Crib 6

06 Nonlinear Least Squares, Gradient Descent

by Alvin Wan . alvinwan.com/cs189/fa17

Note that in the objective functions below, you may choose to featurize your data i.e., replace all x_i with $\phi(x_i)$

1 Nonlinear Least Squares

1. Train a linear model, then fine tune with iterative updates

2 Gradient Descent

1. Take steps along direction of gradient $x_{i+1} = x_i + \eta \nabla_x f(x)$ for learning rate η *(why? see proof in appendix)*
2. η should decrease as a function of i. Commonly used *decay functions*: exponential, step function (e.g., multiply by 0.9 after every 500 steps)

3 Why Step in Direction of Gradient

3.1 Proof

First, why is the gradient the direction of greatest ascent? Take the directional derivative for some loss function f and vector x. For θ, the angle between x and ∇f, we have

$$D_x f = \nabla f \cdot x = |\nabla f||x|\cos(\theta)$$

Note this expression is minimized when $\theta = \pi, \cos(\theta) = -1$. Thus, the direction that decreases f the most, is opposite the gradient vector.

3.2 Why ∇f?

Recall that the gradient step is $x_{i+1} = x_i + \eta \nabla_x f(x)$. Intuitively, the gradient tells us how much change in y occurs, if we perturb x by a little bit. Why does it make sense, then, to update x with "change in y"?

We can look at this another way: Take $f(x + \Delta x)$. Intuitively, we can approximate this point by taking $f(x)$ and extending a tangent line Δx-long. Thus,

$$f(x + \Delta x) \approx f(x) + \langle \nabla f(x), \Delta x \rangle$$

Say we take the gradient step from above, so $\Delta x = -\eta \nabla f(x)$. Then, we have

$$\begin{aligned} f(x + \Delta x) &\approx f(x) + \langle \nabla f(x), -\eta \nabla f(x) \rangle \\ &= f(x) - \eta \langle \nabla f(x), \nabla f(x) \rangle \\ &= f(x) - \eta \|\nabla f(x)\|^2 \\ &\leq f(x) \end{aligned}$$

In other words, taking a gradient step opposite the gradient *tends to* decreases our loss function f.