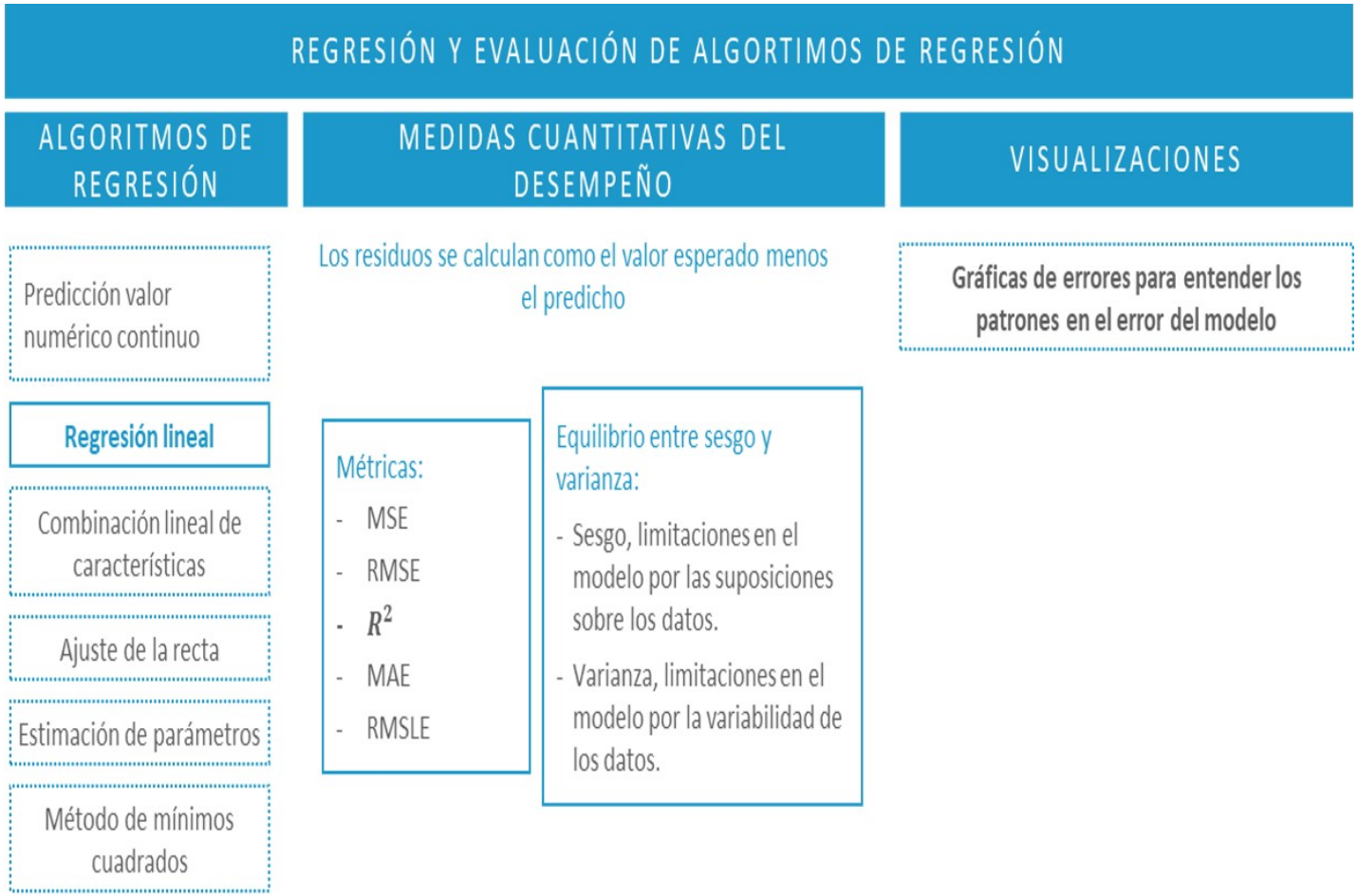
Técnicas de Aprendizaje Automático

Tema 4. Regresión y evaluación de algoritmos de regresión

Índice

Esquema



4.1. Introducción y objetivos

Después de sentar las bases en el aprendizaje automático y en el tratamiento de los datos, pasamos a analizar las técnicas de regresión tradicionales. Antes de adentrarnos en el estudio de los diferentes algoritmos de aprendizaje supervisado, es importante conocer las diferentes formas de medir el desempeño al modelar un resultado numérico continuo. En este tema se pretende proporcionar una comprensión práctica y una intuición para los modelos de regresión que buscan la estructura subyacente de los datos utilizando combinaciones lineales de los predictores.

Como ya se presentó anteriormente, una de las tareas que se puede llevar a cabo con los modelos de aprendizaje supervisado es la **regresión**. El objetivo de estos modelos es predecir un resultado numérico continuo (por ejemplo: predecir el consumo energético de una planta fotovoltaica). Para evaluar la efectividad de estos modelos, normalmente se utiliza alguna medida de precisión. Sin embargo, existen diferentes formas de medir la precisión, cada una con sus propios matices. Para comprender las fortalezas y debilidades de un modelo en particular, depender únicamente de una única métrica es problemático. Las visualizaciones del ajuste del modelo, en particular, los gráficos residuales, son fundamentales para comprender si el modelo es adecuado para su propósito. Estas técnicas se analizan en este capítulo.

El objetivo de este tema es obtener una mayor comprensión sobre los algoritmos de regresión, utilizar las métricas de error más comunes y ser capaz de visualizar los errores de forma gráfica.

- Comprensión de las métricas de error más comunes para medir la precisión del modelo.

Ideas clave

- Identificar el sesgo y la varianza de los datos.
- Identificar y manejar overfitting y underfitting en los modelos de regresión.
- Validar la generalización con visualizaciones de la efectividad de los modelos.

