Lecture 7b: Derivatives and Gradient Descent

Heidi Perry, PhD

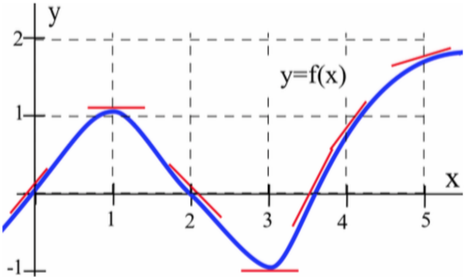
Hack University

heidiperryphd@gmail.com

3/22/2016

Overview

Derivative is the slope of the tangent line



[Hoffman, 2016]

x	$y = f(x)$	$m(x)$
0	0	1
1	1	0
2	0	-1
3	-1	0
4	1	1
5	2	$\frac{1}{2}$

where $m(x)$ is the estimated **slope** of the tangent line to the graph of f(x) aat the point (x, y)

Derivative Basics

Definition of the Derivative

$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Derivative Notation

- $f'(x)$ emphasizes that the derivative is a function related to the function f
- $\mathbf{D}(f)$ emphasizes that f' is a result of an operation on f
- $\frac{df}{dx}$ emphasizes that the derivative is the limit of $\frac{\Delta f}{\Delta x} = \frac{f(x+h)-f(x)}{h}$

[Hoffman, 2016]

Properties of a Function Related to the Derivative

Tangent Line Formula

If $f(x)$ is differentiable at $x = a$ then an equation of the line tangent to f at $(a, f(a))$ is:

$$y = f(a) + f'(a)(x - a)$$

- $f(x)$ is **increasing** if the value increases as the input x move from left to right. $\frac{df}{dx} > 0$
- $f(x)$ is **decreasing** if the value decreases as the input x move from left to right. $\frac{df}{dx} < 0$
- Points where $\frac{df}{dx} = 0$ are **critical points**: (local) maximum, (local) minimum, or inflection point.

Properties of the derivative

- **Theorem:** If a function is differentiable at a point, then it is continuous at that point.
- **Contrapositive form of previous theorem:** If f is not continuous at a point, then f is not differentiable at that point.
- Other times a function is not differentiable:
  - At a cusp or corner.
  - When the tangent line is vertical.

Optimization: Numerical Root-finding Methods

Find minimum (or maximum) of a function by looking for **roots** (points where the function value is zero) of the derivative.

- Bisection
  1. Find interval $[a, b]$ such that $f(a)$ and $f(b)$ have opposite signs.
  2. Bisect interval, $m = \frac{a+b}{2}$.
  3. If $f(m) = 0$, return m. If $f(m) \neq 0$, set $a = m$ or $b = m$ such that condition (1) is met. Repeat until $f(m) < \epsilon$ where ϵ is a threshold that is "close-enough" to zero.
- Newton's method
  1. Pick a starting value x_0.
  2. For each x_n, calculate a new estimate $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ (this "steps" in the direction the tangent line "points" to).
  3. Repeat step 2 until the estimates are "close enough" to a root or until the method fails.

[Hoffman, 2016]

Gradient: Multi-dimensional derivative

Let f be a function of many variables: $f(x_1, x_2, x_3, ..., x_n) = f(\vec{x})$
A partial derivative with respect to any one of the variables is defined as the derivative of the function with all other variables held constant:

$$\frac{\partial f(\vec{x})}{\partial x_i} = \frac{df(x_i)}{dx_i} \text{ where } x_j \neq x_i \text{ are treated as constants}$$

The gradient operator is a vector of partial derivative operators:

$$\nabla = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix} \text{ so } \nabla f(\vec{x}) = \begin{pmatrix} \frac{\partial f(\vec{x})}{\partial x_1} \\ \frac{\partial f(\vec{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\vec{x})}{\partial x_n} \end{pmatrix}$$

The gradient vector points in the direction of greatest **increase** of the function.

Descent Algorithm

Imagine you are a (very intelligent) mouse trying to make it to that lake. You can only see the land maybe a foot around you. How would you proceed?

