeficientes y el valor absoluto de los coeficientes) en la función objetivo:

$$\beta_{elasticnet} = argmin \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

Elastic Net combina las propiedades de Ridge y Lasso, permitiendo la selección de características y reduciendo la magnitud de los coeficientes.

En resumen, las normas de penalización L1 y L2 son herramientas poderosas en la regularización, cada una con sus propias características distintivas. Lasso y Ridge son dos técnicas clave que utilizan estas normas para abordar el sobreajuste y mejorar la capacidad de generalización en modelos de regresión. Elastic Net, al combinar ambas normas, ofrece un enfoque equilibrado para controlar la complejidad del modelo y seleccionar características relevantes. La elección de la técnica de regularización dependerá de la naturaleza de los datos y los objetivos específicos del problema.

Árboles de regresión

Los árboles de regresión son una poderosa herramienta en el campo del aprendizaje automático que se utiliza para modelar relaciones entre variables continuas. A diferencia de los árboles de clasificación, que se enfocan en predecir etiquetas discretas, los árboles de regresión se centran en estimar valores numéricos, lo que los convierte en una herramienta esencial para problemas de regresión.

Un árbol de regresión es una estructura jerárquica en forma de árbol que divide el espacio de características en regiones rectangulares. Cada región se asocia con una predicción numérica, que representa el valor promedio de las muestras de entrenamiento que caen en

esa región. La construcción de un árbol de regresión implica seleccionar iterativamente las divisiones de características que mejor reduzcan el error de predicción en las regiones resultantes.

Las etapas para construir un árbol de regresión son similares a las del árbol de clasificación, solo difiere en los criterios de división, que en vez de utilizar criterios de pureza como gini o entropía, se utilizan métricas de regresión. A continuación, se muestran las etapas:

1. **Selección de la División:** El árbol comienza con un nodo raíz que contiene todos los datos de entrenamiento. En cada paso, se selecciona una característica y un umbral que dividirá los datos en dos grupos.

2. **Cálculo de la Predicción:** Se calcula la predicción numérica para cada región basada en los valores promedio de las muestras en esa región.

3. **Criterio de División:** La elección de la característica y el umbral se realiza de manera que la reducción en el error de predicción sea máxima después de la división. El error se mide en términos de alguna métrica, como la suma de los cuadrados de los residuos (SSE) o la desviación absoluta media (MAD).

4. **Crecimiento del Árbol:** El proceso de selección y división se repite para cada región creada en pasos anteriores, hasta que se cumple algún criterio de detención, como la profundidad máxima del árbol o el número mínimo de muestras en una región.

Aunque comparten algunas similitudes en su estructura, los árboles de regresión y los árboles de clasificación tienen diferencias clave:

1. **Tipo de Salida:** La diferencia fundamental radica en el tipo de salida que generan. Los árboles de regresión predicen valores numéricos continuos, mientras que los árboles de clasificación predicen etiquetas discretas.

2. **Métricas de Evaluación:** En los árboles de regresión, las métricas de evaluación se centran en la precisión de las predicciones numéricas, como el error cuadrático medio (MSE) o la raíz del error cuadrático medio (RMSE). En los árboles de clasificación, se utilizan métricas como la precisión, el recall y el F1-score.

3. **Función de División:** En los árboles de regresión, la función de división se basa en la reducción del error en términos de predicción numérica. En los árboles de clasificación, la función de división busca la mayor pureza de las clases resultantes.

Ventajas y Desventajas

Los árboles de regresión son una técnica popular en el análisis de datos y el aprendizaje automático para resolver problemas de regresión. Como cualquier método, tienen sus ventajas y desventajas. Aquí hay un resumen de las principales ventajas y desventajas de los árboles de regresión:

Ventajas

1. **Interpretabilidad:** Los árboles de regresión son fáciles de entender e interpretar. La estructura jerárquica en forma de árbol es visualmente intuitiva y puede ser explicada fácilmente a las partes interesadas.

2. **No Linealidad:** Pueden capturar relaciones no lineales entre variables predictoras y la variable objetivo. No se limitan a relaciones lineales, lo que los hace útiles en una amplia gama de aplicaciones.

3. **Robustos ante Outliers:** Los árboles de regresión son relativamente robustos ante valores atípicos en los datos. Las divisiones en regiones rectangulares tienden a mitigar el efecto de los valores extremos.

