COMP2261 ARTIFICIAL INTELLIGENCE / MACHINE LEARNING

Gradient Descent

-- batch, stochastic, mini-batch

Dr SHI Lei

Durham University

learning lab

Learning Objectives

- Understand three Gradient Descent strategies

Repeat until convergence {

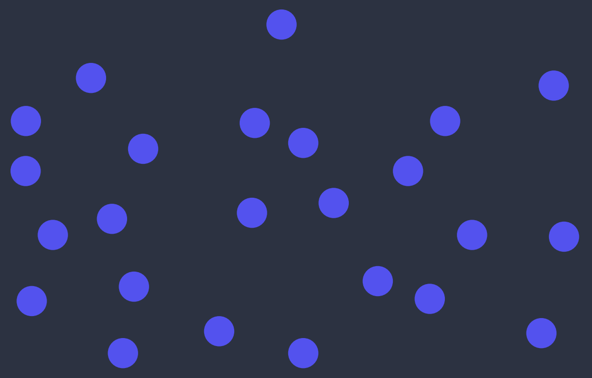
$$\theta_0 := \theta_0 - \alpha \cdot \frac{\partial J}{\partial \theta_0}$$

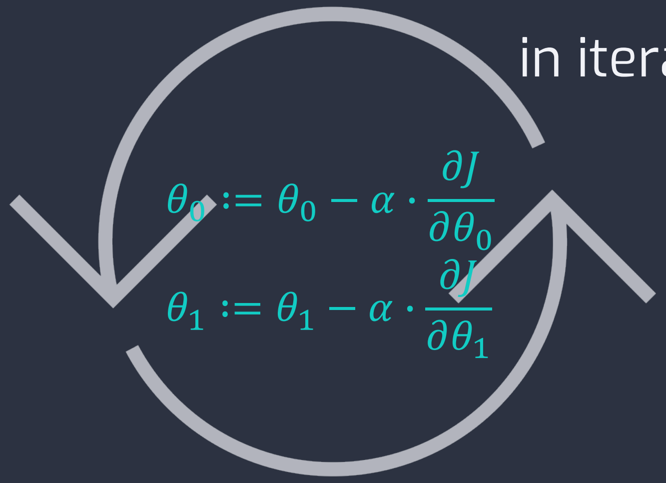
$$\theta_1 := \theta_1 - \alpha \cdot \frac{\partial J}{\partial \theta_1}$$

}

Batch Gradient Descent

- Whole training set used to compute gradient, for each parameter update.
  - Batch -> using the entire batch of training instances.





$$\theta_0 := \theta_0 - \alpha \cdot \frac{\partial J}{\partial \theta_0}$$

$$\theta_1 := \theta_1 - \alpha \cdot \frac{\partial J}{\partial \theta_1}$$

- Drawback…
  - Very slow if the training set is very large.
  - Much worse for high-dimensional problems.
  - Intractable for datasets that don’t fit in memory.

To reduce computational burden and achieve faster iterations…

|0|0|0|0|0|0|0|0
|0|0|0|0|0|0|0|0

Can we not take all instances for each iteration?

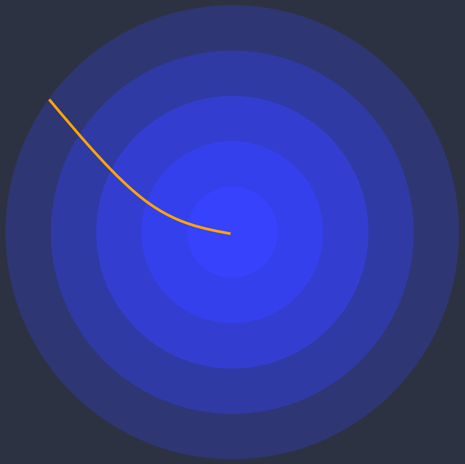
Stochastic Gradient Descent (SGD)

- Only pick randomly one instance from training set, for each parameter update.
  - Stochastic -> random

- SGD performs redundant computations for large datasets.
  - Instances are randomly shuffled and picked for performing iteration.

- Path taken by learning algorithm to reach the minima is usually noisier.
  - It can still reach the minima with significantly shorter training time.

Batch-GD



SGD

Stochastic Gradient Descent (SGD)

- Even though it has reached the minima, it will still keep bouncing around.
  - The final model parameter values may not be the optimal result.

- Decreasing learning rate to reduce noise.
  - As learning rate gets smaller, when algorithm stops, it will be much closer to the minima, the optimal result.

Mini-Batch Gradient Descent

- Use a small random sets of instances (small batch), for each parameter update.
  - Balance between robustness of SGD and efficiency of Batch GD.

Takeaway Points

- Batch Gradient Descent takes the entire dataset
- Stochastic Gradient Descent takes only one instance
- Mini-Batch Gradient Descent, in the between.