4.2. Algoritmos de regresión

Dentro del campo del aprendizaje automático, los algoritmos de regresión son un tipo concreto de **algoritmos de aprendizaje supervisado** y consisten en realizar una predicción de una variable numérica o cuantitativa. En el aprendizaje supervisado, para cada una de las observaciones de las variables predictoras (x_i) existe una medida de la variable respuesta (y_i y_i). Se desea crear un **modelo que relacione las variables y las respuestas** con el objetivo de predecir las respuestas de observaciones en el futuro.

Existen numerosos problemas del mundo real donde la variable que se desea predecir o estimar es una variable numérica continua; por ejemplo, los ingresos de una persona, el precio de venta de un inmueble o algo tan dispar como la fuerza del cemento. De forma general, el aprendizaje supervisado se puede utilizar en todos aquellos casos en los cuales existe **conocimiento de un valor cuantitativo previo.**

Recordamos ahora, la **terminología o notación más común** vista en temas previos:

- Variable respuesta: también conocida como *outcome*, variable dependiente, variable objetivo, *target*, *class*, etc. Se suele denotar por *y*.
- Vector de *p* mediciones predictoras llamado *x*: también conocido como *inputs*, *regressors*, *features*, variables, variables independientes, características, etc.
- Se tienen datos de entrenamiento conocidos como *training data*: $((x_1, y_1),\ldots,(x_N, y_N))$ que son observaciones, ejemplos o instancias de cada una de las medidas de entrada.

Regresión lineal

Antes de avanzar en las métricas existentes para evaluar el error de un modelo de regresión, merece la pena dedicar un espacio a los modelos de regresión lineal. La regresión lineal es un modelo fantástico para empezar porque es uno de los modelos «caballo de batalla» que se utilizan en una amplia variedad de campos, incluidos los negocios, la ciencia y la medicina. Las regresiones lineales son a menudo el primer modelo al que se recurre para comprender cómo se relacionan dos fenómenos diferentes. La simplicidad y la utilidad de la regresión lineal siempre serán valiosas.

La regresión lineal recibe su nombre del hecho de que el resultado es una combinación lineal de los predictores.

La regresión es una técnica (estadística) que intenta modelar la relación entre dos conjuntos de variables. Un conjunto de variables, llamado predictores, incluye variables que creemos que influyen en el valor de otras variables de interés, a las que nos referimos como resultado. También llamamos a los predictores «características» o «covariables». Un análisis de regresión intenta determinar cómo se relacionan estos predictores con el resultado. De manera matemática se puede ver como una función:

$$y = f(x) + \in$$

La función f es lo que llamamos regresión. Al utilizar una regresión, asumimos que el valor de y se puede dividir en dos componentes:

- La primera parte, f(x) , describe cómo los predictores contribuyen al resultado.
- La segunda parte, $\in$, se llama error.

Sin embargo, esta relación no es perfecta. Generalmente hay alguna diferencia entre los valores de f(x) e y porque los predictores normalmente no se pueden usar para

calcular el resultado de manera determinista. El error, $\in$, existe para completar el vacío entre f(x) e y .

Recordemos que el valor de y en un problema de regresión es un valor numérico continuo. Por ejemplo, supongamos que y representa la puntuación de riesgo para un determinado coche. Una puntuación de riesgo mayor supondrá que tendrá una prima superior en el seguro del coche. Supongamos ahora que x representa la potencia del motor. Estamos interesados en determinar cómo la potencia del motor influye en la puntación de riesgo del coche. Para determinar esta relación se puede utilizar una regresión lineal.

Como ya se adelantó, la palabra lineal en la regresión lineal sugiere que el resultado se describe, de alguna manera, como una combinación lineal de los predictores de la siguiente manera:

$$f(x) = y = \beta_0 + \beta_1 X + \epsilon$$

Donde, β1 es el coeficiente (pendiente) asociado a X. Si β1 es grande, un cambio en X implicará un cambio en y y viceversa. β0 es el intercepto (intersección). El intercepto representa el valor de y cuando X=0 . La forma en que se interpreta la intersección puede variar mucho según lo que represente el predictor. A veces, puede que ni siquiera tenga interpretación.

Ambos, β0 y β1, se conocen como los parámetros del modelo. Los **parámetros** son valores que deben aprenderse de los datos y los usamos para calcular la relación entre los predictores y el resultado. Cuando hay un solo predictor en la regresión, nos referimos a esto como **regresión lineal simple**. Si se tienen múltiples predictores en un modelo, lo convierte en una **regresión lineal múltiple**.

En el caso de una regresión lineal múltiple, el resultado expresado como una combinación lineal de los predictores, junto con el error, sería de la siguiente manera:

$$f(x) = y = \beta_0 + \beta_1 X_1 . + \beta_p X_p + \epsilon$$

Una vez que los coeficientes han sido aprendidos, existe una regresión lineal funcional que se puede usar para generar las predicciones. Pero dados solo los datos, no sabemos cuáles deberían ser los parámetros. Una regresión lineal hará su mejor predicción cuando el error, $\in$, sea lo más pequeño posible.

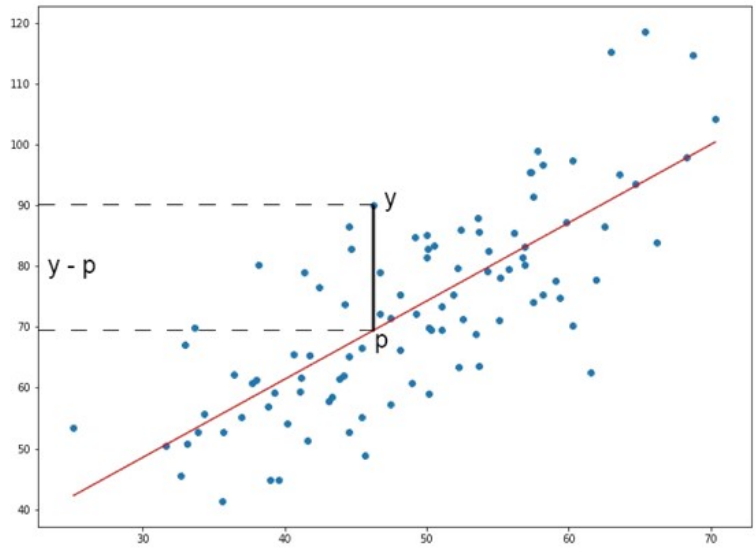


Figura 1. Puntos aleatorios y su línea de regresión lineal. Fuente: Medium, 2018.

