Machine Learning

Lecture 9: Optimization 3

Hao Tang

December 9, 2022

Unconstrained optimization

$$\min_{w} \quad L(w) \tag{1}$$

An example problem with constraints

- The problem

$$\begin{aligned} \min_{w} \quad & L(w) \\ \text{s.t.} \quad & \|w\|^2 \leq 1 \end{aligned} \tag{2}$$

  is an example of a contrained optimization problem.

- The inequality $\|w\|^2 \leq 1$ is called a constraint.

- Solutions that satisfy the constraints are called feasible solutions.

Representing constraints

- We can write the optimization problem as

$$\min_{w} \qquad L(w) + V_{-}(\|w\|^2 - 1), \tag{3}$$

where

$$V_{-}(s) = \begin{cases} 0 & \text{if } s \leq 0 \\ \infty & \text{if } s > 0 \end{cases}. \tag{4}$$

- This does not change anything; both problems are equally hard (or easy) to solve.

Soften the constraints

- We can approximate

$$\min_{w} \quad L(w) + V_{-}(\|w\|^2 - 1) \tag{5}$$

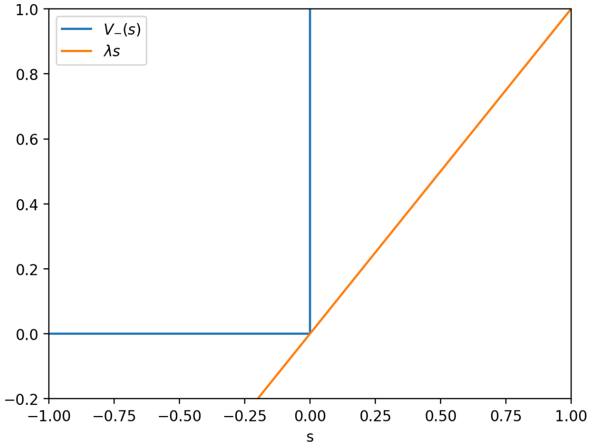
with

$$\min_{w} \quad L(w) + \lambda(\|w\|^2 - 1), \tag{6}$$

for some $\lambda \geq 0$.

- Note that $\lambda s \leq V_{-}(s)$ for all s.

Soften the constraints



Lagrangian

- In general, if you have a optimization problem

$$\begin{aligned} \min_{w} \quad & L(w) \\ \text{s.t.} \quad & h(w) \leq 0 \end{aligned} \tag{7}$$

  the Lagrangian is defined as

$$L(w) + \lambda h(w) \tag{8}$$

  for $\lambda \geq 0$.

- The value λ is called the Lagrange multiplier.

Dual function

- If $\tilde{w}$ is a feasible solution, meaning that $h(\tilde{w}) \leq 0$, then

$$L(\tilde{w}) + \lambda h(\tilde{w}) \leq L(\tilde{w}). \tag{9}$$

- There should be a lowest possible left-hand side,

$$\min_w L(w) + \lambda h(w) \leq L(\tilde{w}) + \lambda h(\tilde{w}) \leq L(\tilde{w}). \tag{10}$$

- We call

$$g(\lambda) = \min_w L(w) + \lambda h(w) \tag{11}$$

the dual function.

Dual function

- We can see that for any λ,

$$g(\lambda) \leq L(w^*) \tag{12}$$

where w^* is the optimal solution for $\min_w L(w)$ subject to $h(w) \leq 0$.

- The proof is the same argument that

$$g(\lambda) = \min_w L(w) + \lambda h(w) \leq L(w^*) + \lambda h(w^*) \leq L(w^*). \tag{13}$$

- In other words, the dual function always has a lower value than the optimal value.

Dual problem

- Since $g(\lambda) \leq L(w^*)$ for any λ,

$$g(\lambda^*) \leq L(w^*) \tag{14}$$

where $\lambda^* = \operatorname{argmax}_{\lambda \geq 0} g(\lambda)$.

- The problem

$$\begin{aligned} \max_{\lambda} \quad & g(\lambda) \\ \text{s.t.} \quad & \lambda \geq 0 \end{aligned} \tag{15}$$

is called the dual problem.

Dual problem

- The dual problem can be written compactly as

$$\max_{\lambda \geq 0} \min_{x} L(x) + \lambda h(x). \tag{16}$$

- For every feasible solution $\hat{x}$, $h(\hat{x}) \leq 0$.
- For every feasible solusion $\hat{x}$, to make $L(\hat{x}) + \lambda h(\hat{x})$ as large as possible, λ has to be zero.
- For the infeasible solusions, $\lambda \to \infty$.

A unigram model

Row, row, row your boat, gently down the stream
Merrily, merrily, merrily, merrily, life is but a dream

- There are 18 words.
- Intuitively,

$$p(\text{row}) = \frac{3}{18} \qquad p(\text{merrily}) = \frac{4}{18} \qquad p(\text{is}) = \frac{1}{18} \tag{17}$$

A unigram model

- There are 13 unique words.
- We refer to the set of unique words $V = \{\text{row}, \text{your}, \text{boat}, \text{gently}, \text{down}, \text{the}, \text{stream}, \text{merrily}, \text{life}, \text{is}, \text{but}, \text{a}, \text{dream}\}$ as the vocabulary.
- We assign each word v a probability β_v.
- The probability of a word is

$$p(w) = \prod_{v \in V} \beta_v^{\mathbb{1}_{v=w}}. \tag{18}$$

A unigram model

- We assume that each word is independent of others.
- This assumption is obviously wrong, but can go really far.
- The likelihood of β given the data is

$$\log p(w_1, \ldots, w_N) = \log \prod_{i=1}^{N} p(w_i) = \log \prod_{i=1}^{N} \prod_{v \in V} \beta_v^{\mathbb{1}_{v=w_i}}. \tag{19}$$

