



Inteligencia Artificial Aplicada para la Economía



Profesores

Profesor Magistral

Camilo Vega Barbosa

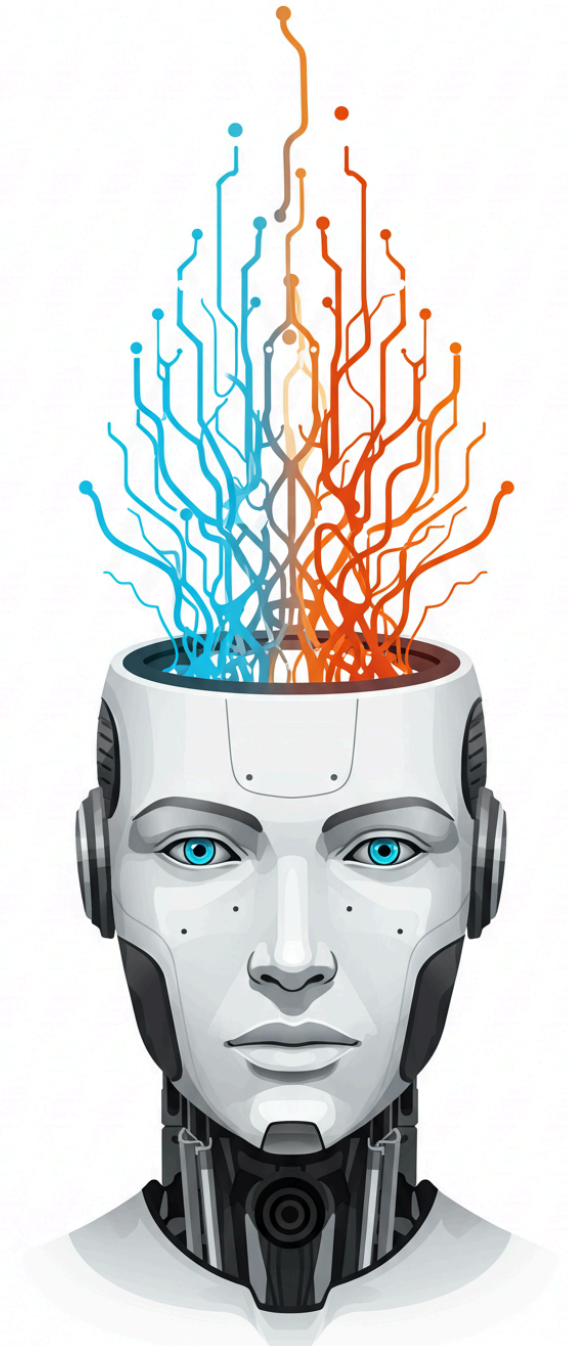
Asistente de Docencia

Daniel Aguirre Salamanca



Regresiones y Regularización en aprendizaje supervisado

Dominando las técnicas fundamentales de
modelado predictivo



Las Regresiones en Aprendizaje Supervisado

Las regresiones son una parte importante del aprendizaje supervisado, siendo herramientas fundamentales para predecir relaciones entre variables. A diferencia de la clasificación que asigna categorías, las regresiones nos permiten predecir valores numéricos continuos.

Tipos principales de regresión

Regresión Lineal

- Modela relaciones lineales entre variables
- Asume una relación directa:
$$y = mx + b$$
- Ideal para predicciones de valores continuos como precios o demanda

Regresión Logística

- A pesar de su nombre, es un modelo de clasificación
- Predice probabilidades usando la función logística:

$$P(y = 1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1x)}}$$

- Transforma predicciones en probabilidades entre 0 y 1

Métricas de Evaluación para Regresiones

Error Cuadrático Medio (MSE)

- Promedio de errores al cuadrado

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Penaliza fuertemente errores grandes
- Siempre positivo, 0 es perfecto

Raíz del Error Cuadrático Medio (RMSE)

- Raíz cuadrada del MSE

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

- Mismas unidades que la variable objetivo
- Más interpretable que MSE

Métricas de Evaluación para Regresiones

Error Absoluto Medio (MAE)