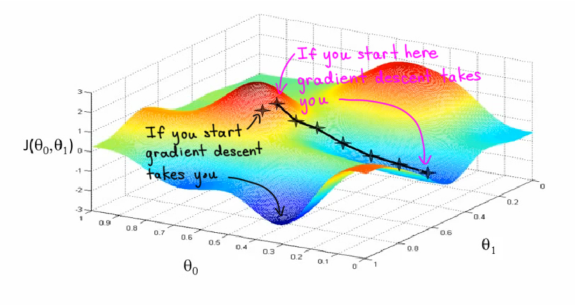
Picture from Laurie at 59 The View From Here

Gradient Descent

Gradient descent algorithm

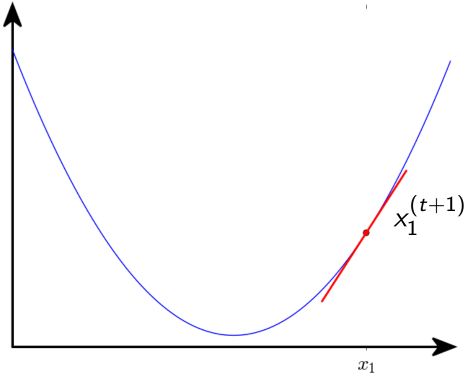
repeat until convergence:

$$\vec{x} := \vec{x} - \alpha \nabla f(\vec{x})$$



Graphic from quinnliu

One-dimensional Gradient Descent: Derivative Term



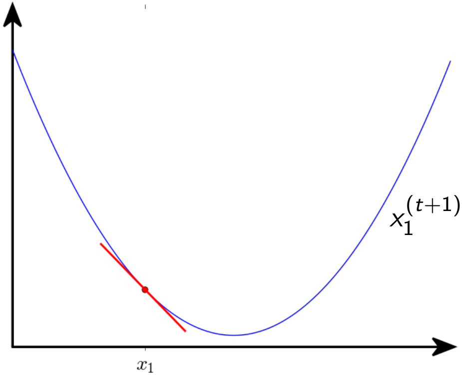
$$x_1^{(t+1)} = x_1^{(t)} - \alpha \tfrac{d}{dx} f(x)$$

$$\tfrac{d}{dx} f(x) > 0$$

$$x_1^{(t+1)} = x_1^{(t)} - \alpha * (\text{positive number})$$

x_i decreases

One-dimensional Gradient Descent: Derivative Term



$$x_1^{(t+1)} = x_1^{(t)} - \alpha \frac{d}{dx} f(x)$$

$$\frac{d}{dx} f(x) < 0$$

$$x_1^{(t+1)} = x_1^{(t)} - \alpha * (\text{negative number})$$

x_1 increases

One-dimensional Gradient Descent: Alpha Term

$$x_1^{(t+1)} = x_1^{(t)} - \alpha \frac{d}{dx} f(x)$$

if α is too small, gradient descent is slow

if α is too large, gradient descent will overshoot the minimum and fail to converge, or even diverge

Apply Gradient Descent Algorithm to Linear Regression

Gradient descent algorithm

repeat until convergence:

$$w_j^{(t+1)} = w_j^{(t)} - \alpha \frac{\partial J}{\partial w_j}\bigg|_{w^{(t)}}$$

for $j = 1$ and $j = 0$

Linear Regression Model

$$h_\theta(x) = \theta_0 + \theta_1 x$$

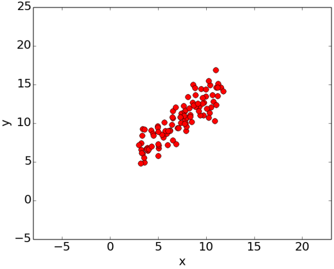
$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

where $x^{(i)}$, $y^{(i)}$ are the observations of the explanatory and response variables respectively

Gradient Descent for Linear Regression Example

The following example, including all graphics, is from Matt Nedrich. Code available on GitHub.

