Comments on Homework Assignments

David Rosenberg

New York University

March 2, 2015

Parameter Tuning

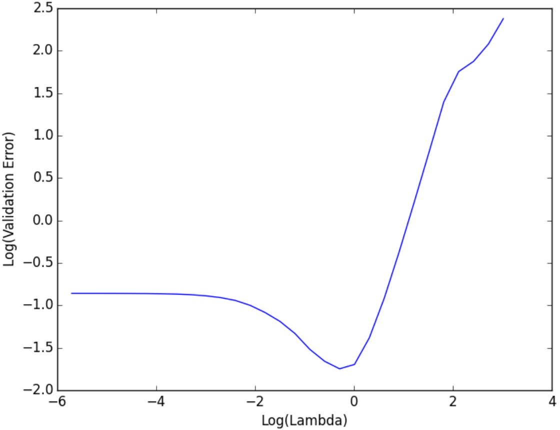
- Can start by trying many different orders of magnitude

$$10^{-5}, 10^{-4}, \ldots, 10^{-1}, 10^{0}, 10^{1}, \ldots, 10^{4}, 10^{5}$$
$$2^{-10}, 2^{-9}, \ldots, 2^{-1}, 2^{0}, 2^{1}, \ldots, 2^{9}, 2^{10}$$

- See where the action is... and zoom in!
- Keep zooming in until things aren't improving on validation set.

Parameter Tuning

- If you want to plot all values on one graph, you may want to take logarithms of your axes.



SGD For Total Loss vs Average Loss

- Suppose we write linear regression objective as

$$J(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

- Then we can do gradient descent using this step direction:

$$-\nabla J(w) = -\sum_{i=1}^{n} 2\left(w^T x_i - y_i\right) x_i$$

- What about stochastic gradient descent?
- Do we just choose a random (x_i, y_i) and step in direction

$$-2\left(w^T x_i - y_i\right) x_i?$$

SGD Step and Gradient Step Should have Same Expectation

- Expectation of gradient step is

$$\begin{aligned}\mathbb{E}\left[-\nabla J(w)\right] &= -\mathbb{E}\left[\sum_{i=1}^{n} 2\left(w^T X_i - Y_i\right) X_i\right] \\ &= -\sum_{i=1}^{n} \mathbb{E}\left[2\left(w^T X_i - Y_i\right) X_i\right] \\ &= -n\mathbb{E}\left[2\left(w^T X - Y\right) X\right]\end{aligned}$$

- Which is n times

$$-\mathbb{E}\left[2\left(w^T X_i - Y_i\right) X_i\right] = -\mathbb{E}\left[2\left(w^T X - Y\right) X\right]$$

- Proper SGD step for this objective is

$$-n \times 2\left(w^T X_i - Y_i\right) X_i$$

- Alternatively, divide original objective by n.

SGD For Total Loss vs Average Loss

- So we had

$$J(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

- Proper SGD step is

$$-n \times 2\left(w^T X_i - Y_i\right) X_i$$

- What if we take step

$$-2\left(w^T X_i - Y_i\right) X_i?$$

- Then we're optimizing

$$J_1(w) = \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

- Does it matter?

SGD For Total Loss vs Average Loss

- The objective functions

$$
\begin{aligned}
J(w) &= \sum_{i=1}^{n}(w^T x_i - y_i)^2 \\
J_1(w) &= \frac{1}{n}\sum_{i=1}^{n}(w^T x_i - y_i)^2
\end{aligned}
$$

  have the same minimizer w^*.
- But they have different minimum values.

SGD For Total Loss vs Average Loss

- The objective functions

$$
\begin{aligned}
J(w) &= \sum_{i=1}^{n}(w^T x_i - y_i)^2 + \lambda \|w\|^2 \\
J_1(w) &= \frac{1}{n}\sum_{i=1}^{n}(w^T x_i - y_i)^2 + \lambda \|w\|^2
\end{aligned}
$$

  do **not** have the same minimizer w^* for the same λ.
- For the same λ, which objective has the minimizer with smaller "complexity" $\|w\|^2$?

Directional Derivatives

Definition

A **directional derivative** of f at x in the direction δx is

$$f'(x;\delta x) = \lim_{h \downarrow 0} \frac{f(x + h\delta x) - f(x)}{h},$$

and it can be $\pm\infty$ (e.g. for discontinuous functions).

- If f is convex and finite near x, then $f'(x;\delta x)$ exists.
- f is differentiable at x iff for some $g (= \nabla f(x))$ and all δx,

$$f'(x;\delta x) = g^T \delta x.$$

Boyd EE364b: Subgradients Slides

Descent Directions and Optimality

Definition

δx is a **descent direction** for f at x if $f'(x;\delta x) < 0$.

- For differentiable f, if $\nabla f(x) \neq 0$, then $\delta x = -\nabla f(x)$ is a descent direction.
- We have a nice characterization for a minimum in terms of directional derivative:

Theorem

If f is convex and finite near x, then either

- *x minimizes f, or*
- *there is a descent direction for f at x.*

Boyd EE364b: Subgradients Slides

$\lambda_{\max}$ for Lasso

- Lasso objective

$$J_\lambda(w) = \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2 + \lambda |w|_1$$

- Is there a $\lambda_{\max}$ such that $\lambda \geqslant \lambda_{\max}$ implies $\arg\min_w J_\lambda(w) = 0$?
- Suppose yes.
- Then $w = 0$ is a minimum of $J_\lambda(w)$.
- Let's see what that means in terms of our directional derivative characterization.

Directional Derivative for Lasso

- Consider a step direction v. For convenience, take v s.t. $|v| = 1$.
- Then directional derivative at $w = 0$ in direction v is

$$J'_\lambda(0; v) = \lim_{h \downarrow 0} \frac{J(hv) - J(0)}{h}.$$

- For $w = 0$ to be a minimizer, need to have $J'_\lambda(0; v) \geqslant 0$ for every direction v.
- Can find $\lambda_{\max}$ by finding conditions on λ for this to be the case.