SISTEMAS INTELIGENTES

Lecture 03: Algoritmos de Regresión

Dr. Edwin Valencia Castillo
Departamento de Sistemas
Facultad de Ingeniería
Universidad Nacional de Cajamarca
2024

Introducción

Este tema profundiza en los algoritmos de aprendizaje supervisado, haciendo énfasis en los algoritmos de regresión, los cuales realizan una estimación o predicción de una variable numérica o cuantitativa.

El objetivo es obtener una mayor comprensión sobre estos algoritmos, utilizar las métricas de error más comunes y ser capaz de visualizar los errores de forma gráfica.

Algoritmos De Regresión

Definición

Los algoritmos de regresión son un tipo concreto de algoritmos de aprendizaje supervisado que consisten en realizar una predicción de una variable numérica o cuantitativa a partir de un vector de mediciones predictoras.

Función de Regresión

La función de regresión ideal es aquella que minimiza el error cuadrático medio (MSE) y se define como la esperanza condicional de la variable respuesta dado el vector de variables predictoras.

1 2 3

Terminología

La variable a predecir se conoce como variable respuesta o variable dependiente. Las mediciones predictoras se conocen como vector de variables predictoras o variables independientes.

$$Y = X_1 + X_2 + X_3$$

Dependent Variable	Independent Variable
Outcome Variable	Predictor Variable
Response Variable	Explanatory Variable

Tipos de Regresión Lineal

✓Regresión lineal simple

- ➢La regresión lineal simple se define mediante la función lineal:

$Y = \beta 0 * X + \beta 1 + \varepsilon$

β0 y β1 son dos constantes desconocidas que representan la pendiente de regresión, mientras que ε (épsilon) es el término de error.

- ➢Puede utilizar la regresión lineal simple para modelar la relación entre dos variables, como las siguientes:
  - Lluvia y rendimiento de los cultivos
  - Edad y estatura en niños
  - Temperatura y expansión del mercurio metálico en un termómetro

Tipos de Regresión Lineal

✓Regresión lineal múltiple

- ➢En el análisis de regresión lineal múltiple, el conjunto de datos contiene una variable dependiente y múltiples variables independientes. La función de línea de regresión lineal cambia para incluir más factores, de la siguiente manera:

$$Y = \beta 0 * x0 + \beta 1 x1 + \beta 2 x2 + \ldots\ldots \beta N x N + \varepsilon$$

A medida que aumenta el número de variables predictivas, las constantes β también aumentan en consecuencia.

- ➢ La regresión lineal múltiple modela múltiples variables y su impacto en un resultado:
  - Lluvia, temperatura y uso de fertilizantes en el rendimiento de los cultivos
  - Dieta y ejercicio sobre enfermedades cardíacas
  - Crecimiento salarial e inflación en las tasas de préstamos hipotecarios

Tipos de Regresión Lineal

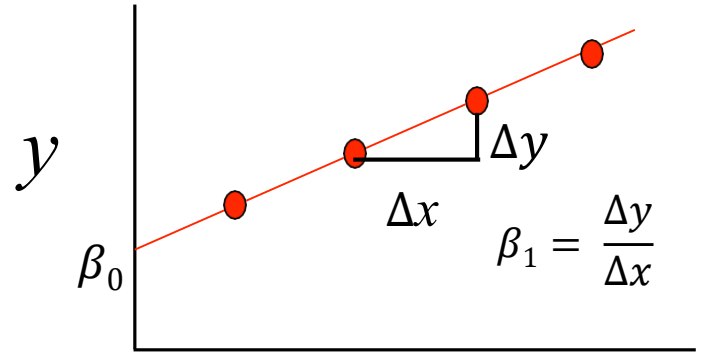
✓Regresión logística

- ➢La regresión logística se utiliza para medir la probabilidad de que se produzca un evento. La predicción es un valor entre 0 y 1, donde 0 indica un evento que es poco probable que ocurra y 1 indica una probabilidad máxima de que suceda. Las ecuaciones logísticas usan funciones logarítmicas para calcular la línea de regresión.
- ➢A continuación, se indican varios ejemplos:
  - La probabilidad de ganar o perder en un partido deportivo
  - La probabilidad de aprobar o reprobar una prueba
  - La probabilidad de que una imagen sea una fruta o un animal

