STOCHASTIC GRADIENT DESCENDENT

sabato 19 ottobre 2024 16:37

RIPETIAMO LA FORMULA AL PASSO J

$$\theta_j := \theta_j - \alpha \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) x_j^{(i)}$$

$$j = 0, 1 \dots M$$

ALCUNE CONSIDERAZIONI

Questa versione è chiamata stocastica o incrementale perché i parametri vengono aggiornati dopo ogni campione di addestramento, quindi il gradiente è una "approssimazione stocastica" del "vero" gradiente della funzione di costo.

- La soluzione ottimale viene generalmente approssimata più velocemente, ma l'algoritmo potrebbe oscillare attorno ad essa senza mai raggiungere la convergenza (problema del zig-zag). Una soluzione vicina a quella ottimale è spesso accettabile.
- Per dataset di grandi dimensioni, il gradient descent stocastico è preferibile rispetto alla versione batch.
- I parametri vengono continuamente aggiornati. Questo significa che la funzione h utilizzerà valori diversi a ogni iterazione.

DIFFERENZE TRA BATCH GD E STOCHASTIC GD

In Batch Gradient Descent, i parametri vengono aggiornati una volta per ogni epoca, utilizzando tutti i campioni del dataset.

- Funzionamento: Ad ogni iterazione, l'algoritmo calcola il gradiente della funzione di costo rispetto a tutti i dati presenti nel dataset, e solo dopo aver elaborato l'intero dataset aggiorna i parametri θ .

È lento ma tende a convergere in modo più uniforme verso il minimo globale.

In Stochastic Gradient Descent, i parametri vengono aggiornati dopo ogni singolo campione di addestramento, non dopo l'intero dataset.

- Funzionamento: L'algoritmo aggiorna i parametri \theta utilizzando un solo esempio x, y alla volta. Quindi, la regola di

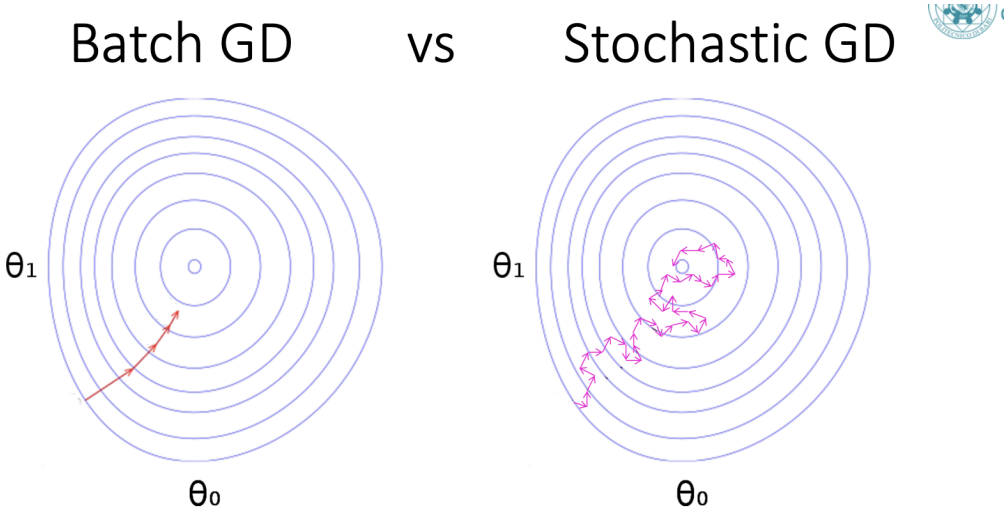
aggiornamento è:

$$\theta_j := \theta_j - 2\left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)}$$

N.B LA DIFFERENZA NEL BATCH GRADIENT TENIAMO LA SOMMATORIA

Riepilogo:

Caratteristica	Batch Gradient Descent	Stochastic Gradient Descent
Aggiornamento dei parametri	Dopo aver elaborato l'intero dataset	Dopo ogni singolo campione
Velocità	Lento su dataset grandi	Molto più veloce
Stabilità del gradiente	Gradiente accurato e stabile	Gradiente rumoroso e approssimato
Convergenza	Convergenza più lenta ma regolare	Convergenza più rapida, ma può oscillare
Adatto per	Dataset piccoli o medi	Dataset molto grandi



Each "jump" is direct toward the minimum, but the relative update step goes through all the examples

Slow, but convergence is sure

Each "jump" may go potentially everywhere, but it is relative to just one training sample

Fast, but convergence is not sure

MINI-BATCH

Nel Mini-batch Gradient Descent, l'algoritmo non aggiorna i parametri dopo ogni singolo campione (come nello Stochastic Gradient Descent), né attende di elaborare l'intero dataset (come nel Batch Gradient Descent). Invece, divide il dataset in piccoli gruppi di dati, chiamati mini-batch, e aggiorna i parametri dopo aver elaborato ogni mini-batch.

Ad esempio, se hai un dataset con 1000 esempi e scegli un mini-batch di dimensione 100, l'algoritmo calcola il gradiente e aggiorna i parametri ogni 100 esempi, invece di attendere tutti i 1000 esempi (Batch) o ogni singolo esempio (Stochastic).