Análisis de regresión

¿Qué es el análisis de regresión?

Es el análisis que se realiza a los modelos de machine learning que tienen como objetivo predecir valores numéricos de forma precisa.

- Analizamos qué tan bueno o malo es un modelo haciendo predicciones numéricas.
  - “¿Cuánto ganará esta persona el próximo año?”
  - “¿Cuántos productos venderá esta tienda el próximo mes?”
- Estos modelos utilizan información pasada para entrenarse y luego predecir valores numéricos.
- Entrenamos diversos modelos y, en base a las métricas de error, elegimos el más adecuado.

En esta clase nos enfocaremos en entender cómo interpretar las métricas más relevantes del análisis de regresión.

Caso - Análisis de regresión

Supongamos el siguiente caso

- Quieres predecir cuánto dinero gastará una persona en tu tienda el próximo mes.
- Para esto usarás información histórica de personas que han comprado:
  - Ingresos, Edad, Frecuencia de compra, Monto comprado, Entre otros
- Con esta información entrenaremos un modelo que prediga un valor numérico continuo (el monto gastado).
  - Probaremos este modelo con nuevas personas.
  - Analizaremos qué tan cerca está la predicción del valor real.
  - Calcularemos el error promedio y otras métricas como MAE, RMSE y R^2.

Interpretación - Análisis de regresión

Supongamos los siguientes resultados para 5 estimaciones

n	Mes	Valor real	Valor estimado	Error absoluto |e(t)|	Error cuadrático e(t)^2	(Valor real - promedio(valor real))^2
1	Ene	210	200	10	100	1156
2	Feb	250	245	5	25	36
3	Mar	230	210	20	400	196
4	Abr	270	260	10	100	676
5	May	260	255	5	25	256
	Suma	**1220**	**1170**	**50**	**650**	**2320**
	Promedio	**244**	**234**	**10**	**130**	**464**

Con estas columnas podemos calcular las métricas más relevantes para el análisis de regresión.

Interpretación - Análisis de regresión

n	Mes	Valor real	Valor estimado	Error absoluto |e(t)|	Error cuadrático e(t)^2	(Valor real - promedio(valor real))^2
1	Ene	210	200	10	100	1156
2	Feb	250	245	5	25	36
3	Mar	230	210	20	400	196
4	Abr	270	260	10	100	676
5	May	260	255	5	25	256
	Suma	**1220**	**1170**	**50**	**650**	**2320**
	Promedio	**244**	**234**	**10**	**130**	**464**

MAE (Mean Absolute Error)

Error absoluto promedio entre los valores reales y los estimados por el modelo.

El MAE mide, en promedio, cuánto se equivoca el modelo, sin importar si el error fue por exceso o por defecto.

$$\text{MAE} = \frac{\sum | e(t) |}{n} = \frac{50}{5} = 10$$

En promedio, el modelo se equivoca por 10 unidades (positivas o negativas) al predecir el valor real.

Interpretación - Análisis de regresión

n	Mes	Valor real	Valor estimado	Error absoluto |e(t)|	Error cuadrático e(t)^2	(Valor real - promedio(valor real))^2
1	Ene	210	200	10	100	1156
2	Feb	250	245	5	25	36
3	Mar	230	210	20	400	196
4	Abr	270	260	10	100	676
5	May	260	255	5	25	256
	Suma	**1220**	**1170**	**50**	**650**	**2320**
	Promedio	**244**	**234**	**10**	**130**	**464**

MSE (Mean Squared Error)

Promedio de los errores al cuadrado entre los valores reales y estimados.

El MSE mide cuánto se equivocó el modelo en promedio, elevando al cuadrado los errores. Esto significa que los errores grandes pesan mucho más que los pequeños.

$$\text{MSE} = \frac{\sum e(t)\ 2}{n} = \frac{650}{5} = 130$$

El modelo se equivoca en promedio 130 unidades al cuadrado. Cuanto menor sea este número, mejor es el ajuste del modelo a los datos

Interpretación - Análisis de regresión

n	Mes	Valor real	Valor estimado	Error absoluto |e(t)|	Error cuadrático e(t)^2	(Valor real - promedio(valor real))^2
1	Ene	210	200	10	100	1156
2	Feb	250	245	5	25	36
3	Mar	230	210	20	400	196
4	Abr	270	260	10	100	676
5	May	260	255	5	25	256
	Suma	**1220**	**1170**	**50**	**650**	**2320**
	Promedio	**244**	**234**	**10**	**130**	**464**

RMSE (Root Mean Squared Error)

El RMSE representa la raíz cuadrada del promedio de los errores al cuadrado.

Mide qué tan lejos, en promedio, están las predicciones de los valores reales, penalizando más fuertemente los errores grandes.

$$\text{RMSE} = \sqrt{MSE} = \sqrt{130} = 11.4$$

El modelo se equivoca en promedio por 11.4 unidades en sus predicciones, considerando el tamaño del error (al cuadrado) y luego volviendo a la misma escala de la variable original.

Interpretación - Análisis de regresión

n	Mes	Valor real	Valor estimado	Error absoluto |e(t)|	Error cuadrático e(t)^2	(Valor real - promedio(valor real))^2
1	Ene	210	200	10	100	1156
2	Feb	250	245	5	25	36
3	Mar	230	210	20	400	196
4	Abr	270	260	10	100	676
5	May	260	255	5	25	256
	Suma	**1220**	**1170**	**50**	**650**	**2320**
	Promedio	**244**	**234**	**10**	**130**	**464**

R2 (Bondad de ajuste)

El R^2 mide qué proporción de la variabilidad del valor real es explicada por el modelo.

Oscila entre 0 y 1, donde 1 indica un modelo perfecto, sin errores, y 0 indica solo errores.

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum e(t)}{\sum (Valor\ real - \bar{y})^2} = 1 - \frac{650}{2320} = 0{,}7198 \approx 72\%$$

El modelo explica aproximadamente el 72% de la variación en los valores reales con base en las variables independientes usadas para predecir.

Software - Análisis de regresión

Todos los indicadores que calculamos paso a paso los entrega el software, por lo que es importante saber cómo interpretarlos (más allá de memorizar fórmulas).

Source	SS	df	MS		Number of obs =	100
					F(1, 98) =	17.47
Model	5.04902329	1	5.04902329		Prob > F =	0.0001
Residual	28.3220135	98	.289000137		R-squared =	0.1513
					Adj R-squared =	0.1426
Total	33.3710367	99	.337081179		Root MSE =	.53759

cholesterol	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
time_tv	.0440691	.0105434	4.18	0.000	.0231461	.0649921
_cons	-2.134777	1.813099	-1.18	0.242	-5.732812	1.463259

En la práctica - Análisis de regresion

¿Qué ocurre en la práctica con el análisis de regresión?

- Corremos varios modelos en paralelo y nos quedamos con el que posee mejores métricas.
- Puede ocurrir que en algunas métricas un modelo es mejor que otro:
  - Comúnmente se prioriza el R2.
  - Muchas veces depende del caso a caso.

Conclusiones

El análisis de regresión es fundamental para implementar modelos de machine learning de clasificación.

Los softwares calculan todo, por lo que es más importante saber cómo interpretar los resultados que memorizar fórmulas.