Regresion Lineal Simple

- ✓Gran parte de las matemáticas se dedican al estudio de variables que están determinísticamente relacionadas entre sí

$$y = \beta_0 + \beta_1 x$$



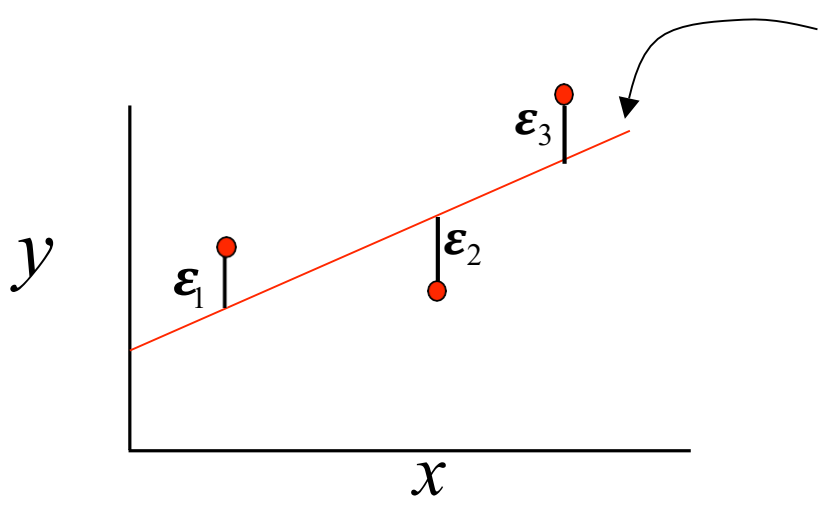
- ✓Pero lo que nos interesa es entender la relación entre variables relacionadas de una manera no determinista

Regresion Lineal Simple

- ✓ Definición: Existen parámetros β_0 , β_1, y , σ^2 tales que para cualquier valor fijo de la variable independiente x, la variable dependiente está relacionada con x a través de la ecuación del modelo:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- ✓ ε ~N(0,σ2) indica que la variable de error ε se distribuye normalmente con media 0 y varianza σ^2. En otras palabras, los errores en el modelo de regresión se asumen normalmente distribuidos alrededor de 0, con una varianza constante σ^2.



