Fundamentos de Redes Neuronales y Descenso del Gradiente

UNIVERSIDAD DE CÓRDOBA.

Andrés Felipe Therán y Daniel José Ortiz

1. Características del Dataset

a) ¿Cuál es el valor de n (número de registros)?

El número de registros contenidos en el dataset es de 20,640.

b) ¿Cuántas características tiene cada observación?

Cada observación tiene 8 características.

c) Según la notación del capítulo, escriba el vector de características x para la primera observación del dataset en forma de columna.

$$x_0 = \begin{bmatrix} 8.3252 \\ 41 \\ 6.984 \\ 1.0238 \\ 322 \\ 2.555 \\ 37.88 \\ -122.23 \end{bmatrix}$$

2. Descripción de Características

El dataset incluye las siguientes 8 características:

- **MedInc:** Ingreso medio del bloque
- **HouseAge:** Edad media de las casas del bloque
- **AveBedrms:** Número promedio de habitaciones por casa
- **AveRooms:** Número promedio de salas por casa
- **Population:** Población por bloque
- **AveOccup:** Número promedio de miembros por casa
- **Latitude:** Latitud del bloque
- **Longitude:** Longitud del bloque

3. Variable Objetivo

¿Qué variable estamos tratando de predecir (etiqueta y)? ¿Qué unidades tiene?

La etiqueta **y** representa el valor medio por casa para los distritos en California, teniendo unidades de cientos de miles de dólares.

4. Tipo de Problema

Según la taxonomía del capítulo, ¿este problema es de regresión o clasificación? Justifique su respuesta.

El problema es de **regresión** debido a que la variable objetivo está distribuida en un rango de valores continuo.

5. Normalización de Características

d) ¿Por qué es importante normalizar las características antes de entrenar una red neuronal?

Si las características se encuentran en escalas muy diferentes, una tasa de aprendizaje que funciona bien para actualizar un peso podría ser demasiado grande o demasiado pequeña para actualizar el otro peso igualmente bien. La normalización asegura que todas las características contribuyan de manera equilibrada al proceso de descenso del gradiente.

e) ¿Qué hace exactamente StandardScaler? Escriba la fórmula matemática.

La estandarización desplaza la media de cada característica para que se centre en cero y cada característica tenga una desviación estándar de 1 (varianza unitaria). La fórmula es:

$$z = (x - \mu) / \sigma$$

f) ¿Por qué usamos fit_transform en el conjunto de entrenamiento pero solo transform en el de prueba?

El método *fit* se utiliza para aprender los parámetros (media y desviación estándar) a partir de los datos de entrenamiento, y el método *transform* utiliza esos parámetros para transformar los datos. Aplicar *fit* sobre los datos de prueba causaría **fuga de datos** (data leakage), comprometiendo la validez de la evaluación del modelo.

6. Inicialización de Pesos

g) ¿Por qué es conveniente inicializar los pesos con valores aleatorios pequeños en lugar de ceros?

Si todos los pesos se inicializan a cero, el parámetro de la tasa de aprendizaje (η) solo afectaría a la escala del vector de peso, no a la dirección. Esto rompe la **simetría** necesaria para que las neuronas aprendan diferentes características.

La inicialización aleatoria pequeña permite que cada peso evolucione de manera independiente.

h) Según la notación del capítulo, ¿qué forma debe tener el vector w?

El vector **w** tiene **n** filas (una por cada característica) y solo una columna, es decir, forma (n, 1).

i) ¿Por qué el sesgo b se inicializa típicamente en cero mientras los pesos no?

El sesgo se puede inicializar en cero porque no afecta la simetría del modelo. Dado que las características han sido normalizadas (centradas en cero), el sesgo comenzará en cero y se ajustará durante el entrenamiento para compensar cualquier desviación en la predicción.

7. Predicción y Función de Activación

j) Explique por qué usamos X @ w en lugar de w^T @ X cuando X tiene forma (m, n).

La operación *w^T @ X* no es posible debido a incompatibilidad dimensional. Para multiplicar matrices, el número de columnas de la primera debe igualar el número de filas de la segunda. Siendo *w^T* de dimensión (1, 8) y *X* de dimensión (20640, 8), no se pueden multiplicar. En cambio, *X @ w* sí es posible: X(20640, 8) × w(8, 1) = (20640, 1).

