1 Optimization

1.1 Batch gradient descent vs. stochastic gradient descent vs. mini batch gradient descent

a. Definition

1) Batch gradient Batch gradient means using all the data point to calculate the gradient.

$cost = \sum_{i=1}^{N}$ -loglikelihood of ith sample

$grad = \frac{\partial(cost)}{\partial \mathbf{w}}$

update all parameter based on gradient

2) Stochastic gradient descent The cost function used in batch gradient descent uses is the summation over all the data points. In stochastic descent the cost function we use only contains one data point, we use one data point to update parameters, iterate over all data points.

For m = 1 : N

cost = - loglikelihood of the ith sample

$grad = \frac{\partial(cost)}{\partial \mathbf{w}}$

update all parameter based on gradient

3) We divide N samples into G = N/k groups so that each group contains k data points

For n = 1 : G

cost = C $\sum_{(n-1)k}^{nk}$ loglikelihood

grad = $\frac{\partial(cost)}{\partial \mathbf{w}}$

update all parameter based on gradient

b. Comparison

	Time per iteration	Convergence time for large data	Sensitivity to parameters	Smoothness
Batch Gradient	Slow for large data	Slower	Moderate	Smooth
Stochastic Gradient	Always fast	Faster	High	Very noisy

c. Practical usage

Shuffle the data before running the stochastic gradient descent

1.2 Newton Method

a. Newton Method Principles

Based on Taylors expansion if we are at x_0, we try to find δx so that $x_0 + \delta x$ is closer to the stationary point.

$$f(x_0 + \delta x) = f(x_0) + f(x_0)\delta x + f(x_0)(\delta x)^2$$

take the derivative

$$df(x_0 + \delta x)/dx = f(x_0) + f(x_0)\delta x$$

therefore

$$\delta x = -\frac{f(x_0)}{f(x_0)}$$
$$X^{(t+1)} = X^t - \frac{f(x_0)}{f(x_0)}$$

b. Matrix Forms

$$x^{(t+1)} = x^t - H^{-1}(f(x^t))f(x^t)$$

where H is the Hessian matrix.

c. Connection with Gradient descent The newton method can be reduce to gradient descent method by taking Hessian matrix as Identity matrix
d. Pros Since it utilizes the second order derivative, it converges much faster than gradient descent.
For quadratic function, the equation from the Taylor expansion is exact, therefore the stationary point can be found using only one step.
e. Cons Need to evaluate the inverse of the Hessian Matrix, so it is computationally expensive.

1.3 Other optimization method

a. Quasi Newton Newton method requires the inverse of the Hessian matrix, which is usually not easy to solve. So we need to find an approximation of the Hessian. Similar to the way we solve for gradient, we can use finite difference method, in which the gradient is

$$gradf(x) = \frac{f(x+\delta x) - f(x)}{\delta x},$$

This is only exact when δx approaches zero. For 2nd order derivative, we can write

$$f^{'}(x) = \frac{f^{'}(x+\delta) - f^{'}(x)}{\delta}$$

Again this is only exact when δ is zero. Based on this idea we replace the Hessian Matrix with an approximation that satisfies the following approximation

$$f(x+\delta x) = f(x) + B\delta x$$

This is quasi newton method. Various Quasi Newton methods exist with different choice of B.

b. Levenburg Marquadt This Method adds a scaled Identity matrix uI to the Hessian, for large u and small Hessian, the method is equivalent to gradient descent with step size 1/u.