- Promedio de errores absolutos

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

- No penaliza tanto errores grandes
- Robusto a valores atípicos

Coeficiente de Determinación (R^2)

- Proporción de varianza explicada

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

- Varía entre 0 y 1 (1 es perfecto)
- Interpretable como porcentaje

El Enfoque de Machine Learning en Regresiones

A diferencia de campos como la estadística o la econometría, el machine learning aborda las regresiones con un enfoque distinto:




Enfoque Tradicional

- Soluciones numéricas directas
- Mínimos cuadrados ordinarios (OLS)
- Foco en inferencia y causalidad
- Énfasis en significancia estadística

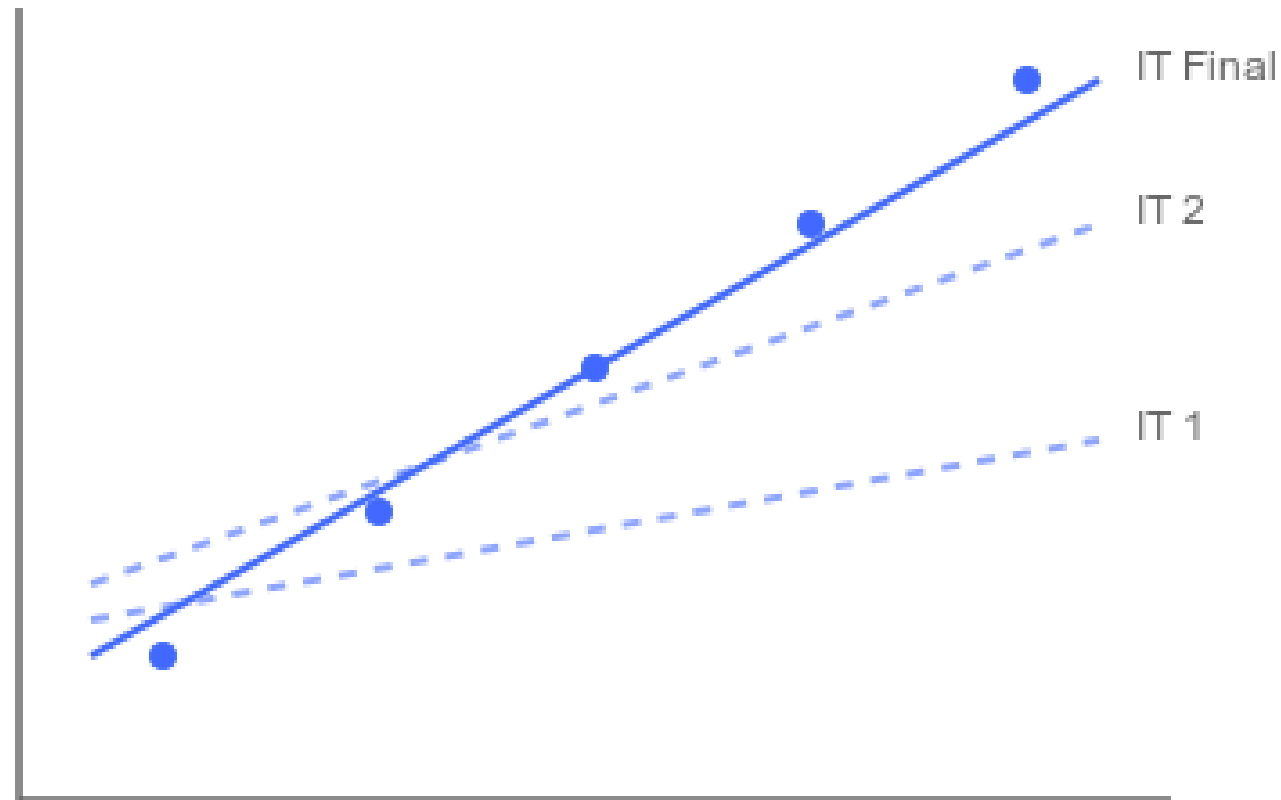


Enfoque ML

- Optimización iterativa
- Descenso del gradiente
- Foco en predicción y generalización
- Énfasis en métricas de error

La clave está en la escalabilidad : el enfoque iterativo permite manejar conjuntos de datos masivos y modelos más complejos.