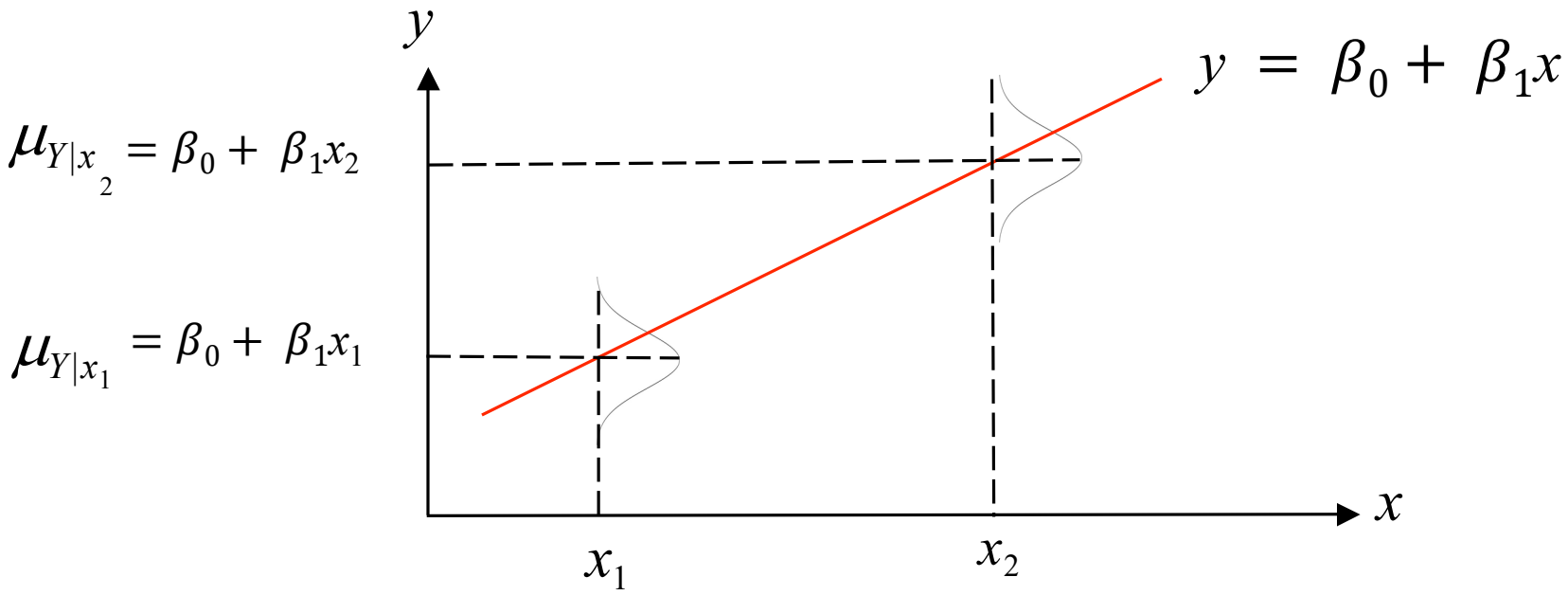
Linea de regression verdadera

$$y = \beta_0 + \beta_1 x$$

Regresion Lineal Simple - Implicaciones

- ✓El valor esperado de Y es una función lineal de X, pero para x fijo, la variable Y difiere de su valor esperado en una cantidad aleatoria
- ✓Formalmente, sea x^* un valor particular de la variable independiente x, entonces nuestro modelo probabilístico lineal dice:
  - ➢$E(Y|x^*) = \mu_{Y|x^*}$ = valor medio de Y cuando x es x^*
  - ➢$E(Y|x^*) = \sigma^2_{Y|x^*}$ = varianza de Y cuando x es x^*

Regresion Lineal Simple – Implicaciones: Interpretación gráfica



- ✓Por ejemplo, si x = altura e y = peso entonces $\mu_{Y|x=60}$ es el peso promedio de todos los individuos de 60 pulgadas de altura en la poblacion

Regresion Lineal Simple – Implicaciones

✓Un ejemplo adicional:

- ➢Supongamos la relación entre la variable independiente altura (x) y la variable dependiente peso (y) se describe mediante un modelo de regresión lineal simple con recta de regresión verdadera

  y = 7.5 + 0.5x y σ=3

Q1: ¿Cuál es la interpretación de β_1 = 0.5?

El coeficiente de x en la ecuación de regresión es 0.5. Esto significa que por cada aumento de 1 unidad en x (peso), se espera que y (altura) aumente en 0.5 unidades.

La desviación estándar σ=3 indica la dispersión de los valores de y alrededor de la línea de regresión. Es decir, los valores de y suelen variar en un rango de aproximadamente 3 unidades alrededor de los valores predichos por la ecuación.

Q2: ¿Si x= 20 cuál es el valor esperado de Y?

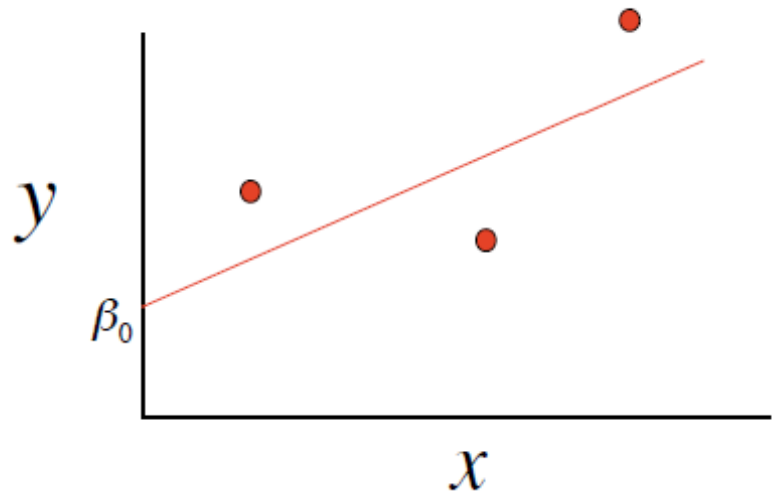
$$\mu_{Y|x=20} = 7.5 + 0.5(20) = 17.5$$

Q3: ¿Si x= 20 cual es P(Y>22)? $P(Y > 22 \mid x = 20) = P\left(\frac{22 - 17.5}{3}\right) = 1 - \phi(1.5) = 0.067$

Regresion Lineal Simple – Estimacion de parámetros del modelo

- ✓ Las estimaciones puntuales de $\hat{\beta}_0$ $\hat{\beta}_1$ se obtienen mediante el principio de los mínimos cuadrados

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$



- ✓ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Regresion Lineal Simple – Valores Predichos y Residuales

- ✓ **Predichos, o ajustados, son valores de y predichos por la línea de regresión de mínimos cuadrados obtenida al introducir $x_1, x_2, \ldots, x_n$ en la línea de regresión estimada**

