Regresión Logística

Minería de Datos (CC3074) - 2026

Regresión Logística

Semestre 01, 2026

El problema de clasificar

La regresión lineal predice un valor continuo.

Pero a veces la pregunta no es ¿cuánto? sino ¿cuál?

- ¿Este correo es spam o no?
- ¿Este paciente tiene la enfermedad?
- ¿Este cliente va a cancelar su suscripción?

Regresión Lineal

La regresión lineal puede predecir valores menores a 0 o mayores a 1.

Eso no tiene sentido si queremos una probabilidad.

Necesitamos un modelo que siempre devuelva un valor entre 0 y 1.

La idea central

La regresión logística no predice una clase directamente.

Predice la probabilidad de pertenecer a la clase positiva.

Luego se aplica un umbral para decidir la clase.

Ejemplo

P(spam | características) = 0.87

→ probabilidad alta → clase: spam

P(spam | características) = 0.12

→ probabilidad baja → clase: no spam

La función sigmoide

La función sigmoide transforma cualquier número real en un valor entre 0 y 1.

σ(z) = 1 / (1 + e^−z)

Comportamiento

| z | σ(z) | | --- | --- | | muy negativo | ≈ 0 | | 0 | 0.5 | | muy positivo | ≈ 1 |

- Si z → −∞, la probabilidad tiende a 0.
- Si z → +∞, la probabilidad tiende a 1.
- En z = 0, la probabilidad es exactamente 0.5.

z

z es la combinación lineal de las variables:

z = β0 + β1·X1 + β2·X2 + ... + βn·Xn

El modelo aprende los valores de β que mejor explican los datos.

El modelo completo

$P(y = 1 \mid X) = \sigma(\beta 0 + \beta 1 \cdot X1 + \ldots + \beta n \cdot Xn)$

Donde:

- $P(y = 1 \mid X)$ → probabilidad de la clase positiva
- σ → función sigmoide
- β → coeficientes aprendidos durante el entrenamiento

De probabilidad a clase

Se define un umbral (por defecto: 0.5).

- Si $P \geq 0.5$ → clase positiva (1)
- Si $P < 0.5$ → clase negativa (0)

Umbral

El umbral se puede ajustar según el problema:

- Detección de enfermedades → umbral bajo (preferir falsos positivos)
- Filtro de spam → umbral alto (evitar falsos positivos)

Cambiar el umbral afecta precisión y recall.

Interpretación de los coeficientes

En regresión lineal, $\beta 1 = 2$ significa "Y aumenta 2 por cada unidad de X".

En regresión logística, la interpretación es diferente.

Odds y log-odds

La regresión logística modela el log-odds:

$\log(P / (1 - P)) = \beta_0 + \beta_1 \cdot X_1 + \ldots$

Un coeficiente positivo → aumenta la probabilidad de la clase 1.

Un coeficiente negativo → disminuye la probabilidad de la clase 1.

Ejemplo

Si $\beta_1 = 1.5$ para la variable "edad":

Aumentar la edad en una unidad multiplica los odds por $e^{1.5} \approx 4.5$.

La probabilidad de la clase positiva aumenta.

Función de costo

No se usa el error cuadrático (MSE) porque genera una función no convexa.

Se usa la entropía cruzada binaria (log loss):

$J = -(1/n) \cdot \Sigma [\, y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y}) \,]$

¿Por qué log loss?

- Penaliza fuertemente las predicciones incorrectas con alta confianza.
- Si el modelo predice P = 0.99 y la clase real es 0 → penalización muy alta.
- Si el modelo predice P = 0.51 y la clase real es 0 → penalización baja.

Minimización

El modelo ajusta los coeficientes β usando gradiente descendente para minimizar J.

No existe solución cerrada como en regresión lineal.

Métricas de evaluación

La exactitud (accuracy) no siempre es suficiente.

Si el 95% de los correos no son spam, un modelo que siempre diga "no spam" tiene accuracy = 95%.

Pero no detecta nada útil.

Matriz de confusión

| | Predicho: 1 | Predicho: 0 | | --- | --- | --- | | Real: 1 | VP (verdadero positivo) | FN (falso negativo) | | Real: 0 | FP (falso positivo) | VN (verdadero negativo) |

Precisión

Precisión = VP / (VP + FP)

De todas las predicciones positivas, ¿cuántas fueron correctas?

Alta precisión → pocos falsos positivos.

Recall (Sensibilidad)

Recall = VP / (VP + FN)

De todos los casos reales positivos, ¿cuántos detectamos?

Alto recall → pocos falsos negativos.

F1-Score

F1 = 2 · (Precisión · Recall) / (Precisión + Recall)

- Balance entre precisión y recall.
- Útil cuando las clases están desbalanceadas.
- Va de 0 a 1 — más cercano a 1 es mejor.

¿Cuándo usar cada métrica?

| Prioridad | Métrica clave | | --- | --- | | Minimizar falsos positivos | Precisión | | Minimizar falsos negativos | Recall | | Balance general | F1-Score | | Clases balanceadas | Accuracy |

Supuestos

Independencia de observaciones

- Cada muestra debe ser independiente.
- No aplica directamente a series de tiempo.

No multicolinealidad severa

- Las variables predictoras no deben estar altamente correlacionadas entre sí.
- Afecta la estabilidad de los coeficientes.

Tamaño de muestra suficiente

- Con pocas observaciones, los coeficientes son inestables.

- Regla práctica: al menos 10 observaciones por variable.

Relación lineal con el log-odds

- El modelo asume que la relación entre X y el log-odds es lineal.
- Si no lo es, el modelo puede tener bajo desempeño.

Extensión multiclase

La regresión logística estándar es binaria.

Para más de dos clases se usan dos estrategias:

One-vs-Rest (OvR)

- Se entrena un modelo por cada clase.
- Cada modelo predice la probabilidad de esa clase vs todas las demás.
- Se elige la clase con mayor probabilidad.

Softmax (Regresión Logística Multinomial)

- Un solo modelo que predice probabilidades para todas las clases simultáneamente.
- La suma de todas las probabilidades es siempre 1.
- Más eficiente y coherente que OvR.

Comparación con otros modelos

| Aspecto | Regresión Logística | Árbol de Decisión | KNN | | --- | --- | --- | --- | | Interpretabilidad | Alta | Alta | Baja | | Frontera de decisión | Lineal | No lineal | No

lineal | | Sensible a escala | Sí | No | Sí | | Requiere normalización | Sí | No | Sí | | Velocidad de entrenamiento | Rápida | Rápida | N/A |

Aplicaciones reales

- Detección de spam
- Diagnóstico médico (¿tiene la enfermedad?)
- Predicción de abandono (churn)
- Aprobación de crédito
- Clasificación de sentimientos (positivo / negativo)

La regresión logística es un buen punto de partida para cualquier problema de clasificación binaria.

¿Qué puede salir mal?

- Variables no escaladas → coeficientes difíciles de comparar
- Clases muy desbalanceadas → el modelo aprende a ignorar la clase minoritaria
- Variables altamente correlacionadas → coeficientes inestables
- Relación no lineal → modelo con bajo desempeño en datos complejos