Machine Learning for Big Data: Constrained Optimization

Lionel Fillatre

fillatre@unice.fr

Topics

- Introduction
- Convexity
- Conditions of optimality
- Conclusion

1 Introduction

Examples of machine learning optimization problems

Linear Classification

$$\arg\min_{w} \sum_{i=1}^{n} ||w||^2 + C \sum_{i=1}^{n} \xi_i$$

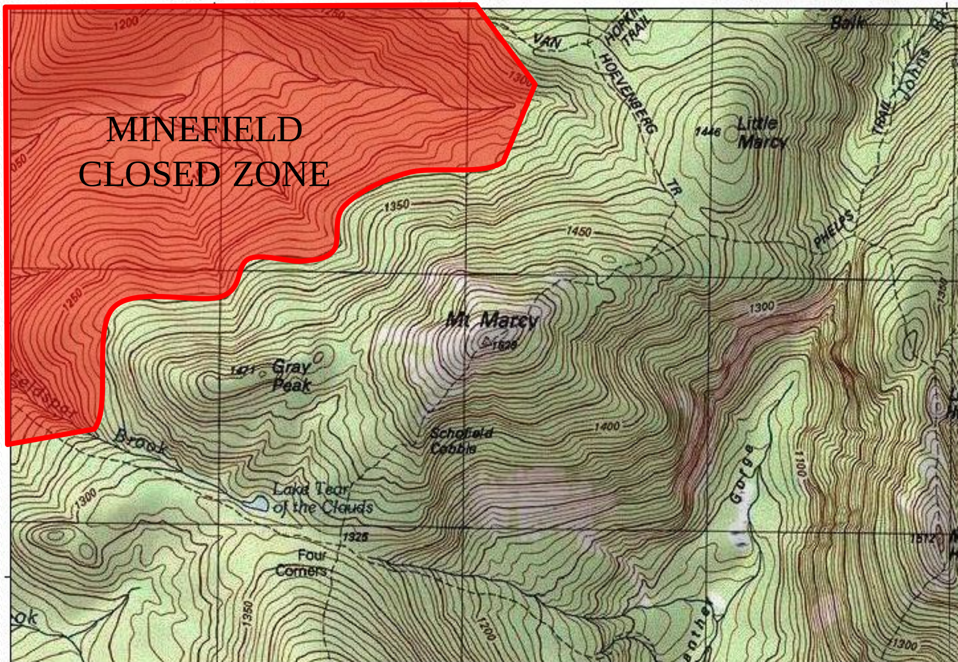
$$\text{s.t. } 1 - y_i x_i^T w \leq \xi_i$$

$$\xi_i \geq 0$$

K-Means

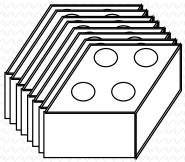
$$\arg \min_{\mu_1,\mu_2,\ldots,\mu_k} J(\mu) = \sum_{j=1}^{k} \sum_{i \in C_j} ||x_i - \mu_j||^2$$

Constrained optimization

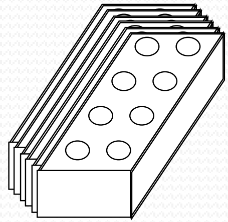


A Production Problem

Weekly supply of raw materials:



8 Small
Bricks



6 Large Bricks

Products:

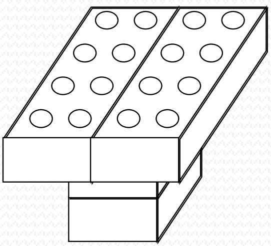
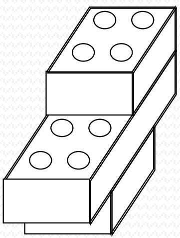


Table
Profit = $20/Table



Chair
Profit = $15/Chair

Linear Programming

- Linear programming uses a mathematical model to find the best allocation of scarce resources to various activities so as to maximize profit or minimize cost.

$$
\begin{aligned}
&\text{Maximize } (\$15)\,Chairs + (\$20)\,Tables \\
&\text{subject to} \\
&\qquad \text{Large Bricks:} \quad Chairs + 2\,Tables \le 6 \\
&\qquad \text{Small Bricks:} \quad 2\,Chairs + 2\,Tables \le 8 \\
&\text{and} \\
&\qquad Chairs \ge 0, \quad Tables \ge 0.
\end{aligned}
$$

Constrained Optimization

$$f(x)\colon \mathbb{R}^n \to \mathbb{R}$$

$$x^* = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} f(x)$$

Subject to:

- Equality constraints defined over $\mathbb{R}^n$: $a_i(x) = 0, i = 1,2,\dots,p$
- Nonequality constraints defined over $\mathbb{R}^n$: $c_j(x) \geq 0, j = 1,2,\dots,q$

- Constraints define a feasible region $x \in \mathcal{F} \subset \mathbb{R}^n$, which should be nonempty.

- An inequality $c_j(x) \leq 0$ is equivalent to $-c_j(x) \geq 0$

➡ The idea is to convert it to an unconstrained optimization.

