Machine Learning for Data Mining (ML4DM)

GRADIENT DESCENT

Bakul Gohel, Ph.D.
DA-IICT, Gandhinagar
Gradient Descent

• Finding a $\mathbf{w}^*$ such that it give minimum (or maximum) value of function $E(\mathbf{w})$

\[ \mathbf{w}^* = \arg\min E(\mathbf{w}) \]

• Minimize/maximize a function $E(\mathbf{w})$ by altering $\mathbf{w}$

• Most ML algorithms involve optimization in which $E(\mathbf{w})$ often referred as objective or criterion and its minimization also referred to as loss function, cost, or error
Gradient Descent

\[ \argmin E(\mathbf{w}) \]

\[ \mathbf{w}^{k+1} = \mathbf{w}^k - \eta \cdot E'(\mathbf{w}^k) \]

Thus we can reduce \( E(\mathbf{w}) \) by moving \( \mathbf{w} \) in small (…?) steps with opposite sign of its derivative with respect to parameters.
Gradient Descent

\[ w^{k+1} = w^k - \eta \cdot E'(w^k) \]

or

\[ w^{k+1} = w^k - \eta \cdot \nabla E(w^k) \]

or

\[ w \leftarrow w + \Delta w \]

\[ \Delta w = \eta \cdot \nabla E(w) \]

\[ w^* = \text{argmin } E(w) \]

\[ \rightarrow E'(w^*) = 0 \]

\[ \rightarrow |w^{k+1} - w^k| = 0 \]

\text{converse is not true}
Gradient Descent

- Multivariate function $E : \mathbb{R}^M \rightarrow \mathbb{R}$

gradient vector direction at particular point
Gradient Descent

\[ E: \mathbb{R}^M \rightarrow \mathbb{R} \]

\[ w = [w_1, w_1 \cdots w_M]^T \]

\[ \nabla E(w) = \begin{bmatrix} \frac{dE}{dw_1} & \frac{dE}{dw_2} & \cdots & \frac{dE}{dw_M} \end{bmatrix}^T \]

(gradient Direction)

\[ w^{k+1} = w^k - \eta \cdot \nabla E(w^k) \]

\[ w^* = \arg \min E(w) \]

\[ \rightarrow \| E'(w) \| = 0 \]

\[ \rightarrow \| w^{k+1} - w^k \| = 0 \]

converse is not true
Gradient Descent


\[ k \leftarrow 1 \]

**do**

\[ w^{k+1} = w^k - \eta \cdot \nabla E(w^k) \]

\[ k \leftarrow k + 1 \]

**while** \[ \|w^{k+1} - w^k\| > \delta \]

**retrun** \[ w^k \]
Local vs Global minima

$E(w)$

This local minimum performs nearly as well as the global one, so it is an acceptable halting point.

Ideally, we would like to arrive at the global minimum, but this might not be possible.

This local minimum performs poorly and should be avoided.
Saddle Point

• Saddle Points: neither maxima nor minima but gradient is zero...
Learning rate or step size

- Parameter update or learning depends on choice of learning rate

\[ w^{k+1} = w^k - \eta \cdot \nabla E(w^k) \]

- Large learning rate
  - “Overshoots” and “oscillating”

- Small learning rate
  - Extremely slow learning

- Solution
  - Start with large \( \eta \) and settle on optimal value
  - Need a schedule for shrinking \( \eta \)
Gradient Descent in DL

• Different variant of Gradient Descent strategies used in ML
  
  • Stochastic gradient descent
  • Momentum
  • Adagrad
  • RMSprop
  • Adam
  • Etc.