otimização

Jacques Wainer

1/4/21

Otimização

$f(w)$ é uma função que recebe um vetor w e devolve um escalar (número).

se $f(w)$ é:

- energia
- custo
- erro
- *loss*

então vc quer minimizar f

se $f(w)$ é:

- ganho
- lucro
- *utility*

voce quer maximizar f

Minimizar e maximizar

minimizar : Obtenha w_0 tal que $f(w_0) \leq f(w)$ para $w \in A$

maximizar : Obtenha w_0 tal que $f(w_0) \geq f(w)$ para $w \in A$

A é a restrição nos valores de w

A pode não ser restrição alguma. Se w tem n dimensões, $A = R^n$ - problemas **sem restrições**

mínimo global: w_0 tal que $f(w_0) \leq f(w)$ para $w \in A$

mínimo local: w_0 tal que $f(w_0) \leq f(w)$ para $|w_0 - w| < \epsilon$ para algum $\epsilon > 0$.

Um video sobre mínimos, máximos e outros pontos críticos

Notação

$$arg \min_{w \in A} f(w)$$

ache w que minimize $f(w)$

subject to $w \in A$

Otimização linear

Otimização linear

ou programação linear

$$f(w) = \alpha_1 w_1 + \alpha_2 w_2 + \ldots + \alpha_n w_n$$

ou

$$f(w) = \alpha^T w$$

para

$$w \in A$$

Restrições convexas

$x, y \in A$ então $\theta x + (1-\theta)y \in A$

conjunto convexo

Restrições lineares

$$\beta_1^T w + c_1 \leq 0$$

$$\beta_2^T w + c_2 \leq 0$$

$$\beta_3^T w + c_3 \leq 0$$

Solução esta nos vértices do politopo (politopo é a extensão multidimensional de um poliedro - que é a extensão 3D de uma figura geometrica 2D)

Otimização convexa

Otimização convexa

minimizar: convexa

maximizar: côncava

função $f()$ é convexa se:

$$f(\theta w + (1-\theta)v) \leq \theta f(w) + (1-\theta)f(v)$$

função convexa

Solução única

se $f()$ é convexo e A é convexo então só existe um mínimo local e ele é o mínimo global!

Não vale se A não é convexo.

Famílias de algoritmos para problemas convexos sem restrições

para a solução de problemas convexos **sem restrições**

- analítico
- descida do gradiente e variações
  - passo pequeno/passo grande (busca linear ou usando segunda derivada (quasi-Newton)
  - diferentes learning rates
  - batch/ minibatch/ stocastic
- sem gradiente

Problemas com restrição

vamos converter problemas com restrições para um outro problema (com mais variáveis) sem restrições.

Problemas não convexos

Problemas não convexos

- há mínimos locais e poucos (ou apenas 1) mínimo global
- ha pontos críticos - derivada em todas as direções = 0 mas não são pontos de mínimo ou máximo
- pontos de sela e ponto de inflexão texto e imagens da khan academy figura de um ponto de sela
- pontos de sela são mínimo em uma direção e máximo em outra. Pode haver uma combinação qualquer de máximo/mínimo e inflexão: num espaço de 40 dimensões um ponto critico (derivada = 0) pode ser máximo para 12 direções, mínimo para 18 direções e inflexão para as restantes 20!.

Tipos de solução para problemas não convexos

- assume que é convexo - acha um mínimo local, e recomeça
- algoritmos de descida do gradiente que são mais insensíveis a pequenos mínimos locais (momento)
- busca em força bruta/cegos: grid e aleatório
- algoritmos tipo genético - PSO e CMA
- otimização bayesiana

Solução analitica de problemas convexos

Solução analítica

$$\frac{\partial f}{\partial w_i} = 0 \quad e \quad \frac{\partial^2 f}{\partial w_i^2} > 0$$

Se voce sabe que f é convexa, então a segunda parte não é importante, o único ponto com derivadas 0 é o mínimo global.

É difícil dizer se uma função é convexa ou não só olhando para a formula.

Regressão linear de 1 variável

Dado N x_i, um dado, e y_i, a saída associada a x_i

Encontre α e β tal que

$$y = \alpha x + \beta$$

quando aplicada a cada x_i e y_i tenha o menor erro possível.

Erro

$\hat{y}_i$ é o valor predito pela equação quando $x = x_i$ e y_i é o valor correto.

- erro quadrado $e_i = (y_i - \hat{y}_i)^2$, ou $e_i = (y_i - \alpha x_i - \beta)^2$
- erro absoluto $e_i = |y_i - \alpha x_i - \beta|$
- o que vc não quer é algo como $e_i = y_i - \alpha x_i - \beta$ que pode ser negativo ou positivo. Não queremos erros negativos compensando erros positivos.
- vamos usar o erro quadrado.
- erro total = $\sum_i e_i$
- erro medio (MSE) = $\frac{1}{N} \sum_i e_i^2$

Minimização

- x_i e y_i não são variáveis, α e β são
- $\langle\alpha, \beta\rangle$ é o vetor de variáveis w.
- ache α, β que minimiza $MSE(\alpha, \beta) = \frac{1}{N}\sum e_i^2$

α

- inicio: $\frac{\partial MSE(\alpha,\beta)}{\partial\alpha} = 0$
- = $\frac{1}{N}\sum\frac{\partial e_i^2}{\partial\alpha}$
- = $\frac{1}{N}\sum 2e_i\frac{\partial e_i}{\partial\alpha}$
- = $\frac{2}{N}\sum e_i(\frac{\partial y_i}{\partial\alpha} - \frac{\partial\alpha x_i}{\partial\alpha} - \frac{\partial\beta}{\partial\alpha})$
- = $\frac{2}{N}\sum e_i(0 - x_i - 0)$
- = $-\frac{2}{N}\sum(y_i x_i - \alpha x_i^2 - \beta x_i)$
- = $-\frac{2}{N}\sum y_i x_i + \frac{2\alpha}{N}\sum x_i^2 - \frac{2\beta}{N}\sum x_i = 0$

etc

$beta$

- início: $\frac{\partial MSE(\alpha,\beta)}{\partial\beta} = 0$
- derivações

Equacões lineares simultaneas

Veja que $\sum x_i, \sum x_i y_i, \sum y_i$ e $\sum x_i^2$ são constantes. E $\frac{\sum x_i}{N}$ é na verdade a média dos x_i.

No final haverá 2 equações e duas incógnitas. A solução é:

$$\alpha = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - (\sum x_i)^2}$$

$$\beta = \frac{\sum y_i}{N} - \alpha\frac{\sum x_i}{N}$$

Regressão linear de multiplas variáveis

É possível fazer as derivações das derivadas para mais de uma dimensão dos dados, mas ai a notação matricial é util. Eu acho que o texto abaixo faz isso. So para voce ver alguma vez derivações algébricas usando matrizes e vetores. texto em ingles

SVD como um problema de otimização

Eu mencionei em aula que o SVD truncado (nos primeiros k valores singulares) é a solução de otimização de reduzir a dimensionalidade de uma matrix de m colunas para k colunas. Isso é chamado do teorema de Eckart-Young

Voce pode ver a demonstração disso e principalmente a formulação do problema de otimização nessa pagina da wikipedia. Vale a pena passar algum tempo entendendo a formulação (nao necessariamente a demonstração).