Machine Learning - MT 2017
8. Optimisation I

Christoph Haase

University of Oxford
October 25, 2017

Outline

Most machine learning methods can (ultimately) be cast as optimization problems.

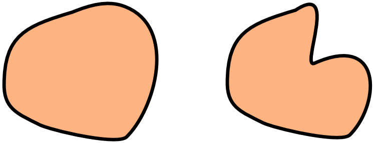
- Convex Optimization
- Recap: Gradients, Hessians
- Gradient Descent
- Stochastic Gradient Descent
- Constrained Optimization

Most machine learning packages such as scikit-learn, tensorflow, octave, torch *etc.*, will have optimization methods implemented. But you will have to understand the basics of optimization to use them effectively.

Convex Sets

A set $C \subseteq \mathbb{R}^D$ is convex if for any $\mathbf{x}, \mathbf{y} \in C$ and $\lambda \in [0, 1]$,

$$\lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{y} \in C$$



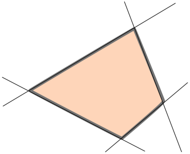
Examples of Convex Sets

- The entire set $\mathbb{R}^D$: since $\lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{y} \in \mathbb{R}^D$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$
- Intersections of convex sets: Given convex sets $C_1, \dots, C_n$, the set $\bigcap_{i=1}^n C_i$ is obviously convex
- Norm balls: For any L-norm $||\cdot||$, the set $B = \{\mathbf{x} \in \mathbb{R}^D : ||\mathbf{x}|| \leq 1\}$ is convex, since for $\mathbf{x}, \mathbf{y} \in B$ we have

  $$||\lambda \cdot \mathbf{x} + (1 - \lambda) \cdot y|| \leq ||\lambda \cdot \mathbf{x}|| + ||(1 - \lambda) \cdot \mathbf{y}|| = \lambda \cdot ||\mathbf{x}|| + (1 - \lambda) \cdot ||\mathbf{y}|| \leq 1$$

- Polyhedra: Given an $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$, a polyhedron is the set $P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A} \cdot \mathbf{x} \leq \mathbf{b}\}$, since for $\mathbf{x}, \mathbf{y} \in P$ we have

  $$\mathbf{A} \cdot (\lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{y}) = \lambda \cdot \mathbf{A} \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{A} \cdot \mathbf{y} \leq \lambda \cdot \mathbf{b} + (1 - \lambda) \cdot \mathbf{b} = \mathbf{b}$$



Examples of Convex Sets

The set of positive semi-definite matrices is convex:

- Recall that $\mathbf{A} \in \mathbb{R}^D$ is positive semi-definite if $\mathbf{A} = \mathbf{A}^\mathsf{T}$ and $\mathbf{x}^\mathsf{T} \cdot \mathbf{A} \cdot \mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^D$
- Set $\mathbb{S}_+^D$ of all such matrices is called the positive semidefinite cone
- $\mathbb{S}_+^D$ is convex, as for $\mathbf{A}, \mathbf{B} \in \mathbb{S}_+^D$, we have

$$\mathbf{x}^\mathsf{T} \cdot (\lambda \cdot \mathbf{A} + (1 - \lambda) \cdot \mathbf{B}) \cdot \mathbf{x} = \lambda \cdot \mathbf{x}^\mathsf{T} \cdot \mathbf{A} \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{x}^\mathsf{T} \cdot \mathbf{B} \cdot \mathbf{x} \geq 0$$

Convex Functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ defined on a convex domain is convex if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$ where f is defined and $0 \leq \lambda \leq 1$,

$$f(\lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{y}) \leq \lambda \cdot f(\mathbf{x}) + (1 - \lambda) \cdot f(\mathbf{y})$$

Examples:

- Affine functions: $f(\mathbf{x}) = \mathbf{b}^\mathsf{T} \cdot \mathbf{x} + c$
- Quadratic functions: $f(\mathbf{x}) = 1/2 \cdot \mathbf{x}^\mathsf{T} \cdot \mathbf{A} \cdot \mathbf{x} + \mathbf{b}^\mathsf{T} \cdot \mathbf{x} + c$, where $\mathbf{A}$ is symmetric positive semidefinite
- Norms: In particular L^p-norms, but any norm will be convex
- Nonnegative weighted sums of convex functions: Given convex functions $f_1, \ldots, f_n$ and $w_1, \ldots, w_n \in \mathbb{R}_{\geq 0}$, the following is a convex function

$$f(\mathbf{x}) = \sum_{i=1}^{k} w_i \cdot f_i(\mathbf{x})$$

Convex Optimization

Given convex functions $f(\mathbf{x}), g_1(\mathbf{x}), \ldots, g_m(\mathbf{x})$ and affine functions $h_1(\mathbf{x}), \ldots h_n$, a convex optimization problem is of the form:

$$\begin{aligned} \text{minimize } & f(\mathbf{x}) & \\ \text{subject to } & g_i(\mathbf{x}) \leq 0 & i \in \{1, \ldots, m\} \\ & h_j(\mathbf{x}) = 0 & j \in \{1, \ldots, n\} \end{aligned}$$

Goal is to find an optimal value of a convex optimization problem:

$$v^* = \min\{f(\mathbf{x}) : g_i(\mathbf{x}) \leq 0, i \in \{1, \ldots, m\}, h_i(\mathbf{x}) = 0, j \in \{0, \ldots, n\}\}$$

Whenever $f(\mathbf{x}^*) = v^*$ then $\mathbf{x}^*$ is an optimal point, which does not need to be unique, and can take values $+\infty$ (in infeasible instances) or $-\infty$ (in unbounded instances)

Classes of Convex Optimization Problems

Linear Programming:

$$\begin{aligned} \text{minimize } & \mathbf{c}^{\mathsf{T}} \cdot \mathbf{x} + d \\ \text{subject to } & \mathbf{A} \cdot \mathbf{x} \leq \mathbf{e} \\ & \mathbf{B} \cdot \mathbf{x} = \mathbf{f} \end{aligned}$$

Quadratically Constrained Quadratic Programming:

$$\begin{aligned} \text{minimize } & \frac{1}{2}\mathbf{x}^{\mathsf{T}} \cdot \mathbf{B} \cdot \mathbf{x} + \mathbf{c}^{\mathsf{T}} \cdot \mathbf{x} + d \\ \text{subject to } & \frac{1}{2}\mathbf{x}^{\mathsf{T}} \cdot \mathbf{Q}_i \cdot \mathbf{x} + \mathbf{r}_i^{\mathsf{T}} \cdot \mathbf{x} + s_i \leq 0 \qquad i \in \{1, \ldots, m\} \\ & \mathbf{A} \cdot \mathbf{x} = \mathbf{b} \end{aligned}$$

Semidefinite Programming:

$$\begin{aligned} \text{minimize } & \operatorname{tr}(\mathbf{C} \cdot \mathbf{X}) \\ \text{subject to } & \operatorname{tr}(\mathbf{A}_i \cdot \mathbf{X}) = b_i \qquad i \in \{1, \ldots, m\} \\ & \mathbf{X} \text{ positive semidefinite} \end{aligned}$$

Here, $\operatorname{tr}(\mathbf{A})$ is the trace of the matrix $\mathbf{A}$

Local Optima are Global Optima

Call $\mathbf{x}$ locally optimal if it is feasible and there is $B > 0$ such that $f(\mathbf{x}) \leq f(\mathbf{y})$ for all feasible $\mathbf{y}$ such that $||\mathbf{x} - \mathbf{y}||_2 \leq B$.

Call feasible $\mathbf{x}$ globally optimal if $f(\mathbf{x}) \leq f(\mathbf{y})$ for all feasible $\mathbf{y}$.

Theorem

For a convex optimization problem, all locally optimal points are globally optimal.