Por lo tanto, el objetivo será encontrar el parámetro β que mejor ajusta la pendiente, por lo tanto, que **minimiza el error**. Entonces, para minimizar el error, necesitamos primeramente una forma de calcularlo. Una función de pérdida en aprendizaje automático es simplemente una medida de qué tan diferente es el valor predicho del valor real. En el caso de la regresión lineal se usará la función de pérdida cuadrática para calcular la pérdida o el error en el modelo. Se puede definir como:

$$L(X) = \sum_{i=1}^{n} (y_i - p_i)^2$$

Una vez definida la función de error, lo único que queda por hacer es minimizarla.

Esto se hace encontrando la derivada parcial de L e igualándola a 0, y luego encontrando una expresión para los parámetros β. Después de hacer los cálculos, nos quedan estas ecuaciones:

$$\beta_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

$$\beta_0 = \overline{y} - \beta_1 \overline{x}$$

Donde $\overline{x}$ es la media de todos los valores de x, e $\overline{y}$ es la media del valor de la salida. Esto se conoce como el **método de mínimos cuadrados**.

4.3. Medidas cuantitativas del diseño

En los problemas de regresión, cuando se está tratando de predecir un valor numérico continuo, los residuos son importantes fuentes de información para evaluar la eficiencia del modelo. Los residuos se calculan como el valor observado (yi menos el valor predicho (y^i), (yi-y^i). Teniendo esto presente, entonces cuando el resultado es un número, el método más común para caracterizar las capacidades predictivas de un modelo es utilizar la **raíz del error cuadrático medio (RMSE)**. Esta métrica es una función de los residuos del modelo, que son los valores observados menos las predicciones del modelo.

El **error cuadrático medio (MSE)** se calcula elevando al cuadrado los residuos y sumándolos. Luego, el RMSE se calcula tomando la **raíz cuadrada del MSE** para que esté en las **mismas unidades que los datos originales**. El valor generalmente se interpreta como qué tan lejos (en promedio) están los residuos de cero o como la distancia promedio entre los valores observados y las predicciones del modelo.

Las definiciones matemáticas de estas métricas son:

- Error cuadrático medio, mean square error **(MSE)**: media de la diferencia entre el valor real y el valor predicho o estimado al cuadrado.

$$MSE = 1/n \sum_{i=1}^{n} \left(y_i - \widehat{y_i}\right)^2$$

- Raíz del error cuadrático medio, Root Mean Square Error **(RMSE)**: la raíz cuadrada de la media de la diferencia entre el valor real y el valor predicho o estimado al cuadrado.

$$RMSE = \sqrt{1/n \sum_{i=1}^{n} \left(y_i - \widehat{y_i}\right)^2}$$

Ideas clave

El **MSE** penaliza fuertemente los errores grandes debido a su naturaleza cuadrática. Por su parte, el **RMSE** también opera de esa manera, pero con una interpretación más intuitiva, ya que se devuelve a la escala original de la variable objetivo.

Algunas características importantes del MSE son:

- **Penalización cuadrática:** el MSE calcula el promedio de los cuadrados de las diferencias entre las predicciones del modelo y los valores reales. Esto significa que los errores más grandes tienen un impacto más significativo en el MSE que los errores más pequeños. Por lo tanto, el MSE es más sensible a los errores grandes que otras métricas.

- **Interpretación en la misma escala:** el MSE se mide en las mismas unidades cuadradas que la variable objetivo, lo que puede dificultar su interpretación directa.

- **Propenso a valores atípicos:** debido a la naturaleza cuadrática de la métrica, el MSE es más sensible a los valores atípicos o a las diferencias extremas entre las predicciones y los valores reales. Esto puede hacer que el MSE sea menos robusto en presencia de valores atípicos que otras métricas.

- **Optimización matemática:** muchos algoritmos de optimización están diseñados para minimizar el MSE durante el entrenamiento de modelos de regresión, lo que hace que el MSE sea una opción popular como función de pérdida en el proceso de entrenamiento.

Algunas características importantes del RMSE son:

- **Interpretación en la misma escala:** el RMSE se mide en las mismas unidades que la variable objetivo, lo que facilita su interpretación en comparación con el MSE.

- **Penalización cuadrática:** al igual que el MSE, el RMSE penaliza más fuertemente los errores grandes debido a su naturaleza cuadrática. Esto significa que los errores más grandes tienen un impacto más significativo en el RMSE que los errores más pequeños.

- **Facilidad de interpretación:** el RMSE proporciona una medida de dispersión similar a la desviación estándar, lo que puede ser útil para comprender la variabilidad de las predicciones del modelo en relación con los valores reales.
- **Optimización matemática:** al igual que el MSE, el RMSE es una métrica comúnmente utilizada en el proceso de entrenamiento de modelos de regresión, ya que muchos algoritmos de optimización están diseñados para minimizarlo.

Otra métrica común es el coeficiente de determinación, comúnmente escrito como R^2. Este valor puede interpretarse como la **proporción de la información de los datos que explica el modelo**. Por tanto, un valor de R^2 de 0,75 implica que el modelo puede explicar tres cuartas partes de la variación del resultado. Existen múltiples fórmulas para calcular este valor (Kvalseth, 1985), aunque la versión más simple encuentra el coeficiente de correlación entre los valores observados y los valores predichos (generalmente denotado por R) y lo eleva al cuadrado. Si bien este es un valor fácilmente interpretable, es importante recordar y tener en cuenta cuando se evalúe un modelo con esta métrica, que, según lo dicho **R^2, es una medida de correlación,** no de precisión. En la Figura 2 se muestra un ejemplo donde el coeficiente R^2 entre los valores observados y los valores predichos es de 0,57, pero se observa que el modelo tiene una tendencia a sobrepredecir valores bajos y subestimar los altos.