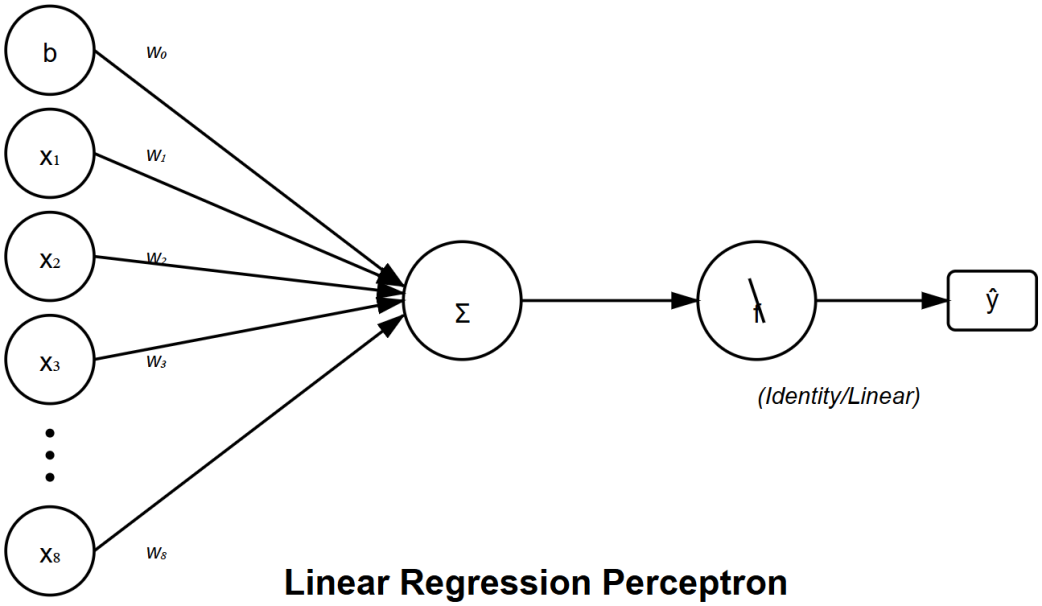
k) En un problema de regresión como este, ¿qué función de activación se usa? ¿Por qué?

La función de activación es **lineal** (identidad) ya que la manera en que se asignan valores a la variable objetivo se asimila a la del modelo de regresión lineal. No se aplica una transformación no lineal porque queremos predecir valores continuos en todo el rango real.

l) ¿Son buenas las predicciones iniciales? ¿Por qué?

Las predicciones son poco acertadas porque los pesos inicializados son aleatorios, lo que hace que sean muy diferentes a los valores óptimos reales. El modelo aún no ha aprendido los patrones en los datos.

m) Dibuje un diagrama mostrando el flujo de datos.



Linear Regression Perceptron

f(z) = z, where z = $w_0 \cdot b + w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + ... + w_8 \cdot x_8$

8. Función de Pérdida MSE

n) ¿Por qué elevamos al cuadrado las diferencias en lugar de usar el valor absoluto?

Elevar al cuadrado es consistente con el proceso del descenso del gradiente. El MSE es una función **diferenciable** en todos los puntos (el valor absoluto no lo es en cero), lo que facilita el cálculo del gradiente. Además, penaliza más fuertemente los errores grandes.

p) ¿Qué ventaja tiene el MSE para el descenso del gradiente comparado con el MAE?

La ventaja del MSE sobre el MAE es que es una **función convexa**, lo que garantiza que todo mínimo local es un mínimo global. Esto asegura convergencia al óptimo cuando se usa descenso del gradiente con una tasa de aprendizaje apropiada.

q) ¿Qué significa que la función de pérdida se "estabilice"?

Que la función de pérdida se estabilice significa que el gradiente con respecto a los parámetros es cercano a cero, indicando que los parámetros han alcanzado un **mínimo**. El modelo ha **convergido** y las actualizaciones adicionales no producen mejoras significativas.

9. Cálculo del Gradiente

r) Demuestre que: $\partial L/\partial w = (2/m)\ X^T\ (\hat{y} - y)$

$$\frac{\partial L}{\partial w_j} = \frac{\partial}{\partial w_j}\frac{1}{n}\sum_i \left(y^{(i)} - \sigma\left(z^{(i)}\right)\right)^2 = \frac{1}{n}\frac{\partial}{\partial w_j}\sum_i \left(y^{(i)} - \sigma\left(z^{(i)}\right)\right)^2$$

$$= \frac{2}{n}\sum_i \left(y^{(i)} - \sigma\left(z^{(i)}\right)\right)\frac{\partial}{\partial w_j}\left(y^{(i)} - \sigma\left(z^{(i)}\right)\right)$$

$$= \frac{2}{n}\sum_i \left(y^{(i)} - \sigma\left(z^{(i)}\right)\right)\frac{\partial}{\partial w_j}\left(y^{(i)} - \sum_i \left(w_j x_j^{(i)} + b\right)\right)$$