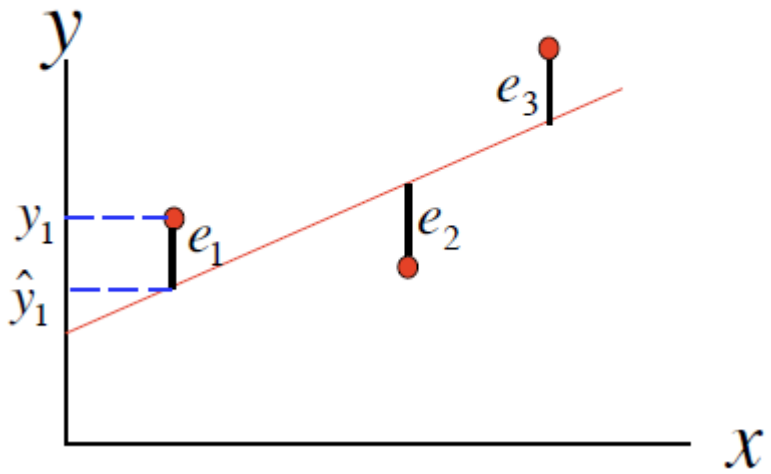
$$\hat{y}_1 = \hat{\beta}_0 - \hat{\beta}_1 x_1$$

$$\hat{y}_2 = \hat{\beta}_0 - \hat{\beta}_1 x_2$$

- ✓ **Residuos son las desviaciones de los valores observados y predichos**

$$e_1 = y_1 - \hat{y}_1$$

$$e_2 = y_2 - \hat{y}_2$$



Regresion Lineal Simple: Metricas de error

- ✓Existen diferentes métricas de error que se pueden aplicar en un problema de regresión para obtener la función f(x) ideal. Las más comunes o habituales y sus definiciones matemáticas son:
  - **Error cuadrático medio, mean square error (MSE): se define como la media de la diferencia entre el valor real y el valor predicho o estimado al cuadrado.**

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

  - **Error absoluto medio, mean absolute error (MAE): se define como la diferencia en valor absoluto entre el valor real y el valor predicho.**

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

Regresion Lineal Simple: Metricas de error

- **Raíz del error cuadrático medio, root mean square Error (RMSE): se define como la raíz cuadrada de la media de la diferencia entre el valor real y el valor predicho o estimado al cuadrado.**

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

  - Esta métrica comparada con el error absoluto medio (MAE) amplifica y penaliza los errores grandes. Por otro lado, en MAE cada error contribuye al total del error en función de su valor absoluto.

- **Logaritmo de la raíz del error cuadrático medio, root mean logarithmic square error (RMLSE):**

$$RMLSE = \sqrt{1/n \sum_{i=1}^{n} (\log(y_i + 1) - \log(\hat{y}_i + 1))^2}$$

  - Esta métrica penaliza una under-prediction más que una over-prediction

Regresion Lineal Multiple

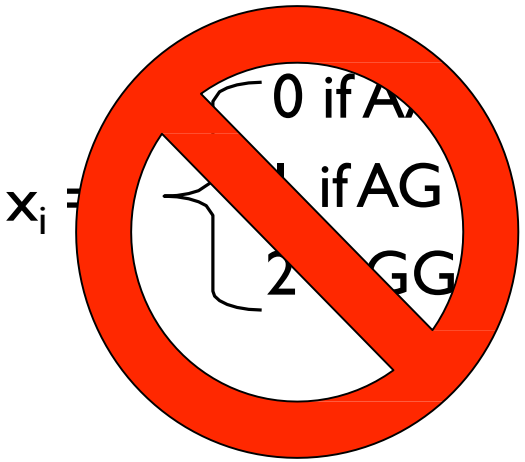
- ✓Es una extensión del modelo de regresión lineal simple a dos o más variables independientes

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

- ✓Coeficientes de Regresión Parciales: $\beta i \equiv$ efecto en la variable dependiente al aumentar la i - ésima variable independiente en 1 unidad, manteniendo constantes todos los demás predictores.
- ✓En resumen, βi nos dice cuánto cambiará la variable dependiente si cambiamos la i-ésima variable independiente en una unidad, suponiendo que todas las demás variables independientes permanecen sin cambios. Este concepto es crucial para entender cómo cada variable independiente afecta la variable dependiente en un contexto donde múltiples variables están influyendo simultáneamente.

Variables categóricas independientes

- ✓Las variables cualitativas se incorporan fácilmente en el marco de regresión a través de variables dummy (ficticias)
- ✓Ejemplo simple: el sexo se puede codificar como 0/1
- ✓¿Qué pasa si mi variable categórica contiene tres niveles?



Variables categóricas independientes

- ✓La codificación previa daría lugar a la colinealidad
- ✓La solución es configurar una serie de variables dummy. En general, para k niveles, se necesitan k-1 variables dummy

$$x_1 = \begin{cases} 1 & \text{if AA} \\ 1 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 2 & \text{if AG} \\ 0 & \text{otherwise} \end{cases}$$

	x_1	x_2
AA	1	0
AG	0	1
GG	0	0

PREGUNTAS??

Dr. Edwin Valencia Castillo
Departamento de Sistemas
Facultad de Ingeniería
Universidad Nacional de Cajamarca
2024