Recommender Systems Mathematical Foundations

Professor Robin Burke

Spring 2019

Outline

- Derivatives
- Gradients
- Gradient descent example

Derivative

- A generalization of the idea of slope
  - A line has a slope
  - https://en.wikipedia.org/wiki/Slope#/media/File:Gradient_of_a_line_in_coordinates_from_-12x%2B2_to_%2B12x%2B2.gif
- Does a curve has slope
  - Yes, but not "a slope"
  - Different slopes at different points
  - https://en.wikipedia.org/wiki/Slope#/media/File:Tangent_function_animation.gif
  - "Instantaneous slope"

Derivative

- A function f' that yields the slope of the curve f at each point
- Examples
  - f = 5x
  - f' = ?
  - f = 5x + 10
  - f' = ?
  - f = -5x – 10
  - f' = ?
  - f = 5x + 6y
  - f' = ?
  - Other notation
    - $\frac{df}{dx}$
    - Make it possible to think about functions of multiple variables

Derivative

- $f=3x^2$
- $f' = 6x$
- Classic proof
  - slope = rise / run
  - slope = f(x+Δ) - f(x) / Δ
  - $f'(x) = \frac{(3x^2+6x\Delta+3\Delta^2)-3x^2}{\Delta} = \frac{6x\Delta+3\Delta^2}{\Delta} = 6x + 3\Delta$
  - Now we let $\Delta \to 0$

General exponent rule

- Except n=0
  - $f(x) = ax^n$
  - $f'(x) = anx^{n-1}$
- Lots of other rules for computing derivatives
  - Take a calculus class!

Derivative of $5x^{-3}$

- A: $-20x^{-4}$
- B: $15x^{-2}$
- C: $15x^{-4}$
- D: $-15x^{-4}$

Slope at a point

- Evaluate the derivative function at that point
- Slope of $3x^2$ where $x = 7$?
  - $f'(7) = 6(7) = 42$
- Important question
  - when is the slope = 0
  - $6x = 0$,
  - only when $x = 0$

For the purposes of this class, why is slope = 0 important?

- A: Because that's the coolest point
- B: Because that's the maximum or minimum of the function
- C: Because that's where the function becomes undefined
- D: Because that's the maximum or minimum of the function if it is convex

Derivatives

- $f(x,y) = 3x^2 - 2y^2$
- $\frac{df}{dx} = 6x$
- $\frac{df}{dy} = -4y$
- More technically correct Partial Derivatives
- $\frac{\delta f}{\delta x} = 6x$
- $\frac{\delta f}{\delta y} = -4y$

Gradient

- Vector of all the partial derivatives
- $y = f(x_{1,} x_2, x_3, x_4) = \sum_{i=1}^{4} x_i^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2$
- $\frac{\partial y}{\partial x_1} = ?$
- A: 2x
- B: $2x_1$
- C: $2x_i$
- D: $2x_1 + x_2^2 + x_3^2 + x_4^2$

Gradient

- Vector-valued function
  - Separate function for each dimension
- $\nabla f = [2x_1, 2x_2, 2x_3, 2x_4]$
- We can evaluate at a given point
  - $\nabla f(1,1,1,2) = [2, 2, 2, 4]$

Our big hairy equation

$$J = \frac{1}{2} \sum_{(i,j)\in S} e_{ij}^2 = \frac{1}{2} \sum_{(i,j)\in S} \left(r_{ij} - \sum_{s=1}^{k} u_{is} \cdot v_{js} \right)^2$$

- r_{ij} – fixed – this is the training data
- u_{is} is the association between user i and one of the k latent factors
  - History / Romance in our example from last week
- v_{js} is the association between item j and one of the k latent factors
- What is $\frac{\partial J}{\partial u_{is}}$?

Computing $\frac{\partial J}{\partial u_{is}}$

- Key point
  - only user i matters
  - can simplify the function
- $\frac{\partial J}{\partial u_{is}} = \frac{\partial}{\partial u_{is}} \frac{1}{2} \sum_{j \in I} \left(r_{ij} - \sum_{s=1}^{k} u_{is} v_{sj} \right)^2$
- $= \frac{\partial}{\partial u_{is}} \frac{1}{2} \sum_{j \in I} \left(r_{ij}^2 - 2 r_{ij} \sum_{s=1}^{k} u_{is} v_{sj} \quad + \left(\sum_{s=1}^{k} u_{is} v_{sj} \right)^2 \right)$
- $= \frac{1}{2} \sum_{j \in I} \left(-2 r_{ij} v_{sj} + 2 v_{sj} \sum_{s=1}^{k} u_{is} v_{sj} \right)$
- $= \sum_{j \in I} \left(-r_{ij} v_{sj} + v_{sj} \sum_{s=1}^{k} u_{is} v_{sj} \right) = \sum_{j \in I} -v_{sj} \left(r_{ij} - \sum_{s=1}^{k} u_{is} v_{sj} \right)$
- $= \sum_{j \in I} -v_{sj} e_{ij}$

This is our prediction foirmula

e_{ij} is the error on r_{ij}

Our big hairy gradient

- A bunch of those $\frac{\partial J}{\partial u_{is}}$ terms
- A bunch of $\frac{\partial J}{\partial v_{sj}}$ terms
  - derivation is very similar
- Now we know the gradient of our loss function at every point
  - that is: we can calculate at each point, what is the steepest uphill direction
  - downhill = - uphill
  - we can calculate the steepest downhill direction

Example

- Let's factorize by gradient descent

Thursday

- Factorization at a faster pace
  - Add in regularization
  - Alternating least squares