

Resumen del capítulo: métricas de regresión

Coeficiente de determinación

El **coeficiente de determinación** o la **métrica R2 (R-al cuadrado)** divide el ECM del modelo entre el ECM de la media y luego resta el valor obtenido de uno. Si la métrica aumenta, la calidad del modelo también mejora.

R2 se calcula de la siguiente manera:

$$R^2 = 1 - \frac{\text{Modelo ECM}}{\text{Media ECM}}$$

- R2 es igual a uno solo si el valor ECM es cero. Dicho modelo predeciría perfectamente todas las respuestas.
- R2 es cero: el modelo funciona tan bien como la media.
- Cuando R2 es negativo, la calidad del modelo es muy baja.
- R2 no puede tener valores mayores a uno.

Para calcular esta métrica, puedes usar la *función* `r2_score()` de la librería `sklearn.metrics`:

```
from sklearn.metrics import r2_score

print("R2 =", r2_score(target, predicted))
```

Error absoluto medio

Necesitamos establecer la notación convencional para *Data Science*:

$$y_i$$

- El valor objetivo para la observación con número de serie i en la muestra utilizada para medir la calidad. El subíndice indica el número de serie de la observación.

$$\hat{y}_i$$

- El valor de predicción para la observación con número de serie i (en la muestra de prueba, por ejemplo).

EAM (error absoluto medio) es otra métrica de evaluación. Es algo similar al *ECM*, pero no está al cuadrado. Vamos a escribirlo usando las notaciones convencionales para Data Science.

El error de una observación se escribe de la siguiente manera:

$$\text{Error} = y_i - \hat{y}_i$$

Para cancelar la diferencia entre subajuste y sobreajuste, se calcula el **error absoluto**.

$$\text{Error absoluto} = |y_i - \hat{y}_i|$$

Para recopilar los errores a lo largo de la muestra, agreguemos la siguiente notación:

N

- El número de observaciones en la muestra.

$$\sum_{i=1}^N$$

- La suma de todas las observaciones de la muestra (i varía en el rango de 1 a N).

La fórmula para EAM, o error absoluto medio, es:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

Para calcular esta métrica, puedes usar la función `mean_absolute_error()` de la librería `sklearn.metrics`:

```
from sklearn.metrics import mean_absolute_error

mae = mean_absolute_error(target, predicted))
```

Para calcular el valor ECM, usamos el valor de la media como una constante.

El modelo constante debe seleccionarse de manera que te permita obtener el valor de métrica EAM más bajo posible. Necesitamos encontrar el valor a, en el que se alcanza el mínimo:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - a|$$

El mínimo se alcanza cuando **a** es igual a la mediana del objetivo.

A diferencia del *EAM*, la métrica *RECM* es más sensible a valores grandes: los **errores significativos** afectan fuertemente el valor final de *RECM*. De esa forma, puedes cambiar el valor de una métrica sin cambiar otra.