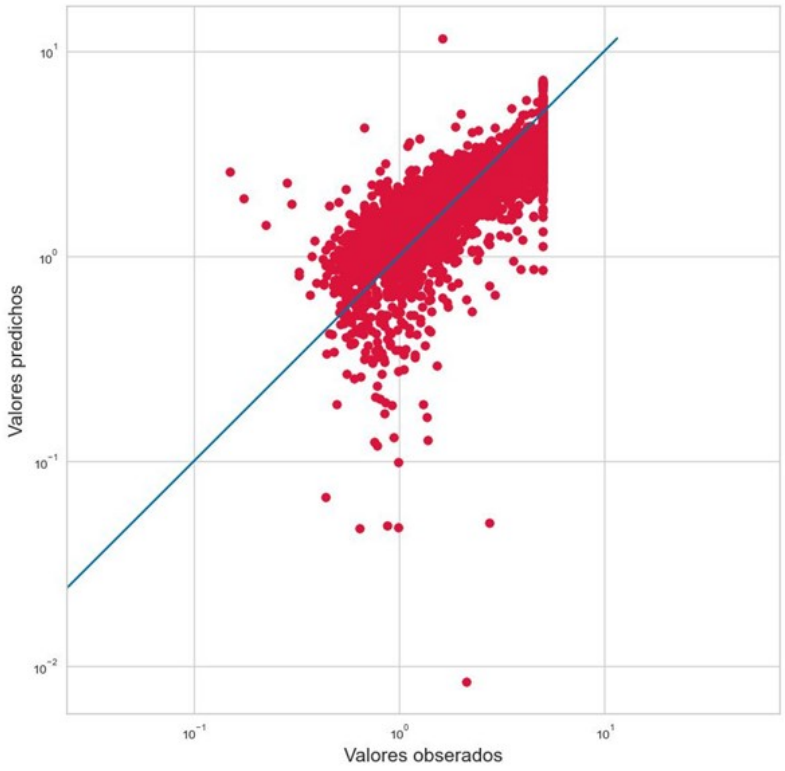


Figura 2. Gráfico de los valores observados y predichos donde el valor de R^2 es moderado (57 %).

Fuente: elaboración propia.

En este gráfico, las predicciones no son uniformemente precisas. La línea azul de referencia indica dónde serían las predicciones iguales.

También es importante darse cuenta de que R^2 es dependiente de la variación del resultado. Utilizando la interpretación de que esta estadística mide la proporción de la varianza explicada por el modelo, hay que recordar que el denominador de esa proporción se calcula utilizando la varianza muestral del resultado. Por ejemplo, supongamos que el resultado de un conjunto de pruebas tiene una varianza de 4,2. Si el RMSE de un modelo predictivo fuera 1, el R^2 sería aproximadamente del 76 %. Si tuviéramos otro conjunto de pruebas con exactamente el mismo RMSE, pero los

resultados de la prueba fueran menos variables, los resultados se verían peor. Por ejemplo, si la varianza del conjunto de prueba fuera 3, el R^2 sería del 67 %.

Junto con estas métricas, existen otras dos más que también pueden ser útiles en algunos contextos.

El **error absoluto medio, Mean Absolute Error (MAE)**, se define como la diferencia en valor absoluto entre el valor real y el valor predicho. Esta métrica es menos sensible a los valores atípicos en comparación con otras métricas. Se calcula de la siguiente manera:

$$MAE = 1/n\sum_{i=1}^{n}|y_i - \widehat{y_i}|$$

Es simple pero efectiva, ya que calcula la diferencia absoluta promedio entre las predicciones del modelo y los valores reales. MAE es una **medida de la magnitud promedio de los errores en las predicciones**, sin considerar su dirección, lo que significa que tanto las sobreestimaciones como las subestimaciones contribuyen por igual al error total.

Algunas características importantes de MAE son:

- **Interpretación directa:** MAE es fácil de interpretar, ya que representa la magnitud promedio de los errores en las unidades originales de la variable objetivo. Por ejemplo, si se están prediciendo los precios de las casas en euros, un MAE de 10 000€ significa que, en promedio, las predicciones están desviadas en 10 000€ del precio real.

- **Robustez a los valores atípicos:** MAE es menos sensible a los valores atípicos en comparación con otras métricas de regresión, como el error cuadrático medio (MSE) o el error cuadrático medio raíz (RMSE), ya que no penaliza los errores más grandes más fuertemente.

- **Fácil de calcular:** MAE es computacionalmente eficiente y fácil de implementar.

La **raíz del logaritmo cuadrático del error medio**, RMSLE (Root Mean Squared Logarithmic Error):

$$RMSLE = \sqrt{1/n \sum_{i=1}^{n} \left(\log\left(y_1 + 1\right) - \log\left(\widehat{y_i} + 1\right)\right)^2}$$

RMSLE es una métrica comúnmente utilizada en problemas de regresión cuando las variables objetivo tienen una escala muy amplia y las diferencias relativas son más importantes que las diferencias absolutas. Esta métrica es especialmente útil cuando se trabaja con variables que tienen valores muy dispersos o siguen distribuciones de cola larga, como los precios de las acciones, las ventas minoristas, las métricas de rendimiento en publicidad, etc.

La fórmula de RMSLE se calcula tomando el logaritmo de los valores predichos y reales antes de calcular el error cuadrático medio, lo cual tiene el efecto de penalizar más fuertemente las diferencias en las predicciones más pequeñas (*underprediction*) en comparación con las diferencias en las predicciones más grandes (*overprediction*).

Algunas características importantes de RMSLE son:

- **Interpretación:** al tomar el logaritmo de los valores, se convierte la diferencia relativa entre los valores predichos y reales en una escala más uniforme y fácilmente interpretable.
- **Robustez:** RMSLE es menos sensible a las diferencias en las predicciones de las muestras con valores muy altos o bajos, lo que la hace más robusta en presencia de valores atípicos o datos dispersos.

Finalmente, se presentan dos últimas métricas que tienen la característica común de que ambas miden el porcentaje de error: **MAPE** y **MPE**.