$$= \frac{2}{n}\sum_i \left(y^{(i)} - \sigma\left(z^{(i)}\right)\right)\left(-x_j^{(i)}\right) = -\frac{2}{n}\sum_i \left(y^{(i)} - \sigma\left(z^{(i)}\right)\right)x_j^{(i)}.$$

$$= \frac{2}{n}\sum_i \left(y^{(i)} - \sigma\left(z^{(i)}\right)\right)\frac{\partial}{\partial w_j}\left(y^{(i)} - \sum_i \left(w_j x_j^{(i)} + b\right)\right)$$

$$= \frac{2}{n}\sum_i \left(y^{(i)} - \sigma\left(z^{(i)}\right)\right)\left(-x_j^{(i)}\right) = -\frac{2}{n}\sum_i \left(y^{(i)} - \sigma\left(z^{(i)}\right)\right)x_j^{(i)}.$$

$$\nabla_w \mathrm{MSE}(w) = \begin{pmatrix} \frac{\partial}{\partial w_0}\mathrm{MSE}(w) \\ \frac{\partial}{\partial w_1}\mathrm{MSE}(w) \\ \vdots \\ \frac{\partial}{\partial w_n}\mathrm{MSE}(w) \end{pmatrix} = \frac{2}{m}X^{\mathrm{T}}(Xw - y)$$

Demuestre también que: $\partial L/\partial b = (2/m)\ \Sigma(\hat{y}_i - y_i)$

$$\frac{\partial L}{\partial b} = \frac{1}{n}\frac{\partial}{\partial b}\sum_i \left(y^{(i)} - z^{(i)}\right)^2$$

$$= \frac{2}{n}\sum_i \left(y^{(i)} - z^{(i)}\right)\frac{\partial}{\partial b}\left(y^{(i)} - z^{(i)}\right)$$

$$= \frac{2}{n}\sum_i \left(y^{(i)} - z^{(i)}\right)\left(-\frac{\partial z^{(i)}}{\partial b}\right)$$

$$= \frac{2}{n}\sum_i \left(y^{(i)} - z^{(i)}\right)(-1)$$

$$= -\frac{2}{n}\sum_i \left(y^{(i)} - z^{(i)}\right) = \frac{2}{n}\sum_i \left(z^{(i)} - y^{(i)}\right)$$

t) ¿Por qué el gradiente "señala la dirección de máxima pendiente ascendente"?

El gradiente mide y señala el cambio de cualquier función con respecto a una variable. Matemáticamente, el gradiente apunta en la dirección donde la función **crece más rápidamente**. En la metáfora de la montaña, sería la dirección más empinada hacia arriba, hacia un máximo local.

u) ¿Qué significa el signo negativo en $w \leftarrow w - \eta\nabla L$?

El signo negativo invierte la dirección del gradiente, haciendo que nos movamos en la dirección de **máxima pendiente descendente**. Así, la regla de actualización **minimiza** la pérdida con cada iteración, acercándonos al valle (mínimo) en lugar de la cima.

10. Actualización de Parámetros

w) ¿Por qué dw debe tener la misma forma que w?

Cuando se efectúa una resta de vectores, es necesario verificar que tengan la misma cantidad de componentes y la misma forma. La operación $w - \eta \times dw$ requiere que ambos vectores sean compatibles dimensionalmente.

x) ¿Qué sucede con los gradientes si todas las predicciones son exactamente iguales a los valores reales?

Los vectores gradiente pasan a convertirse en **vectores nulos** (todos sus componentes son cero). Esto indica que hemos alcanzado el óptimo perfecto y no hay más actualizaciones necesarias.

11. Tasa de Aprendizaje

z) ¿Qué es la tasa de aprendizaje (η)? ¿Qué problemas pueden surgir?

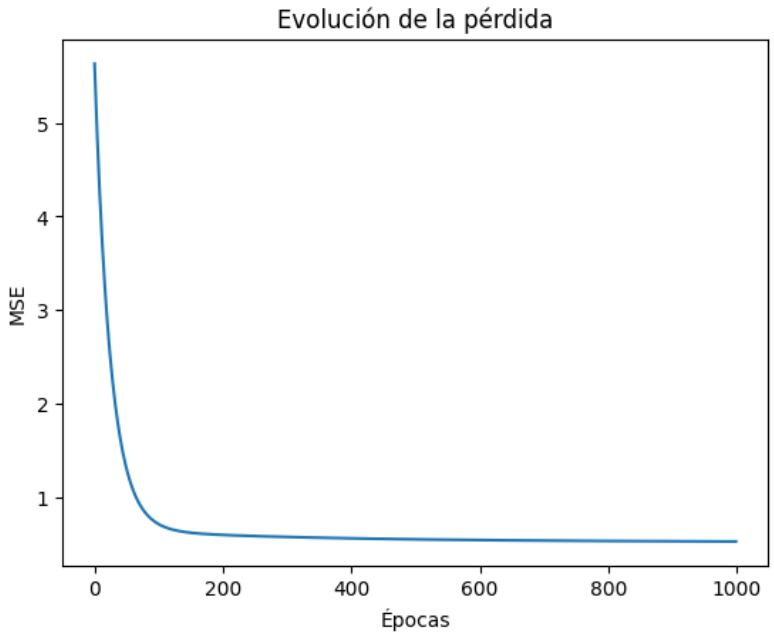
La tasa de aprendizaje controla el **tamaño del paso** en cada actualización del descenso del gradiente.

