# (Opcional) Lectura: Evaluación de Modelos de Aprendizaje Automático

Tiempo estimado: 30 minutos

### Definir la división de entrenamiento/prueba

Entender la división de entrenamiento/prueba es fundamental en el aprendizaje automático, particularmente en escenarios de aprendizaje supervisado. Este concepto implica dividir su conjunto de datos en dos partes: los conjuntos de entrenamiento y de prueba. Mientras que el conjunto de entrenamiento educa al modelo sobre los patrones dentro de los datos, el conjunto de prueba evalúa su capacidad para generalizar a datos nuevos y no vistos. Por ejemplo, imagina que estás prediciendo precios de casas basándote en características como tamaño, habitaciones y ubicación. Dividirías tus datos en conjuntos de entrenamiento y prueba, permitiendo que el modelo aprenda de uno y sea evaluado en el otro. Este proceso asegura que puedas evaluar el rendimiento del modelo con precisión.

En este escenario, el 80% del conjunto de datos se asigna para entrenamiento (`X_train and y_train`), mientras que el 20% restante se asigna para prueba (`X_test and y_test`). Al especificar el parámetro random_state, la división se vuelve reproducible. Esto significa que ejecutar el código con el mismo valor de random_state producirá consistentemente la misma división, asegurando consistencia a través de múltiples ejecuciones.

### Evaluar modelos de clasificación utilizando precisión, una matriz de confusión, precisión y recuperación

**Precisión**: La precisión cuantifica la proporción de predicciones correctas entre todas las predicciones generadas por el modelo. Por ejemplo, al predecir precios de casas, la precisión indica el porcentaje de precios pronosticados correctamente entre todos los precios dentro del conjunto de datos. Se calcula determinando la proporción de instancias que fueron predichas correctamente sobre el total de instancias.

$$\text{Precisión} = \frac{\text{Número total de predicciones}}{\text{Número de predicciones correctas}}$$

$$\text{Precisio´n} = \text{Nu´mero de predicciones correctasNu´mero total de predicciones}$$

**Matriz de confusión**: Una matriz de confusión ofrece una visión concisa de cuán bien se desempeña un modelo de clasificación. Al comparar las predicciones del modelo con los resultados reales, determinamos las cuentas de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Si bien aplicar directamente una matriz de confusión para predecir precios de casas puede parecer poco convencional debido a su naturaleza continua, puedes discretizar los precios en categorías (por ejemplo, barato, moderado, caro) y luego utilizar la matriz de confusión para comparar las categorías predichas con las reales.

Una matriz de confusión proporciona un resumen de los resultados de predicción en un problema de clasificación. La matriz muestra las cuentas de `verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN).`

| **Real/predicho** | **Precio alto** | **Precio bajo** |
|---|---|---|
| Precio alto | TP | FN |
| Precio bajo | FP | TN |

**Precisión**: La precisión es cuántas `predicciones positivas correctas` hizo el modelo de todas sus `predicciones positivas`. En el ámbito de la predicción de precios de casas, la precisión refleja la exactitud de los precios de casas predichos entre todas las predicciones. Sin embargo, dada la naturaleza continua de los precios de las casas, la precisión comúnmente se define dentro de un nivel de tolerancia específico. Por ejemplo, podrías establecer un umbral (por ejemplo, dentro del 5% del precio real) y calcular la precisión basándote en las predicciones que caen dentro de ese rango.

$$\text{Precisión} = \frac{TP}{TP + FP}\text{Precisio´n} = \text{TP} + \text{FP}\,\text{TP}$$

**Recuperación**: La recuperación mide cuántas `predicciones positivas verdaderas` se hicieron de todas las `instancias positivas reales`. En el escenario de la predicción de precios de casas, la recuperación indica el porcentaje de precios de casas predichos correctamente entre todos los precios de casas reales. Típicamente, la recuperación se define dentro de un nivel de tolerancia específico para acomodar la naturaleza continua de los precios de las casas.

$$\text{Recuperación} = \frac{TP}{TP + FN}\text{Recuperacio´n} = \text{TP} + \text{FN}\,\text{TP}$$

**Aquí hay un ejemplo ilustrativo:**

Suponga que su conjunto de datos ha sido dividido en conjuntos de entrenamiento y prueba, y ha construido un modelo de clasificación. Ejecuta el modelo en el conjunto de prueba y obtiene la siguiente matriz de confusión:

| **Real/predicho** | **Precio alto** | **Precio bajo** |
|---|---|---|
| Precio alto | 80 | 20 |
| Precio bajo | 10 | 90 |

**De esta matriz de confusión:**

**Verdaderos positivos (TP)** = 80 (Precio alto predicho como Precio alto) En los datos proporcionados, de 100 casas caras, el modelo identificó correctamente 80 como de precio alto.

**Verdaderos negativos (TN)** = 90 (Precio bajo predicho como Precio bajo) En los datos dados, de 100 casas no caras, el modelo identificó correctamente 90 como de precio bajo.