El error porcentual absoluto medio (MAPE) es el porcentaje equivalente de MAE. La ecuación se parece a la del MAE, pero con ajustes para convertir todo en

porcentajes.

$$MAPE = \frac{100}{n} \sum \left| \frac{y - \hat{y}}{y} \right|$$

Así como MAE es la magnitud promedio del error producido por su modelo, MAPE es qué tan lejos están las predicciones del modelo de sus resultados correspondientes en promedio. Al igual que MAE, MAPE también tiene una interpretación clara, ya que los porcentajes son más fáciles de conceptualizar para las personas. Tanto MAPE como MAE son robustos a los efectos de valores atípicos gracias al uso del valor absoluto.

Sin embargo, a pesar de todas sus ventajas, estamos más limitados en el uso de MAPE que MAE. Muchas de las debilidades del MAPE en realidad surgen de la división, ya que al tener que escalar todo con los valores reales, MAPE no está definido para los puntos de datos donde el valor es 0. De manera similar, MAPE puede crecer inesperadamente si los valores reales son excepcionalmente pequeños.

Finalmente, el MAPE está sesgado hacia predicciones que son sistemáticamente menores que los valores reales mismos. Es decir, MAPE será menor cuando la predicción sea menor que la real en comparación con una predicción que sea mayor en la misma cantidad. Veamos esto con un ejemplo. Si tuviéramos un valor de predicción $\hat{y}$=10 más pequeño que el valor real y=20, y con n=1, entonces el valor de MAPE sería del 50 %. En cambio, si tuviéramos un valor de predicción $\hat{y}$=20 mayor que el valor real y=10, y con n=1, entonces el valor de MAPE sería del 100 %.

Por su parte, el error porcentual medio (MPE) es exactamente igual al de MAPE. La única diferencia es que carece de la operación de valor absoluto.

$$MPE = \frac{100}{n} \sum \left(\frac{y - \hat{y}}{y} \right)$$

Aunque el MPE carece de la operación de valor absoluto, en realidad es su ausencia lo que hace que el MPE sea útil. Dado que los errores positivos y negativos se cancelarán, no podemos hacer ninguna afirmación sobre qué tan bien se desempeñan las predicciones del modelo en general. Sin embargo, si hay más errores negativos o positivos, este sesgo aparecerá en el MPE. A diferencia de MAE y MAPE, MPE nos resulta útil porque nos permite ver si nuestro modelo subestima sistemáticamente (más error negativo) o sobreestima (error positivo).

Resumidamente, se han presentado las principales métricas de evaluación en los modelos de regresión con sus principales fortalezas y debilidades. A modo de resumen se presenta la siguiente tabla (Tabla 1) que recopila sus principales características.

Acrónimo	Nombre completo	Cálculo de los residuos	Robusta valores anómalos
MSE	Error cuadrátrico medio	Cuadrado	No
RMSE	Raíz del error cuadrático medio	Cuadrado	No
MAE	Error absoluto medio	Valor absoluto	Sí
RMSLE	Raíz del logaritmo cuadrático del error medio	Cuadrado	Sí
MAPE	Error porcentual absoluto medio	Valor absoluto	Sí
MPE	Error porcentual medio	N/A	Sí

Tabla 1. Tabla resumen de las principales características de las métricas de error para modelos de regresión presentadas. Fuente: elaboración propia.

Es importante recordar que todas las métricas presentadas se obtienen a partir del cálculo de los residuos producidos por los modelos; es decir, la diferencia entre el valor real y el estimado. Para cada una de ellas, se utiliza la magnitud de la métrica para decidir si el modelo está funcionando bien o no. De manera general, los valores de métrica de error pequeños indican una buena capacidad predictiva, mientras que los valores grandes sugieren lo contrario. A la hora de decidir qué métrica utilizar es

importante tener en cuenta la naturaleza de los datos. Los valores atípicos pueden ser claves a la hora de tomar esta decisión, ya que algunos dominios de datos pueden ser más propensos a tener valores atípicos, mientras que otros pueden no verlos con tanta frecuencia.

4.4. Equilibrio entre la varianza y el sesgo

El equilibrio o el compromiso sesgo-varianza es un compromiso que permite analizar el rendimiento y la eficacia en las tareas de aprendizaje supervisado.

En estadística y aprendizaje automático, el equilibrio sesgo-varianza describe la relación entre la complejidad de un modelo, la precisión de sus predicciones y qué tan bien puede hacer predicciones sobre datos que no se utilizaron para entrenar el modelo. En general, a medida que aumentamos la cantidad de parámetros ajustables en un modelo, este se vuelve más flexible y puede adaptarse mejor a un conjunto de datos de entrenamiento. Se dice que tiene menor error o sesgo. Sin embargo, para modelos más flexibles, tenderá a haber una mayor variación en el ajuste del modelo cada vez que tomamos un conjunto de muestras para crear un nuevo conjunto de datos de entrenamiento. Se dice que existe una mayor varianza en los parámetros estimados del modelo.

El dilema o problema de sesgo-varianza es el conflicto que se da al intentar minimizar simultáneamente estas dos fuentes de error que impiden que los algoritmos de aprendizaje supervisado se generalicen más allá de su conjunto de entrenamiento (Wikipedia, 2024):

- El **error de sesgo** es un error debido a suposiciones erróneas en el algoritmo de aprendizaje. Un sesgo alto puede hacer que un algoritmo pierda las relaciones relevantes entre las características y los resultados objetivo (desajuste).
- La **varianza** es un error de la sensibilidad a pequeñas fluctuaciones en el conjunto de entrenamiento. Una alta varianza puede resultar de un algoritmo que modela el ruido aleatorio en los datos de entrenamiento (sobreajuste).

Ideas clave

La descomposición sesgo-varianza es una forma de analizar el error de generalización esperado de un algoritmo de aprendizaje con respecto a un problema particular como una suma de tres términos, el sesgo, la varianza y una cantidad llamada error irreducible, resultante del ruido en el problema mismo.