4. **Flexibilidad:** Pueden manejar múltiples características de entrada y se pueden usar en problemas de alta dimensionalidad.

5. **Tratamiento Automático de Variables:** No requieren transformaciones manuales de las variables predictoras, como normalización o estandarización, ya que trabajan en base a divisiones de características.

Desventajas

1. **Sobreajuste:** Los árboles de regresión tienden a sobre ajustar los datos de entrenamiento, especialmente cuando el árbol es profundo. Pueden capturar el ruido en los datos y generar predicciones poco confiables en nuevos datos.

2. **Estabilidad:** Los pequeños cambios en los datos de entrenamiento pueden resultar en árboles completamente diferentes. Esto hace que los árboles sean inestables y sensibles a variaciones en los datos.

3. **Limitaciones en la extrapolación:** Los árboles de regresión pueden tener dificultades para extrapolar más allá del rango de valores observados en los datos de entrenamiento.

4. **No Considera Relaciones Globales:** Debido a su naturaleza jerárquica, los árboles de regresión pueden no capturar relaciones globales complejas en los datos, especialmente si la estructura de división no lo permite.

En general, los árboles de regresión son una herramienta valiosa y versátil en el kit de herramientas de análisis de datos y aprendizaje automático. Sin embargo, es importante considerar cuidadosamente las ventajas y desventajas al seleccionarlos para resolver un problema específico, y en muchos casos, combinarlos con otras técnicas puede ser beneficioso para mejorar su rendimiento y robustez.

Hiperparámetros

Los hiperparámetros son valores ajustables que se configuran antes de entrenar un modelo y que afectan la forma en que el modelo se ajusta a los datos. En el caso de los árboles de regresión, los hiper parámetros controlan aspectos clave del proceso de construcción del árbol y, por lo tanto, influyen en la precisión, complejidad y capacidad de generalización del modelo. Aquí hay una explicación de algunos de los hiper parámetros más importantes en los árboles de regresión:

Profundidad Máxima (Max Depth)

La profundidad máxima de un árbol de regresión controla la cantidad de divisiones que puede tener el árbol. Un árbol más profundo puede capturar relaciones más complejas en los datos, pero también es más propenso al sobreajuste, especialmente si la profundidad es demasiado grande. Limitar la profundidad puede ayudar a controlar el sobreajuste.

Número Mínimo de Muestras en una Hoja (Min Samples Leaf)

Este hiper parámetro establece el número mínimo de muestras que deben estar en una hoja del árbol. Si se establece un valor alto, el árbol tendrá hojas con más muestras, lo que puede resultar en un modelo más estable pero menos flexible. Un valor bajo puede llevar al sobreajuste.

Número Mínimo de Muestras para Dividir (Min Samples Split)

Determina el número mínimo de muestras requeridas para que un nodo pueda dividirse. Establecer un valor alto evita divisiones en regiones con pocas muestras, lo que puede ayudar a prevenir el sobreajuste.

Número Máximo de Nodos (Max Nodes)

Establece el número máximo de nodos en el árbol. Controla la complejidad total del árbol y, por lo tanto, su capacidad de adaptación. Limitar el número de nodos puede ayudar a evitar el sobreajuste.

Función de Criterio (Criterion)

Es el criterio utilizado para medir la calidad de una división. Los dos criterios más comunes son el Error Cuadrático Medio (MSE) y la Desviación Absoluta Media (MAE). El MSE es más sensible a errores grandes, mientras que el MAE es más robusto ante valores atípicos.

Ajustar estos hiper parámetros de manera adecuada es crucial para obtener un árbol de regresión bien calibrado y con un buen rendimiento en datos nuevos. Sin embargo, también es importante considerar el riesgo de sobreajuste y encontrar un equilibrio que permita al modelo generalizar correctamente. La selección y ajuste de hiper parámetros a menudo implica el uso de técnicas como la validación cruzada para evaluar el rendimiento del modelo en diferentes conjuntos de datos.

Métricas de Regresión

Las métricas de regresión desempeñan un papel fundamental en la evaluación de la precisión y el rendimiento de los modelos de regresión. Estas métricas permiten medir cuán cerca están las predicciones del modelo de los valores reales, lo que es crucial para entender la efectividad y la utilidad de un modelo en la práctica. Una elección adecuada de métricas puede guiar la selección de hiper parámetros y la comparación entre diferentes algoritmos, brindando información valiosa sobre qué tan bien se ajusta un modelo a los datos.