Optimización Iterativa en Regresión



El modelo mejora iterativamente, ajustando sus parámetros hasta encontrar la línea que mejor se ajusta a los datos.

Proceso de Entrenamiento en Regresiones

El entrenamiento de una regresión es un proceso secuencial donde buscamos los coeficientes óptimos (β) que minimicen el error en nuestras predicciones.

1. Entrenamiento del Modelo

- Partimos de nuestro conjunto de entrenamiento ($X_{\text{train}}, y_{\text{train}}$)
- El modelo busca los coeficientes que minimicen el error:

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n (y_i - X_i \beta)^2$$

Proceso de Entrenamiento en Regresiones

2. Evaluación en Prueba

- Usamos los coeficientes encontrados (β) para predecir:


$$\hat{y}_{test} = X_{test}\hat{\beta}$$


- Calculamos el error en datos nunca vistos:

$$MSE_{test} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2$$

Este proceso nos permite validar si los coeficientes realmente generalizan bien a datos nuevos.

Regresión Logística como Clasificador

 La regresión logística, a pesar de su nombre, es una herramienta poderosa para problemas de clasificación binaria. Su verdadera fortaleza radica en que no solo clasifica, sino que nos da la probabilidad de pertenencia a cada clase.

 Esta característica es especialmente útil en aplicaciones donde necesitamos entender la certeza de nuestras predicciones, como en:

- Detección de fraude en transacciones
- Predicción de riesgo crediticio
- Diagnóstico médico
- Predicción de abandono de clientes

De Probabilidades a Clases con Regresión Logística

Proceso

- La regresión logística predice $P(y=1|x)$
- Definimos un umbral (típicamente 0.5)
- Si $P(y=1|x) \geq \text{umbral} \rightarrow \text{Clase 1}$
- Si $P(y=1|x) < \text{umbral} \rightarrow \text{Clase 0}$

Ejemplo

- Predicción de riesgo crediticio
- $P(\text{default}=1) = 0.7$
- Umbral = 0.5
- Como $0.7 > 0.5 \rightarrow \text{Alto riesgo}$

Nota: La elección del umbral puede ajustarse según el balance deseado entre falsos positivos y negativos.

El Problema del Sobreajuste y la Regularización

El Dilema del Aprendizaje:

Nuestros modelos pueden caer en la trampa de memorizar en lugar de aprender. Es como un estudiante que se aprende las respuestas de memoria, pero no entiende realmente la materia.

La Solución :

La regularización actúa como un "presupuesto" para nuestro modelo: cada variable tiene un costo. Al igual que compramos solo lo necesario cuando tenemos un presupuesto limitado, el modelo aprende a usar solo las variables verdaderamente importantes.

Ridge y Lasso: Dos Enfoques de Regularización 🔍

Regresión Ridge (L2)

- Añade término cuadrático de penalización:
$$\min_{\beta} \sum (y_i - X_i\beta)^2 + \lambda \sum \beta_j^2$$
- Reduce coeficientes pero no los elimina
- Ideal cuando hay multicolinealidad

Regresión Lasso (L1)

- Usa penalización en valor absoluto:
$$\min_{\beta} \sum (y_i - X_i\beta)^2 + \lambda \sum |\beta_j|$$
- Puede eliminar variables completamente
- Excelente para selección de características

Regularización en Clasificación

Mejorando el Logit:

La regresión logística regularizada combina lo mejor de dos mundos: la capacidad de clasificar con la robustez de la regularización.

$$\underbrace{\log L(\beta)}_{\text{Logit original}} - \underbrace{\lambda \|\beta\|}_{\text{Penalización}} = \sum_{i=1}^n [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] - \lambda \|\beta\|$$

Clasificación:

- Ajustamos el umbral de decisión (default: 0.5)
- La regularización nos da probabilidades más robustas
- Pero, podemos elegir umbrales según nuestras necesidades (Arriesgado o conservador).

Recursos del Curso

Plataformas y Enlaces Principales

 **GitHub del curso**

 github.com/CamiloVga/IA_Aplicada

 **Asistente IA para el curso**

 **Google Notebook LLM**