**Falsos positivos (FP)** = 10 (Precio bajo predicho como Precio alto) En los datos dados, de 100 casas no caras, el modelo clasificó incorrectamente 10 como de precio alto.

**Falsos negativos (FN)** = 20 (Precio alto predicho como Precio bajo) En los datos dados, de 100 casas caras, el modelo clasificó incorrectamente 20 como de precio bajo.

**Ahora calcule las métricas de evaluación:**

$$\text{Exactitud} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{80 + 90}{80 + 90 + 10 + 20} = \frac{170}{200} = 0.85 = 85\%$$
Exactitud = TP+TN+FP+FNTP+TN = 80+90+10+2080+90 = 200170 = 0.85 = 85%

$$\text{Precisión} = \frac{TP}{TP + FP} = \frac{80}{80 + 10} = \frac{80}{90} \approx 0.89 = 89\%$$
Precisio´n = TP+FPTP = 80+1080 = 9080 ≈ 0.89 = 89%

$$\text{Recuperación} = \frac{TP}{TP + FN} = \frac{80}{80 + 20} = \frac{80}{100} \approx 0.80 = 80\%$$
Recuperacio´n = TP+FNTP = 80+2080 = 10080 ≈ 0.80 = 80%

Estas métricas te ayudan a comprender el rendimiento de tu modelo de clasificación, lo que te permite tomar decisiones informadas sobre mejoras o ajustes del modelo.

### Evaluar un modelo de regresión utilizando el error cuadrático medio y otros tipos de términos de error

Evaluar un modelo de regresión implica `evaluar qué tan bien el modelo predice la variable objetivo`. Aquí, discutiremos cómo evaluar un modelo de regresión utilizando el error cuadrático medio (MSE) y otras métricas de error como el error absoluto medio (MAE), el error cuadrático medio de la raíz (RMSE) y el R-cuadrado ($R^2$).

- **Error cuadrático medio (MSE):**
  - El MSE calcula la media de los errores cuadrados, representando la discrepancia cuadrada promedio (diferencia) entre los valores estimados y el valor real.
  - Fórmula:
    $$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$MSE = n1 ∑i=1n(yi − yi^)2
  - Un MSE más bajo indica un mejor rendimiento del modelo.
- **Error absoluto medio (MAE):**
  - El MAE calcula el tamaño promedio de los errores en un conjunto de predicciones, sin tener en cuenta su dirección.
  - Fórmula:
    $$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$MAE = n1 ∑i=1n |yi − yi^|
  - Un MAE más bajo indica un mejor rendimiento del modelo.
- **R-cuadrado ($R^2$):**
  - El $R^2$ nos indica cuánto de los cambios en la variable dependiente son explicados por los cambios en las variables independientes.
  - Fórmula:
    $$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$R2 = 1 − ∑i=1n(yi−y¯)2∑i=1n(yi−yi^)2
  - Valores más altos de $R^2$, que van de 0 a 1, significan un mejor rendimiento del modelo.
- **Error cuadrático medio de la raíz (RMSE):**
  - El RMSE, como la raíz cuadrada del MSE, proporciona una visión del tamaño típico de los errores.
  - Fórmula:
    $$\text{RMSE} = \sqrt{\text{MSE}}$$RMSE = MSE
- **Error absoluto porcentual medio (MAPE):**
  - El MAPE cuantifica la precisión como un porcentaje del error.
  - Fórmula:
    $$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$MAPE = n1 ∑i=1n ||yiyi−yi^|| × 100

- **Error absoluto mediano:**
  - El error absoluto mediano es robusto a los valores atípicos porque utiliza la mediana en lugar de la media.
  - Útil cuando tus datos tienen valores atípicos que podrían sesgar las métricas de error.

### Ejemplo de conjunto de datos

Aquí hay un pequeño conjunto de datos para ilustración:

| Tamaño (pies cuadrados) | Número de habitaciones | Ubicación | Precio real ($) | Precio predicho ($) |
|---|---|---|---|---|
| 1500 | 3 | 1 | 300000 | 310000 |
| 1600 | 3 | 2 | 320000 | 315000 |
| 1700 | 4 | 1 | 350000 | 345000 |
| 1800 | 4 | 3 | 370000 | 375000 |
| 1900 | 5 | 2 | 400000 | 390000 |

## Error cuadrático medio (MSE):

**Cálculo:**

MSE = (1/5) * [ $(300000 - 310000)^2$ + $(320000 - 315000)^2$ + $(350000 - 345000)^2$ + $(370000 - 375000)^2$ + $(400000 - 390000)^2$ ]

MSE = (1/5) * [ $(-10000)^2$ + $(5000)^2$ + $(5000)^2$ + $(-5000)^2$ + $(10000)^2$ ]

MSE = (1/5) * [ 100,000,000 + 25,000,000 + 25,000,000 + 25,000,000 + 100,000,000 ]