Las métricas de regresión son esenciales por varias razones:

- **Evaluación del Rendimiento:** Permiten una evaluación cuantitativa y objetiva de cuán bien el modelo está haciendo predicciones sobre nuevos datos.
- **Comparación de Modelos:** Ayudan a comparar diferentes modelos y enfoques para determinar cuál se ajusta mejor a los datos y proporciona predicciones más precisas.
- **Selección de Hiper parámetros:** Ayudan a ajustar los hiper parámetros del modelo para optimizar su rendimiento.
- **Toma de Decisiones:** Son fundamentales para tomar decisiones informadas basadas en las predicciones del modelo.

A continuación, se presentan algunas métricas de regresión comunes junto con sus fórmulas y su interpretación:

Métrica	Descripción	Fórmula
Error Cuadrático Medio (MSE)	Promedio de los cuadrados de las diferencias entre las predicciones y los valores reales. Penaliza más los errores grandes.	$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$
Raíz del Error Cuadrático Medio (RMSE)	Raíz cuadrada del MSE y tiene las mismas unidades que la variable objetivo. Proporciona una medida más intuitiva del error.	$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$

Desviación Absoluta Media (MAE)	Promedio de las diferencias absolutas entre las predicciones y los valores reales. Menos sensible a valores atípicos.	$MAE = \frac{\sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert}{n}$
Error Absoluto Porcentual Medio (MAPE)	Promedio de los errores porcentuales absolutos entre las predicciones y los valores reales. Útil para interpretar errores en términos relativos.	$MAPE = \frac{\sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert}{\sum_{i=1}^{n} \lvert y_i \rvert} \cdot 100$
Coeficiente de determinación (R2)	Proporción de la varianza en la variable objetivo que es explicada por el modelo. El R2 más cercano a 1 indica un mejor ajuste.	$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$

Las métricas de regresión proporcionan una forma cuantitativa de medir y comparar la precisión de los modelos de regresión. Cada métrica ofrece una perspectiva única sobre el rendimiento del modelo y su capacidad para hacer predicciones precisas. Al elegir métricas apropiadas y comprender su interpretación, los científicos de datos pueden evaluar de manera efectiva la calidad de sus modelos y tomar decisiones informadas para mejorar su rendimiento en aplicaciones del mundo real.

¡Manos a la obra! - Aplicando lo aprendido

Aplicaremos lo aprendido utilizando Python. Para ello:

- Analiza el ejercicio resuelto en el archivo 02 - Ejercicio guiado
- Replica el proceso con tu propio modelo, utilizando el dataset **diabetes**.

Preguntas de proceso

Reflexiona:

- ¿Qué es un problema de regresión y en qué se diferencia de un problema de clasificación?
- ¿Cuál es el objetivo fundamental en un problema de regresión?
- ¿Por qué es importante comprender el contexto del problema antes de abordar un problema de regresión?
- ¿Qué es la regularización en el contexto de la regresión lineal y cuál es su propósito?
- Describe la diferencia entre L1 (Lasso) y L2 (Ridge) regularización en la regresión lineal.
- ¿Por qué la regularización puede ayudar a prevenir el sobreajuste en la regresión lineal?
- ¿Cuál es la diferencia fundamental entre un árbol de regresión y un árbol de clasificación?
- ¿Cómo se elige la característica óptima para dividir en un nodo de un árbol de regresión?
- ¿Cuál es el criterio de parada en un árbol de regresión y cómo ayuda a evitar el sobreajuste?
- ¿Qué representa el Coeficiente de Determinación (R^2) en las métricas de regresión?
- ¿Cómo se interpreta un valor de R2 cercano a 0? ¿Y un valor cercano a 1?
- ¿Cuándo podría ser preferible usar el Error Absoluto Porcentual Medio (MAPE) en lugar del MSE o MAE?
- ¿Por qué es importante considerar tanto el rendimiento de entrenamiento como el de prueba al evaluar un modelo de regresión?

Referencias bibliográficas

1. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
2. Python Data Science Handbook: https://jakevdp.github.io/PythonDataScienceHandbook
3. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
4. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html
5. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html
6. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html
7. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

¡Continúa aprendiendo y practicando!