- **Si es demasiado pequeña:** el descenso del gradiente será muy lento, requiriendo muchas iteraciones para converger.
- **Si es demasiado grande:** puede sobrepasarse el mínimo, oscilar, no converger, o incluso divergir.

12. Resultados y Evaluación

cc) Entrene el modelo con learning_rate=0.01 y epochs=1000. Grafique el historial de pérdida (época vs MSE).

rate=0.01, epochs=1000

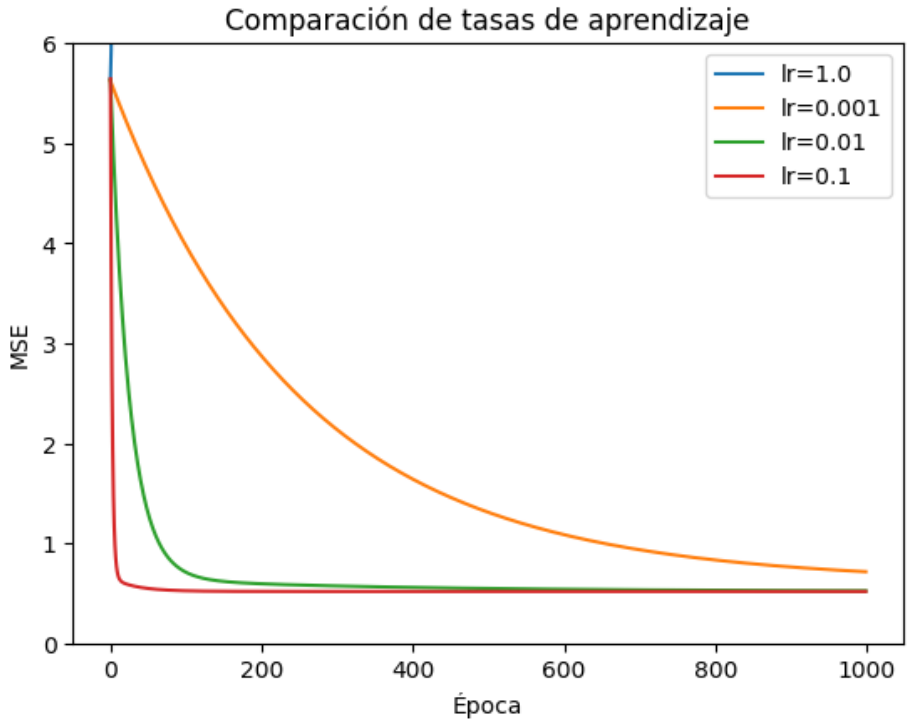


dd) ¿El modelo convergió? ¿Cómo lo sabe observando la gráfica?

Sí, el modelo convergió porque la función de pérdida tiende y permanece en un valor cercano a un mínimo constante, sin cambios significativos en iteraciones posteriores.

ee) Experimente con diferentes tasas de aprendizaje: 0.001, 0.01, 0.1, 1.0. ¿Cuál funciona mejor?

Dentro del rango considerado, la tasa de aprendizaje con la mejor evolución de la función de pérdida es **0.1**. La función de pérdida con la tasa de **1.0 diverge**, confirmando que una tasa demasiado alta causa inestabilidad.



gg) Calcule el MSE en el conjunto de prueba. ¿Qué indica sobre la generalización?

Resultados:

- **MSE de entrenamiento:** 0.5245
- **MSE de prueba:** 0.5544

El valor del MSE en el conjunto de prueba es **ligeramente mayor** al de entrenamiento. Esto indica que el modelo tiene un desempeño similar en datos no vistos, mostrando **buena generalización**. La diferencia pequeña sugiere que no hay sobreajuste significativo, aunque podría indicar que el modelo es algo simple (underfitting leve) y podría beneficiarse de más complejidad.

hh) Cree un scatter plot comparando y_test vs predicciones. ¿Qué observa?

Los puntos tienden a agruparse en cierto segmento de la línea diagonal y = x, lo que implica que no están perfectamente distribuidos.