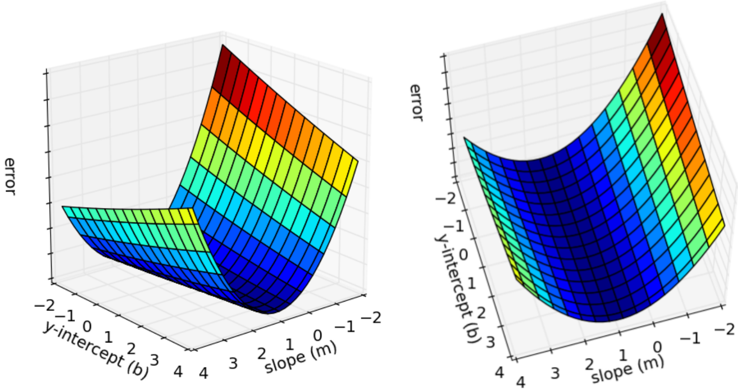
Scatter Plot of Data



Error term to minimize:

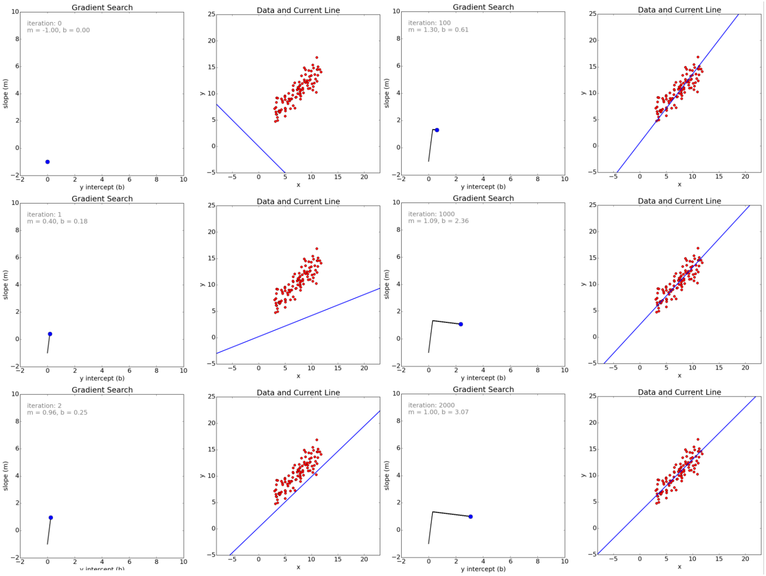
$$J(m, b) = \frac{1}{N} \sum_{i=1}^{N} (y_i - (mx_i + b))^2$$

Gradient Descent for Linear Regression Example

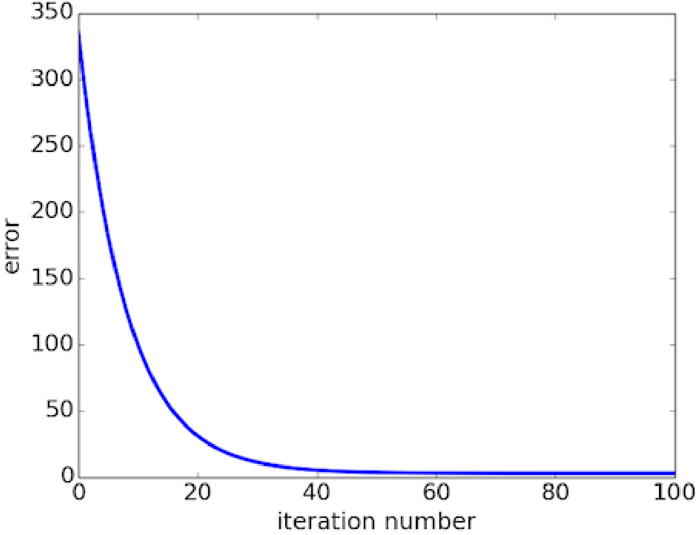


Start gradient search at point $(b, m) = (0, -1)$

Gradient Descent for Linear Regression Example



Gradient Descent for Linear Regression Example



Exercise

IPython notebook: Gradient Descent For Linear Regression

References

Dale Hoffman (2016)
Contemporary Calculus, http://contemporarycalculus.com/

Joel Grus (2015)
Data Science from Scratch, O'Reilly

Andre Ng (2016)
Machine Learning Course - Stanford University Coursera

Recommended Reading

Data Science From Scratch, Chapter 8
Contemporary Calculus, Chapter 2

Articles for discussion:
Four Pitfalls of Hill Climbing