MSE = 275,000,000 / 5

MSE = 55,000,000

## Error absoluto medio (EAM):

**Cálculo:**

EAM = (1/5) * (|300000 - 310000| + |320000 - 315000| + |350000 - 345000| + |370000 - 375000| + |400000 - 390000|)

EAM = (1/5) * (10000 + 5000 + 5000 + 5000 + 10000)

EAM = 35000 / 5

EAM = 7000

## Error cuadrático medio (RMSE):

**Cálculo:**

MSE = (1/5) * [$(300000 - 310000)^2$ + $(320000 - 315000)^2$ + $(350000 - 345000)^2$ + $(370000 - 375000)^2$ + $(400000 - 390000)^2$]

MSE = (1/5) * [$(-10000)^2$ + $(5000)^2$ + $(5000)^2$ + $(-5000)^2$ + $(10000)^2$]

MSE = (1/5) * [100,000,000 + 25,000,000 + 25,000,000 + 25,000,000 + 100,000,000]

MSE = 275,000,000 / 5

MSE = 55,000,000

RMSE = sqrt(55,000,000)

RMSE ≈ 7,416.20

## Más sobre R-cuadrado ($R^2$):

R-cuadrado, o el coeficiente de determinación, es una `medida estadística que representa la proporción de la varianza en la variable dependiente (precio de la casa) que puede ser predicha a partir de las variables independientes` (tamaño, número de habitaciones y ubicación). Sirve como un indicador clave de qué tan bien se ajusta un modelo de regresión a los datos.

### Fórmula

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \text{R2} = 1 - \sum \text{i=1n(yi−y¯)2} \sum \text{i=1n(yi−yi^)2}$$

donde,

- $(y_i)$(yi) = Valor real de la variable dependiente (precio de la casa)
- $(\hat{y}_i)$(y^i) = Valor predicho de la variable dependiente
- $(\bar{y})$(y¯) = Media de los valores reales de la variable dependiente
- $(n)$(n) = Número de observaciones
- $(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2)$(∑i=1n(yi − y^i)2) = Suma de cuadrados de los residuos (SSR) o suma de cuadrados residual (RSS)
- $(\sum_{i=1}^{n}(y_i - \bar{y})^2)$(∑i=1n(yi − y¯)2) = Suma total de cuadrados (TSS)

**Interpretación**

- **$R^2 = 1$**: El modelo explica toda la variabilidad de los datos de respuesta alrededor de su media. Las predicciones coinciden perfectamente con los datos reales.
- **$R^2 = 0$**: El modelo no tiene en cuenta ninguna variación en los datos de respuesta alrededor de su media. Las predicciones son tan buenas como simplemente usar la media de los datos reales.
- **$0 < R^2 < 1$**: El modelo explica una parte de la variabilidad, con valores más altos indicando un mejor ajuste.
- **$R^2 < 0$**: Esto puede ocurrir si el modelo es peor que una línea horizontal (media de los valores reales), lo que típicamente indica un modelo incorrecto.

## Cálculo numérico

Supongamos el siguiente conjunto de prueba hipotético y predicciones:

- Precios reales
  $(y_{\text{test}})$(ytest) = [370000, 320000]
- Precios predichos
  $(y_{\text{pred}})$(ypred) = [360000, 310000]

**Pasos y cálculos:**

1. Calcular la media de $(y_{\text{test}})$: $[\bar{y} = \frac{\sum y_{\text{test}}}{n}](\text{ytest}) : [\bar{y} = n\sum \text{ytest}]$

   Dado $(y_{\text{test}})$(ytest) = [370000, 320000]
   $\bar{y}$ y¯ = (370000 + 320000)/2 = 345000

2. Calcular la suma total de cuadrados $((SS_{\text{tot}}))$: $SS_{\text{tot}} = \sum_{i=1}^{n}(y_i - \bar{y})^2((\text{SStot})) : \text{SStot} = \sum i=1n(yi - \bar{y})2$
   $SS_{\text{tot}}$ SStot = $(370000 - 345000)^2 + (320000 - 345000)^2$
   $SS_{\text{tot}}$ SStot = $(25000)^2 + (-25000)^2$
   $SS_{\text{tot}}$ SStot = (625000000 + 625000000) = 1250000000

3. Calcular $(R^2)$: $[R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}](\text{R2}) : [\text{R2} = 1 - \text{SStotSSres}]$
   $R^2$ R2 = 1 - (200000000/1250000000)
   $R^2$ R2 = 1 - 0.16
   $R^2$ R2 = 0.84

# Conclusión

- Es importante dividir tu conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba.
- Una matriz de confusión evalúa el rendimiento y la precisión de un problema de clasificación.
- El error cuadrático medio es útil para evaluar modelos de regresión.
- El valor $R^2$ nos indica cuánto de la variación de la variable dependiente puede ser comprendido por los cambios en la variable independiente.

## Autores

Niveditha Pandith TS
Malika Singla