Data 100, Discussion 11

Suraj Rampure

Wednesday, November 6th, 2019

Agenda

- Gradients
- Convexity
- Logistic Regression and Cross Entropy Loss

summary stats

$\hat{c}$ = mean(y) (L_2)

$\hat{c}$ = median(y) (L_1)

Gradients and Loss

Recall, in order to find the optimal value of our parameter $\hat{\beta}$, we typically need to minimize some **empirical risk**.

- Empirical risk depends on our choice of loss function: for instance, L_1, L_2, Huber, cross entropy.
- Different choices of loss functions will lead to different values of $\hat{\beta}$.

Sometimes, we're able to find an **analytical** solution for the minimizing value of $\hat{\beta}$.

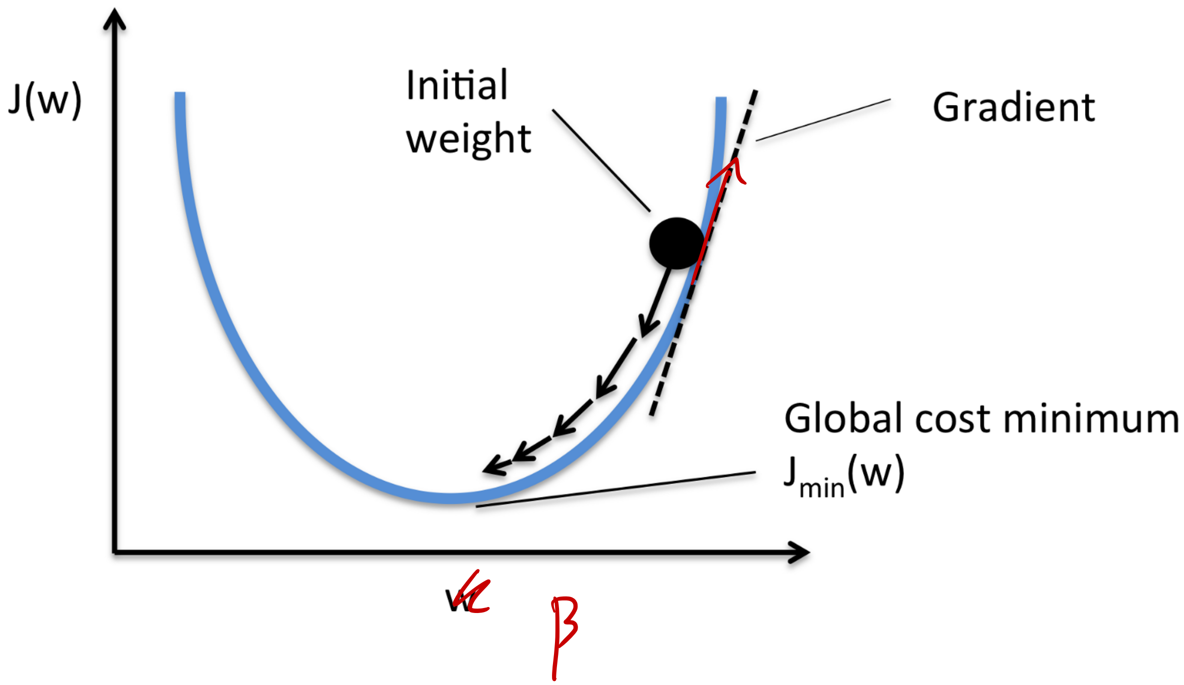
- For instance, $\hat{\beta}_{ridge} = (X^TX + \lambda I)^{-1}X^Ty$.
- As we look at more and more complex loss functions, though, this becomes less common, and so we need to look at **numerical techniques** (like gradient descent).

Gradient Descent

multivariate equivalent to derivative

Goal: Identify the global minimum of a function.

- We know that any minimum of a function occurs where the gradient is 0.
- Hence, gradient descent tries to find the point at which the gradient is 0, by **moving in the opposite direction of the gradient**, iteratively.



Gradient descent update equation:

$$\beta^{(t+1)} = \beta^{(t)} - \alpha \nabla R(X, y, \beta^{(t)})$$

estimate of β at time step $t+1$

est. time t

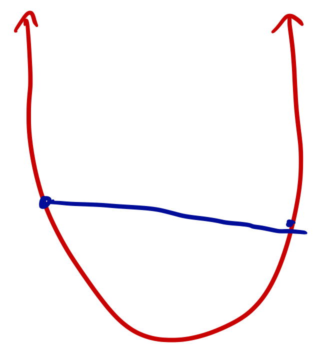
learning rate

"how large of a step to take"

gradient of empirical risk

Convexity

Formally: f is convex iff, for all $x_1, x_2 \in \text{Domain}(f)$ and for all $t \in [0, 1]$,