2 Karush-Kuhn-Tucker conditions

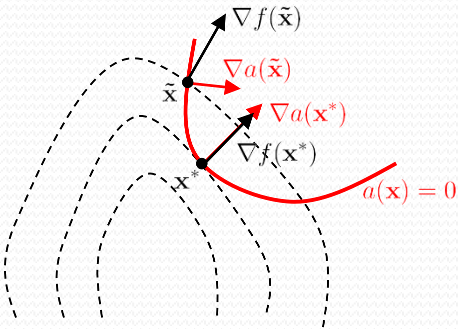
Equality constraints

- Minimize *f*(**x**) subject to: $a_i(x) = 0$ for $i = 1,2, \dots, p$

- Main result (necessary condition): the gradient of *f*(**x**) at a local minimizer x^* is equal to the linear combination of the gradients of $a_i(\mathbf{x})$ with Lagrange multipliers λ_i^* as the coefficients

$$\nabla f(\mathbf{x}^*) = \sum_{i=1}^{p} \lambda_i^* \nabla a_i(\mathbf{x}^*)$$

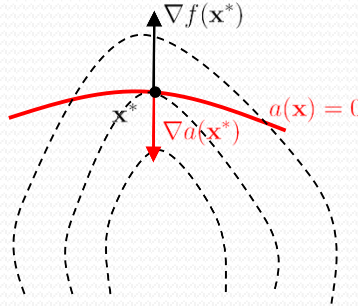
Geometric interpretation



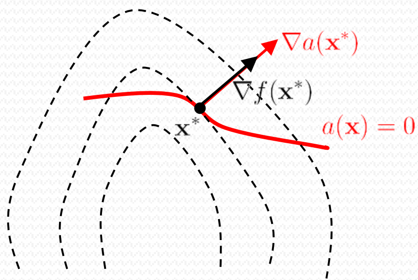
$f_3 > f_2 > f_1$

$\tilde{\mathbf{x}}$ is not a minimizer

x* is a minimizer, $\lambda^*>0$



$f_3 > f_2 > f_1$ x* is a minimizer, $\lambda^*<0$



$f_3 > f_2 > f_1$

x* is not a minimizer

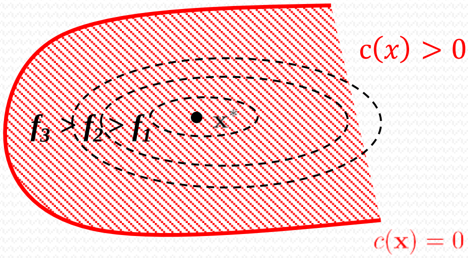
Inequality constraints

- Minimize *f(**x**)* subject to: $c_j(x) \geq 0$ for $j = 1,2, \ldots, q$

- Main result (necessary condition): the gradient of *f(**x**)* at a local minimizer x^* is equal to the linear combination of the gradients of $c_j(x)$ which are active ($c_j(x)=0$ for all $j \in A \subset \{1,2, \ldots, q\}$) with KKT multipliers $\mu_j^* \geq 0$ as the coefficients

$$\nabla f(\mathbf{x}^*) = \sum_{j \in A} \mu_j^* \nabla c_j(\mathbf{x}^*)$$

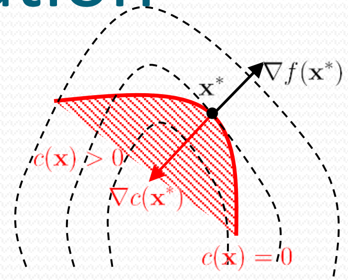
Geometric interpretation



No active constraints at x^*,

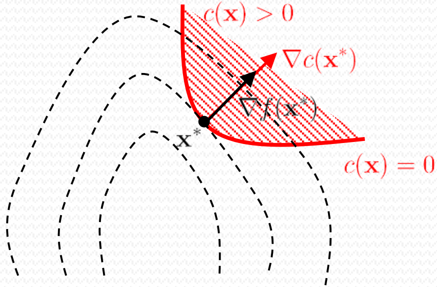
$$\nabla f(x^*) = 0$$

x* is a minimizer



$f_3 > f_2 > f_1$

x* is not a minimizer, μ<0



$f_3 > f_2 > f_1$

x* is a minimizer, μ>0

Lagrangian funtion and KKT

- We can introduce the Lagrangian function

$$L(x,\lambda,\mu) = f(x) - \sum_{i=1}^{p} \lambda_i a_i(x) - \sum_{i=1}^{q} \mu_j c_j(x)$$

- The necessary conditions for the local minimizer $x^* \in \mathcal{F}$ are
  1. $\nabla_x L(x^*,\lambda,\mu) = 0$
  2. $a_i(x^*) = 0$ for all $i = 1, \dots, p$
  3. $\mu_j c_j(x^*) = 0$ for all $j = 1, \dots, q$

- These are Karush-Kuhn-Tucker conditions (KKT)

KKT: sufficient conditions

- If the objective $f(x)$ is convex, the inequality constraints $c_j(x)$ are convex and the equality constraints $a_i(x)$ are affine, the KKT conditions are sufficient.