4.5. Visualización de los errores

Una de las mejores formas de evaluar un modelo de regresión es utilizando gráficos. Uno de los gráficos que más información proporciona visualmente es un **diagrama de dispersión de dos dimensiones** donde el eje x son los valores reales y el eje y, los valores predichos o estimados (o viceversa). En la Figura 3 se muestra un ejemplo de visualización de los valores reales y la línea de regresión.

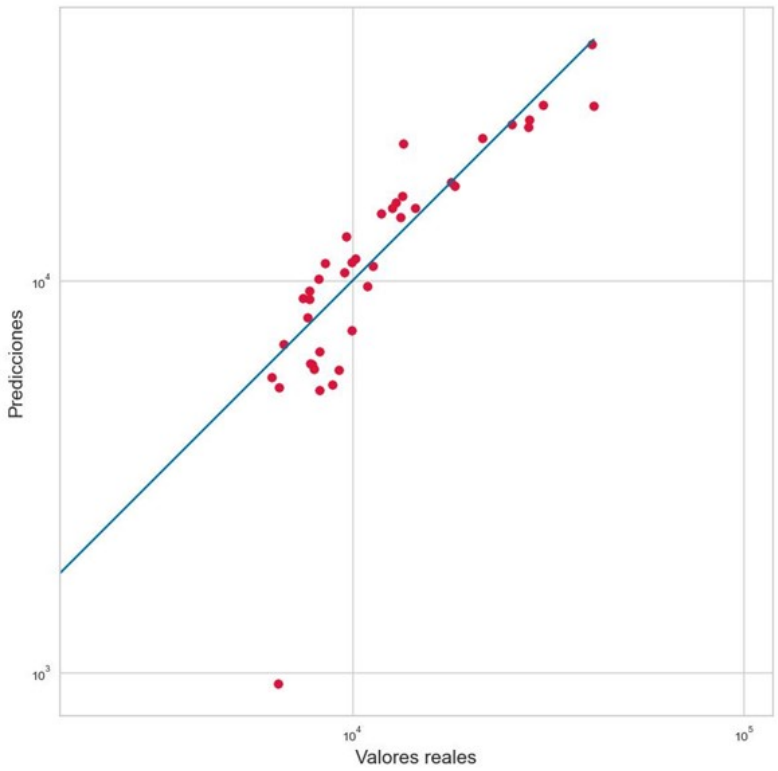


Figura 3. Gráfico de dispersión que muestra la relación entre las predicciones y los valore reales. Fuente: elaboración propia.

Cada uno de los puntos corresponde a una de las estimaciones o predicciones realizadas. En el caso de que las **predicciones o estimaciones fueran perfectas**, caerían directamente sobre la diagonal del gráfico. Por otro lado, los puntos que caen por encima de la diagonal son sobreestimaciones de las predicciones (*overpredictions*), mientras que los puntos que caen por debajo de la diagonal son predicciones que se quedan cortas (*underpredictions*).

Además de tener un gráfico de dispersión de los valores reales frente a los valores predichos, suele ser útil tener una gráfica de los residuos. Una gráfica de residuos es una representación gráfica de los residuos (errores) en un modelo de regresión lineal. Los residuos son las diferencias entre los valores observados de la variable dependiente y los valores predichos obtenidos del modelo de regresión lineal. En términos simples, un gráfico de residuos muestra qué tan lejos están las predicciones de los puntos de datos reales.

El gráfico de residuos se crea trazando los residuos en el eje vertical frente a los valores predichos o variables independientes en el eje horizontal. Esta representación visual ayuda a identificar patrones, tendencias o anomalías en los datos, que pueden indicar problemas potenciales con el modelo de regresión, como no linealidad, heterocedasticidad o la presencia de valores atípicos. En la Figura 4 se presenta un ejemplo de este tipo de gráfico.

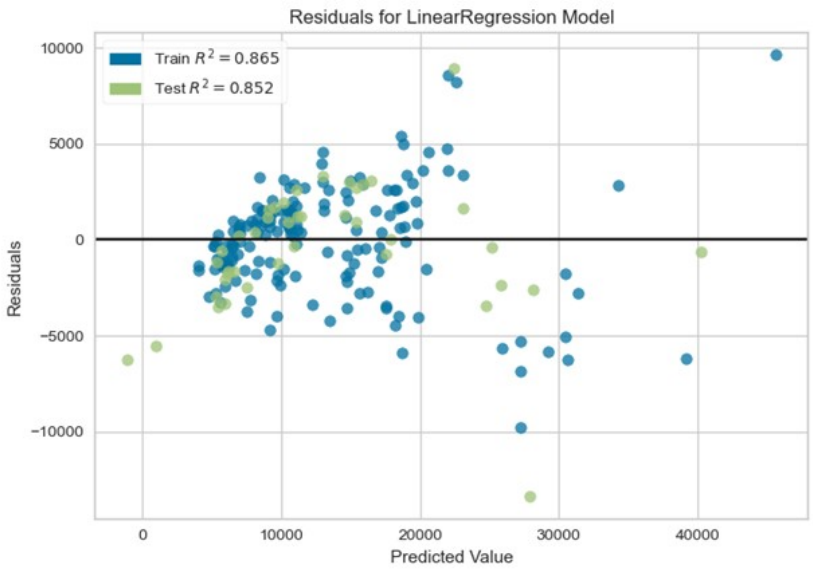


Figura 4. Gráfico de residuos que muestra la relación entre las predicciones y los residuos. Fuente: elaboración propia.

Un uso común de los gráficos de residuos es analizar la varianza del error del modelo de regresión. Si los puntos están dispersos aleatoriamente alrededor del eje horizontal, un modelo de regresión lineal suele ser apropiado para los datos; de lo contrario, es más apropiado un modelo no lineal.

4.6. Cuaderno de ejercicios

- Describe las funciones de la librería *scikit-learn* que permiten hacer uso de las siguientes métricas de evaluación: MSE, RMSE, R2, MAE y RMSLE. Calcula cada una de ellas para los siguientes datos: y_true = [3, 5, 2.5, 7] , y_ pred = [2.5, 0.0, 2, 8] .