- Suppose $\mathbf{x}$ is locally optimal and $\mathbf{y} \neq \mathbf{x}$ is such that $f(\mathbf{y}) < f(\mathbf{x})$
- Now $f(\mathbf{z}) < f(\mathbf{x})$ does not hold for any $\mathbf{z}$ such that $||\mathbf{x} - \mathbf{z}||_2 \leq B$
- Set $\mathbf{z} = \lambda \cdot \mathbf{y} + (1 - \lambda) \cdot \mathbf{x}$ with $\lambda = \frac{B}{2 \cdot ||\mathbf{x} - \mathbf{y}||_2}$
- We have $||\mathbf{x} - \mathbf{z}||_2 \leq B$, since

$$||\mathbf{x} - \mathbf{z}||_2 = ||\mathbf{x} - (\lambda \cdot \mathbf{y} + (1 - \lambda) \cdot \mathbf{x})||_2 = ||\lambda \cdot (\mathbf{x} - \mathbf{y})||_2 = B/2$$

- Convexity of f gives the desired contradiction $f(\mathbf{z}) < f(\mathbf{x})$:

$$f(\mathbf{z}) = f(\lambda \cdot \mathbf{y} + (1 - \lambda) \cdot \mathbf{x}) \leq \lambda \cdot f(\mathbf{y}) + (1 - \lambda) \cdot f(\mathbf{x}) < f(\mathbf{x})$$

Linear Programming

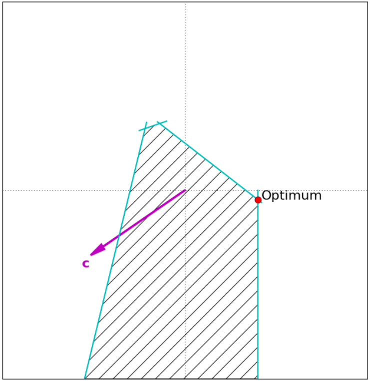
Looking for solutions $\mathbf{x} \in \mathbb{R}^n$ to the following optimization problem

minimize $\quad \mathbf{c}^\mathsf{T}\mathbf{x}$

subject to:

$$\mathbf{a}_i^\mathsf{T}\mathbf{x} \leq b_i, \quad i = 1, \ldots, m$$

$$\bar{\mathbf{a}}_i^\mathsf{T}\mathbf{x} = \bar{b}_i, \quad i = 1, \ldots, l$$



- No analytic solution
- Efficient algorithms exist, both in theory and practice

Linear Model with Absolute Loss

Suppose we have data $\langle(\mathbf{x}_i, y_i)\rangle_{i=1}^{N}$ and that we want to minimise the objective:

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} |\mathbf{x}_i^{\mathsf{T}}\mathbf{w} - y_i|$$

Let us introduce ζ_i one for each datapoint

Consider the linear program in the $D + N$ variables $w_1, \ldots, w_D, \zeta_1, \ldots, \zeta_N$

$$\text{minimize} \quad \sum_{i=1}^{N} \zeta_i$$

subject to:

$$\mathbf{w}^{\mathsf{T}}\mathbf{x}_i - y_i \le \zeta_i, \qquad i = 1, \ldots, N$$

$$y_i - \mathbf{w}^{\mathsf{T}}\mathbf{x}_i \le \zeta_i, \qquad i = 1, \ldots, N$$

Minimising the Lasso Objective

For the Lasso objective, *i.e.*, linear model with ℓ_1-regularisation, we have

$$\mathcal{L}_{\text{lasso}}(\mathbf{w}) = \sum_{i=1}^{N} (\mathbf{w}^\mathsf{T} \mathbf{x}_i - y_i)^2 + \lambda \sum_{i=1}^{D} |w_i|$$

- Quadratic part of the loss function can't be framed as linear programming
- Lasso regularization does not allow for closed form solutions
- Can be rephrased as quadratic programming problem
- Alternatively resort to general optimisation methods