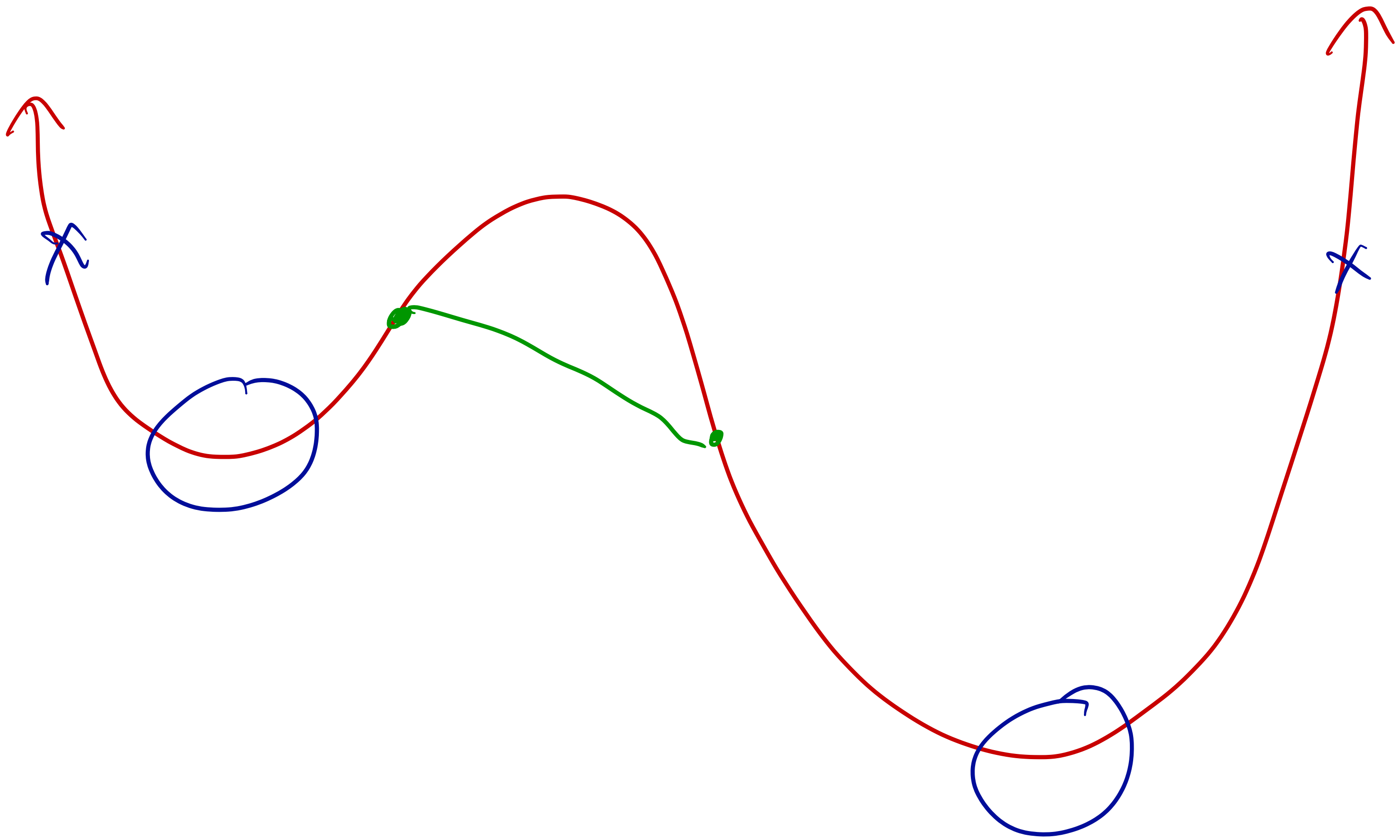


$$tf(x_1) + (1-t)f(x_2) \geq f(tx_1 + (1-t)x_2)$$

- A more meaningful interpretation: f is convex iff **any secant line drawn between two points on f lies on or above the curve**.
- Alternative definition: the **second derivative is always non-negative**. ("concave up")
- These are definitions in one dimension, but they also apply in multiple dimensions.

Why do we care?

- Any local minimum of a convex function is also a global minimum.
- **Gradient descent works well for convex functions** because we are guaranteed that the point where the gradient is 0 is a global minimum.
- This is not necessarily the case for a non-convex function!



not convex

Logistic Regression and Cross Entropy Loss

For this, we'll refer to the lecture slides.

Links to Demos

- https://www.benfrederickson.com/numerical-optimization/
- https://alykhantejani.github.io/images/gradient_descent_line_graph.gif