- Since β is a probability vector, we have the assumption

$$\sum_{v \in V} \beta_v = 1. \tag{20}$$

A unigram model

- We arrive at the optimization problem

$$\begin{aligned} \min_{\beta} \quad & -\sum_{i=1}^{N} \sum_{v \in V} \mathbb{1}_{v=w_i} \log \beta_v \\ \text{s.t.} \quad & \sum_{v \in V} \beta_v = 1 \end{aligned} \tag{21}$$

- Its Lagrangian is

$$F = -\sum_{i=1}^{N} \sum_{v \in V} \mathbb{1}_{v=w_i} \log \beta_v + \lambda \left(\sum_{v \in V} \beta_v - 1 \right). \tag{22}$$

A unigram model

- Solving the optimality condition gives

$$\frac{\partial F}{\partial \beta_k} = \sum_{i=1}^{N} \mathbb{1}_{k=w_i} \frac{1}{\beta_k} - \lambda = 0 \implies \beta_k = \frac{1}{\lambda} \sum_{i=1}^{N} \mathbb{1}_{k=w_i}. \tag{23}$$

- The dual problem is

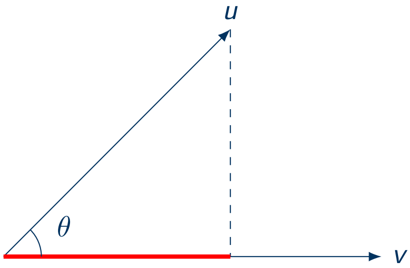
$$\max_{\lambda \geq 0} \quad - \sum_{i=1}^{N} \sum_{v \in V} \mathbb{1}_{v=w_i} \log \frac{1}{\lambda} \sum_{j=1}^{N} \mathbb{1}_{v=w_j} + \lambda \left(\sum_{v \in V} \frac{1}{\lambda} \sum_{i=1}^{N} \mathbb{1}_{v=w_i} - 1 \right). \tag{24}$$

A unigram model

$$\sum_{v \in V} \beta_v = \sum_{v \in V} \frac{1}{\lambda} \sum_{i=1}^{N} \mathbb{1}_{v=w_i} = 1 \implies \lambda = \sum_{v \in V} \sum_{i=1}^{N} \mathbb{1}_{v=w_i} = N \tag{25}$$

$$\beta_k = \frac{\sum_{i=1}^{N} \mathbb{1}_{k=w_i}}{\sum_{v \in V} \sum_{i=1}^{N} \mathbb{1}_{v=w_i}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{k=w_i} \tag{26}$$

Projection



$$\|u\| \cos\theta = \|u\| \frac{u^\top v}{\|u\|\|v\|} = \frac{u^\top v}{\|v\|} \tag{27}$$

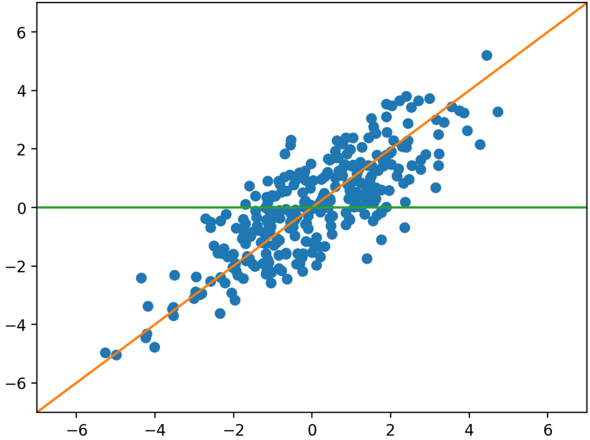
Projection

- The projection of x onto w is $\frac{x^\top w}{\|w\|}$.
- If we have N data points $\{x_1, \ldots, x_N\}$, then the sum of the (squared) projection is

$$\sum_{i=1}^{N} \left(\frac{x_i^\top w}{\|w\|} \right)^2 = \frac{w^\top X X^\top w}{w^\top w}. \tag{28}$$

- The sum of squared projection can be seen as the spread of the data.

Maximal projection



Maximal projection

- We want to find the maximum direction to project.
- The optimization problem is

$$\max_{w} \frac{w^\top X X^\top w}{w^\top w}. \tag{29}$$

Maximal projection

- The problem is scale invariant.

$$\frac{(aw)^\top XX^\top (aw)}{(aw)^\top (aw)} = \frac{w^\top XX^\top w}{w^\top w}. \tag{30}$$

- The problem is equivalent to

$$\max_w w^\top XX^\top w \qquad \text{s.t. } \|w\|^2 = 1. \tag{31}$$

Maximal projection

- The Lagrangian is

$$F = w^\top X X^\top w + \lambda(1 - \|w\|^2). \tag{32}$$

- Finding the optimal solution gives

$$\frac{\partial F}{\partial w} = (XX^\top + XX^\top)w - 2\lambda w = 0 \implies XX^\top w = \lambda w. \tag{33}$$

- It turns out that λ is an eigenvalue, and w an eigenvector.

Maximal projection

- Plugging the solution back to the objective,

$$\frac{w^\top X X^\top w}{w^\top w} = \frac{\lambda w^\top w}{w^\top w} = \lambda \tag{34}$$

- Since the goal is to find the maximal projection, this is now equivalent to finding the largest eigenvalue of $XX^\top$.

Maximal projection

- The term

$$\frac{w^\top XX^\top w}{w^\top w} \tag{35}$$

is called the Rayleigh quotient.

- The optimal w is called the first principal component.

- We will learn more about this when we talk about principal component analysis.