SOLUCIÓN:

```
from sklearn.metrics import mean_squared_error

from sklearn.metrics import r2_score

from sklearn.metrics import mean_absolute_error

from sklearn.metrics import mean_squared_log_error

from sklearn.metrics import root_mean_squared_error



y_true = [3, 5, 2.5, 7]

y_pred = [2.5, 0.0, 2, 8]



mean_squared_error(y_true, y_pred)

root_mean_squared_error(y_true, y_pred)

r2_score(y_true, y_pred)



mean_absolute_error(y_true, y_pred)

root_mean_squared_log_error(y_true, y_pred)
```

- El siguiente código hace uso del dataset *fetch_california_housing*, disponible en *sklearn.datasets.fetch_california_housing* y genera un modelo de regresión lineal para la predicción del precio de vivienda. En las variables *y_test*, *y_pred* se almacenan los resultados de las observaciones y las predicciones, respectivamente. A partir de estos valores, se debe calcular MSE, RMSE y R2. Explica la b ondad del modelo. ¿Qué significan los resultados obtenidos?

```
import pandas as pd

import numpy as np

from sklearn.datasets import fetch_california_housing

import matplotlib.pyplot as plt

from pandas.plotting import scatter_matrix

from sklearn.metrics import r2_score

from sklearn.linear_model import LinearRegression

from sklearn.pipeline import Pipeline

from sklearn.preprocessing import StandardScaler

from sklearn.model_selection import train_test_split

housing = fetch_california_housing(as_frame=True)

housing = housing.frame

housing.head()

X = housing.iloc[:,:-1]

y = housing.iloc[:,-1]
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

regression_pipeline = Pipeline([

('scaler', StandardScaler()),

('regressor', LinearRegression())

])

regression_pipeline.fit(X_train,y_train)

y_pred = regression_pipeline.predict(X_test)
```

SOLUCIÓN:

MSE: 0.5558915986952442

R2: 0.575787706032451

- Con los resultados de predicción del ejercicio anterior realiza una visualización de los errores frente a las observaciones. En el gráfico, además de los datos indicados, debe aparecer la línea de ajuste óptimo.

SOLUCIÓN:

Deberías obtener un gráfico como el de la Figura 1 de este mismo tema.

4.7. Referencias bibliográficas

Hastie, T., Tibshirani, R. y Friedman, J. (2009). *The Elements of Statistical Learning* (2ª ed.). Springer.

James, G., Witten, D., Hastie, T y Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Kuhn, M. y Johnson, K. (2013). *Applied predictive modeling* (Vol. 26, p. 13). Springer.

Kvalseth, T (1985). Cautionary Note About R2. *American Statistician, 39*(4), 279–285.

Menon, A. (2008). Linear Regression Using Least Squares. Towards Data Science. Medium. https://towardsdatascience.com/linear-regression-using-least-squares-a4c3456e8570

Pino, A. E., Chichande, B. S. y Tovar, Y. J. (2019). Determinación de modelos predictivos para los indicadores de competitividad empresarial aplicando regresión lineal. *Revista Ibérica de Sistemas e Tecnologías de Información, E18*, 94-107.

Wikipedia. (2024). *Bias–variance tradeoff*. https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff

Métricas de error para regresión de scikit-learn

Scikit-learn (s.f.). *Metrics and scoring: quantifying the quality of predictions: Regression metrics* https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics

En este recurso se encuentra la documentación de la librería *scikit-learn* y su módulo de métricas de error. El módulo *sklearn.metrics* implementa varias funciones de pérdida, puntuación y utilidad para medir el rendimiento de la regresión.

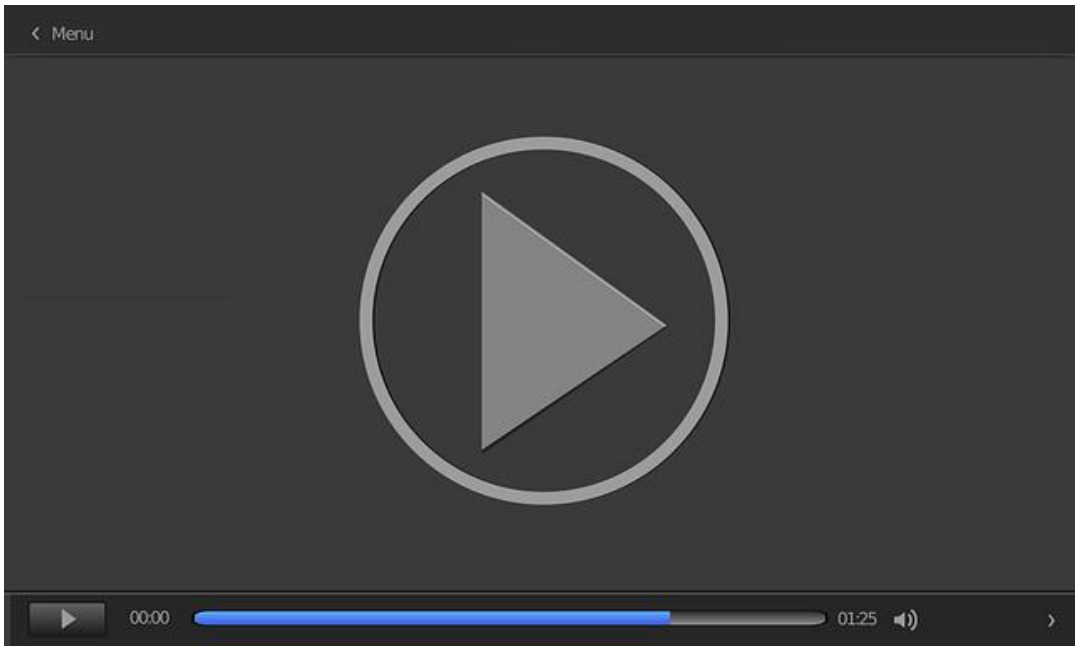
Métricas de error para regresión de Keras

Keras. (s.f.). *Regression metrics*. https://keras.io/api/metrics/regression_metrics/

En este recurso se encuentra la documentación de la librería Keras 3 y su módulo de métricas de error. El módulo *keras.metrics* implementa las funciones de error estudiadas en este tema para medir el rendimiento de la regresión.

