Para estudiantes de Educación Básica y Media
UNIVERSIDAD DE CHILE

Clase 10: Regresión lineal y regularización

FM849 - Programación Científica para Proyectos de Inteligencia Artificial (IA)

9 de enero de 2026

Motivación

En esta clase, estudiaremos uno de los problemas de aprendizaje supervisado que busca aproximar fenómenos que se pueden modelar mediante una función lineal.

Estos fenómenos tienen una particularidad: la variable que nos interesa predecir puede tomar infinitos valores.

Pensemos en algunos ejemplos:

- Entender en salud cómo algunos factores pueden influir sobre el valor que toma la presión arterial de un paciente.
- Entender en campañas de marketing cómo la publicidad en distintos medios (TV, redes sociales, etc.) puede influir sobre el volumen de ventas.

Fundamentos de la regresión lineal

Tratemos de entender las regresiones lineales intuitivamente. Para ello, imaginemos que queremos vender una consola de videojuegos.

Por ahora, juntamos información sobre las siguientes variables:

- Cantidad de juegos físicos o digitales en unidades (x_1).
- Número de accesorios incluidos (p. ej., cámaras, controles, etc.) en unidades (x_2).
- Meses de uso (x_3).
- Estado, en una escala del 1 al 10, donde 1 significa que tiene muchos imperfectos, y 10 significa que está casi nueva (x_4).

Fundamentos de la regresión lineal

Si quisiéramos modelar linealmente el precio (y) en función de las variables descritas anteriormente ($X = (x_1, x_2, x_3, x_4)$), la forma de hacerlo es incorporando coeficientes que multipliquen a cada variable, tal y como se hace en las funciones lineales. Estos coeficientes los escribimos con la letra β:

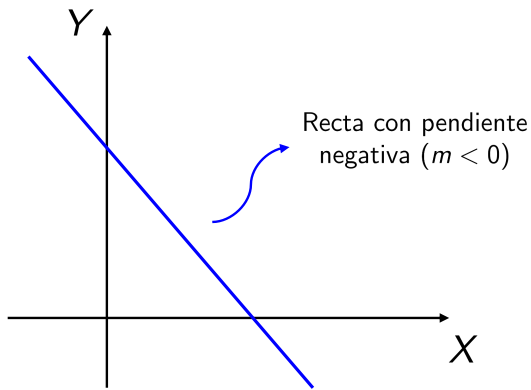
$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot x_4$$

Para entender los efectos de cada variable, pensemos específicamente en qué pasa cuando aumenta x_3 (meses de uso). ¿Debería subir o bajar el precio?

Asignación de coeficientes

En el caso específico de los meses de uso, si aumentan, obviamente debería bajar el precio final, porque la consola pierde valor. La forma en que esto se refleje en la ecuación de la recta es asignando un coeficiente negativo a β_3 (que acompaña a x_3).

Esto es porque gráficamente, si pensamos en sólo una variable, la recta $y = mx + n$ (función lineal afín) es estrictamente decreciente (que es lo que buscamos, porque queremos que el precio baje cuando los meses suban), si y sólo si la pendiente m es negativa:



Relación entre variables

Otro de los factores que debemos tener en cuenta es que las variables pueden relacionarse entre sí.

Tomando el mismo ejemplo, podemos notar que intuitivamente, debiese existir una relación directa entre los meses de uso (x_3) y el estado de la consola (x_4), porque a mayor uso, suele estar en peor estado.

En un modelo de regresión lineal, no queremos que hayan variables que se expliquen perfectamente entre sí, porque puede llevar a resultados erróneos. Esto tiene una explicación matemática que no profundizaremos, porque necesitan más herramientas y no es el objetivo del curso.

Detengámonos un momento...

Nosotros planteamos un modelo de regresión lineal que considera variables que suenan relevantes, pero ¿es realmente un buen modelo?

- ¿Por qué no separamos los juegos físicos y digitales en dos variables? Puede que los juegos físicos sean más solicitados.
- ¿Realmente todos los juegos valen lo mismo?
- Lo mismo ocurre con los accesorios...
- ¿Hay alguna variable que no hayamos considerado y podamos medir?