- In this case, x^* is the solution of the constrained problem (global constrained minimizer)

Example of KKT

- Let us solve

$$\min_{x \in \mathbb{R}} x^2$$
$$\text{s.t. } x \geq 2$$

- Lagrangian function: $L(x, \mu) = x^2 - \mu(x - 2)$

- KKT conditions:
  - $\nabla_x L(x^*, \mu^*) = 2x^* - \mu^* = 0$ which yields $x^* = \frac{\mu^*}{2}$
  - $\mu^*(x^* - 2) = 0$, which yields ($\mu^* = 0$, so $x^* = 0$) or ($x^* = 2$, so $\mu^* = 4 \geq 0$)

- The solution is $x^* = 2$ because $x^* \geq 2$ and KKT conditions are sufficient in this case (convex problem).

3 Dual Lagrangian

Dual Lagrangian

- The Lagrange dual function $g: \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ is defined as the minimum of the Lagrangian over *x*:

$$\mathrm{g}(\lambda, \mu) = \inf_{x \in \mathbb{R}^n} L(x, \lambda, \mu) = \inf_{x \in \mathbb{R}^n} \left(f(x) - \sum_{i=1}^{p} \lambda_i a_i(x) - \sum_{i=1}^{q} \mu_j c_j(x) \right)$$

- Observe that:
  - the infimum is unconstrained (as opposed to the original constrained minimization problem)
  - g is concave regardless of original problem (infimum of affine functions)
  - g is defined oved a domain *dom(g)* which depends on the problem
  - g can take the value $-\infty$ for some λ, μ

Lower Bound of Dual Lagrangian

- Let $p^* = \inf_{x \in \mathcal{F}} f(x)$ the optimal value of the constrained problem
- We can show that

$$p^* = \min_{x \in \mathcal{F}} \max_{\mu \geq 0} L(x, \lambda, \mu)$$

- It is shown that, if $\mu_j \geq 0$ for all $j = 1, \ldots, q$,

$$\mathrm{g}(\lambda, \mu) \leq p^*$$

for all λ_i, $i = 1, \ldots, p$

- The idea is to maximize $\mathrm{g}(\lambda, \mu)$ with respect to the Lagragian multipliers.

Dual Lagrangian

- The Dual Lagrangian problem is defined as

$$\max_{(\lambda,\mu)\,\in\, dom(g)} g(\lambda,\mu)$$

$$\text{s.t. } \mu_j \geq 0 \text{ for all } j = 1, \dots, q$$

- This problem finds the best lower bound on p^* obtained from the dual function.
- It is a convex optimization (maximization of a concave function and linear constraints).
- The optimal value is denoted d^*.
- λ, μ are dual feasible if $\mu_j \geq 0$ for all j and $(\lambda, \mu) \in dom(g)$. In general, the latter implicit constraints can be made explicit in problem formulation.

Strong Duality

- Under certain assumptions (not studied in this course), strong duality holds:

$$d^* = p^*$$

- It is very desirable because we can solve a difficult problem by solving the dual problem

- The strong duality does not hold in general

- The strong duality usually holds for convex problems

- Conditions that guarantee strong duality in convex problems are called constraint qualifications.

Comments on the dual approch

- This dual approach is not guaranteed to succeed. However, it does for a certain class of functions.
- In these cases it often leads to a simpler optimization problem.
- Particularly in the case when the dimension of x is much larger than the number of constraints.
- The expression of x in terms of the Lagrange multipliers may give some insight into the optimal solution.

Example of dual approach

- Let us solve

$$\min_{x \in \mathbb{R}} x^2$$
$$\text{s.t. } x \geq 2$$

- Lagrangian function: $L(x, \mu) = x^2 - \mu(x - 2)$

- Dual Lagrangian function:

$$\mathrm{g}(\mu) = \inf_{x \in \mathbb{R}} L(x, \mu) = \inf_{x \in \mathbb{R}} \left(\left(x - \frac{\mu}{2} \right)^2 - \frac{\mu^2}{4} + 2\mu \right) = -\frac{\mu^2}{4} + 2\mu,$$

i.e. $\mathrm{g}(\mu) = L(x^*, \mu) = -\left(\frac{\mu}{2} - 2 \right)^2 + 4$ for $x^* = \frac{\mu}{2}$

- $\max_{\mu \geq 0} g(\mu) = \max_{\mu \geq 0} \left(-\left(\frac{\mu}{2} - 2 \right)^2 + 4 \right) = 4$ for $\mu^* = 4$.

- The solution is $x^* = 2$.

8 Conclusion

Conclusion

- Constrained optimization is very important in machine learning to deal with constrained problems
- A huge number of applications including classification with Support Vector Machine
- Many extensions including regularization, sparsity, etc.