Tutorial sobre Análisis de Regresión Lineal

statisticsfun. (2012, febrero 5). *An Introduction to Linear Regression Analysis* [Vídeo]. YouTube. https://www.youtube.com/watch?v=zPG4NjIkCjc



Accede al vídeo:

https://www.youtube.com/embed/zPG4NjIkCjc

En este vídeo se presenta una introducción estadística al análisis de regresión lineal. El vídeo permitirá recordar los conceptos estadísticos asociados a un análisis de regresión lineal, para después poder entender cómo aplicar este método a las tareas de aprendizaje automático.

Simulación de Regresión

PhET Interactive simulations, University of Colorado at Boulder. *Least-Squares Regression*. https://phet.colorado.edu/en/simulations/least-squares-regression

El proyecto PhET Interactive Simulators de la Universidad de Colorado Boulder crea simulaciones interactivas gratuitas de matemáticas y ciencias. Los simuladores de PhET se basan en una extensa investigación educativa e involucran a los estudiantes a través de un entorno intuitivo similar a un juego donde aprenden a través de la exploración y el descubrimiento. En esta simulación se presenta un entorno para comprender cómo opera una regresión lineal y cada uno de los elementos involucrados en ella.

Test

1. ¿Cuáles de las siguientes métricas no se usan para evaluar los problemas de regresión?

A. Error absoluto medio.

B. Error cuadrático medio.

C. Raíz del error cuadrático medio.

D. Raíz del error logarítmico cuadrático.

2. Indica la afirmación correcta sobre el Error absoluto medio (MAE):

A. MAE mide la diferencia promedio entre predicciones y valores reales.

B. MAE es una medida de la precisión del modelo de regresión.

C. MAE penaliza de manera cuadrática las diferencias entre predicciones y valores reales.

D. MAE es insensible a los valores atípicos en los datos.

3. ¿Puede mejorar el R^2 (de entrenamiento) de un modelo de regresión lineal múltiple después de eliminar o agregar características?

A. Sí, el R^2 (de entrenamiento) de un modelo de regresión lineal múltiple puede mejorar después de eliminar o agregar características. Esto ocurre porque el valor de R^2 es una medida de qué tan bien las variables independientes del modelo explican la variabilidad de la variable dependiente.

B. Sí, el R^2 (de entrenamiento) de un modelo de regresión lineal múltiple puede mejorar después de eliminar o agregar características. Esto ocurre porque el valor de R^2 es una medida sobre cuántas variables independientes explican la variabilidad de la variable dependiente.

C. No, el R^2 (de entrenamiento) de un modelo de regresión lineal múltiple siempre puede empeorar después de eliminar o agregar características. Esto ocurre porque el valor de R^2 es una medida que está altamente correlacionada con el número de características empleadas en el entrenamiento.

D. No, el R^2 (de entrenamiento) de un modelo de regresión lineal múltiple permanece estable para los diferentes conjuntos de características empeladas en el entrenamiento. Esto es debido a que R^2 es una medida que relaciona los datos con los resultados.

4. Indica la afirmación correcta sobre el error cuadrático medio (MSE):

A. MSE mide el promedio de los cuadrados de los errores.

B. MSE es una medida robusta que ignora los valores atípicos.

C. MSE proporciona una medida lineal de la precisión del modelo.

D. MSE es menos sensible a los errores grandes en comparación con MAE.

5. Indica la afirmación correcta sobre la raíz del error cuadrático medio (RMSE):

A. RMSE es más sensible a los errores grandes en comparación con MAE en problemas de regresión de aprendizaje automático.

B. RMSE proporciona una medida de error en diferentes unidades que la variable objetivo en problemas de regresión.

C. RMSE calcula el logaritmo de la raíz cuadrada del promedio de los cuadrados de las diferencias entre predicciones y valores reales.

D. RMSE es una medida de error robusta que ignora la magnitud de las diferencias entre las predicciones y los valores reales.

6. ¿Cuál es el objetivo de la regresión?

A. Comprender el valor de los predictores X.

B. Comprender el valor del error.

C. Comprender el valor de las salidas Y.

D. Comprender la relación entre el predictor X y la salida Y.

7. ¿Qué indica un coeficiente de regresión negativo en un modelo lineal?

A. La variable predictora tiene un efecto negativo en la variable de respuesta.

B. No hay relación entre la variable predictora y la variable de respuesta.

C. La variable predictora tiene un efecto positivo en la variable de respuesta.

D. El modelo no es adecuado para hacer predicciones.

8. ¿Qué representa el intercepto en un modelo de regresión lineal?

A. El valor esperado de la variable de respuesta cuando todas las variables predictoras son cero.

B. La pendiente de la línea de regresión.

C. La suma de los cuadrados de los residuos.

D. La varianza explicada por el modelo.

9. ¿Cuál de las siguientes afirmaciones describe correctamente el compromiso sesgo-varianza en un modelo de machine learning?

A. El compromiso sesgo-varianza se refiere a la relación entre el sesgo y la varianza de un modelo, donde un sesgo alto puede llevar a un subajuste y una varianza alta puede llevar a un sobreajuste.

B. El compromiso sesgo-varianza solo se aplica a modelos de regresión y no a otros tipos de modelos de machine learning.

C. Un modelo con bajo sesgo y alta varianza generalmente se ajusta bien a los datos de entrenamiento y generaliza bien a nuevos datos.

D. El compromiso sesgo-varianza no es relevante en el contexto de machine learning y no afecta el rendimiento del modelo.

10. ¿Cuál de las siguientes visualizaciones es más adecuada para evaluar los errores en un modelo de regresión?

A. Gráfico de barras de frecuencia de errores absolutos.

B. Gráfico de dispersión de residuos vs. valores predichos.

C. Gráfico de línea de errores relativos vs. iteraciones de entrenamiento.

D. Gráfico de pastel de errores cuadráticos medios.