Estos cuestionamientos separan a un buen científico de datos de alguien que simplemente aplica recetas. Anteriormente, hablamos de la importancia de entender el dominio del problema, y los efectos negativos que puede tener en la sociedad el mal uso de estas técnicas.

¿En qué otros factores deberíamos fijarnos?

La regresión lineal tiene **supuestos** que debemos cumplir para que los resultados sean confiables. Estos son los siguientes:

1. El modelo debe ser lineal en los parámetros (coeficientes). Esto no se exige para las variables, dado que se pueden transformar (p. ej., $\sqrt{x}$, $\log(x)$, etc.).

Ejemplo

La recta $y = \beta_0 + \beta_1 x^2$ cumple este supuesto, porque podemos aplicar la transformación $\varphi(x) = \sqrt{x}$ y escribir $y = \beta_0 + \beta_1 \varphi(x)$.

La recta $y = \beta_0 + \beta_1 \beta_2 x$ no cumple este supuesto, porque $\beta_1 \beta_2$ no es lineal.

¿En qué otros factores deberíamos fijarnos?

2. La muestra debe ser aleatoria. Si no es así, los resultados pueden estar sesgados, por la baja variabilidad que existe (dilema sesgo-varianza).
3. No debe existir multicolinealidad perfecta entre las variables explicativas (variables que se expliquen completamente entre sí).
4. No debe existir ninguna relación entre el error (denotado por ε) y las variables explicativas.

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \boxed{\varepsilon} \leftarrow \text{residuo no medible}$$

Pueden pensar ε como todo aquello que no puede ser medido o explicado por las variables que tenemos.

¿En qué otros factores deberíamos fijarnos?

5. Por último, los errores deben tener varianza constante. Esto significa que la dispersión de los errores debe ser similar a lo largo de todo el rango de valores predichos por el modelo.

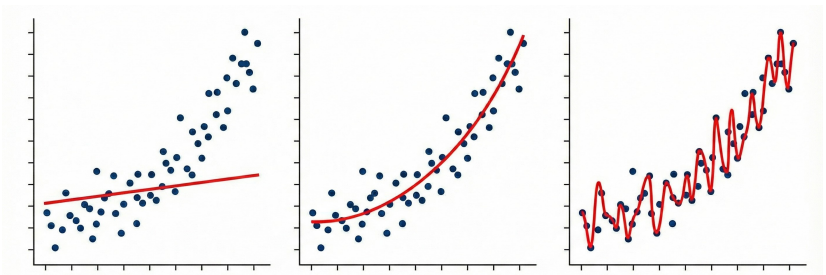
Ejemplo

Una consola vieja podría tener más variación en el precio que una nueva. En ese escenario, no se cumple el último supesto.

¿Por qué podría pasar esto? Piensen que las consolas, si son viejas, pueden ya no ser tan útiles para el dueño/a y podría venderla por necesidad. Por otro lado, alguien podría pensar que tiene más valor porque es de colección. En ambos casos, aumenta mucho la variabilidad del precio con respecto a las nuevas, que se venden a un precio más estable.

Regularización

En la clase pasada hablamos de los problemas de ajuste. Las rectas también lo tienen. La interpolación de Lagrange demuestra que cualquier conjunto de puntos se puede ajustar perfectamente con un polinomio, lo que no es deseable, porque llevará a que el modelo sólo será bueno en el conjunto de entrenamiento.



Underfitting
(ajuste de mala calidad)

Buen *fit*

Overfitting
(ajuste de "memorización")

Regularización

La opción por excelencia para evitar el sobreajuste en regresiones lineales es la **regularización**, que consiste en agregar un término de penalización al aprendizaje que busca evitar que los coeficientes (β) tomen valores muy grandes.

Existen tres tipos principales de regularización:

- **Ridge** (L_2): penaliza el cuadrado de los coeficientes.
- **Lasso** (L_1): penaliza el valor absoluto de los coeficientes.
- **Elastic Net**: combinación de las dos anteriores.

La elección de la regularización depende del problema, pero en general, Ridge es más utilizada cuando se tienen muchas variables correlacionadas, mientras que Lasso es útil para selección de variables, ya que puede reducir algunos coeficientes a cero.