Statistical Learning. Convex optimization

Jose Ameijeiras Alonso
Departamento de Estatística, An. Mat. e Otimización (USC)

Máster Interuniversitario en Tecnologías de Análisis de Datos Masivos: Big Data

Introduction

- In many problems in statistical estimation and regression the solution requires either iterative methods or numerical optimization.

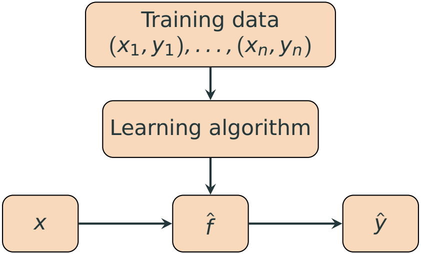
Introduction

- Example: Suppose that we observe a quantitative response Y and a preditor variable X and we assume that there is some relationship between Y and X. The relation can be written in general:

$$Y = f(X) + \epsilon$$

  - f is some fixed but unknown function
  - ϵ is a random error term
- How do we estimate f?
- We want to find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y).

Introduction

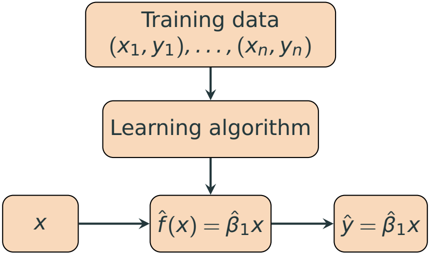


- To evaluate the performance of a statistical learning method, we need some way to measure how well its predictions actually match the observed data.
- The most commonly-used measure is the mean squared error (MSE), given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Introduction

- Example: Let us consider one very simple assumption: $f(X) = \beta_1 X$



$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i)^2$$

- The objective is to choose the value $\hat{\beta}_1$ which minimizes the MSE

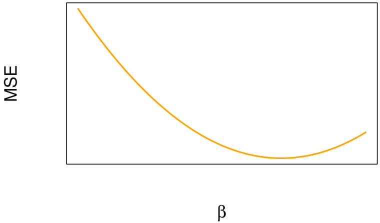
Introduction

- Example: Let us consider one very simple assumption: $f(X) = \beta_1 X$
- Given $(x_1, y_1), \ldots, (x_n, y_n)$, the objective is to choose the value $\hat{\beta}_1$ which minimizes the MSE

$$\hat{\beta}_1 = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta x_i)^2 \equiv \arg\min_{\beta} \|y - x^t \beta\|_2^2$$

where $x = (x_1, \ldots, x_n)^t$ and $y = (y_1, \ldots, y_n)^t$

- Here we show the representation of the MSE (as a function of β) for a given training set $(x_1, y_1), \ldots, (x_n, y_n)$



Introduction

- A mathematical optimization problem, has the form

$$\boxed{\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \;\; i = 1, \ldots, m \\ & h_j(x) = 0, \;\; j = 1, \ldots, p \end{array}}$$

- The vector $x = (x_1, \ldots, x_d)^t \in \mathbb{R}^d$ is the optimization variable
- The function $f_0 : \mathbb{R}^d \rightarrow \mathbb{R}$ is the objective function
- The functions $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$, $i = 1, \ldots, m$ are the inequality constraint functions
- The functions $h_j : \mathbb{R}^d \rightarrow \mathbb{R}$, $j = 1, \ldots, p$ are the equality constraint functions
- The domain of the problem is:

$$\mathcal{D} = \text{dom}(f_0) \cap \text{dom}(f_1) \cap \ldots \cap \text{dom}(f_m) \cap \text{dom}(h_1) \cap \ldots \cap \text{dom}(h_p)$$

Introduction

- A mathematical optimization problem, has the form

$$\boxed{\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0,\ i = 1, \ldots, m \\ & h_j(x) = 0,\ i = j, \ldots, p \end{array}}$$

- The feasible set is the set of points in $\mathcal{D}$ that satisfy all the constraints (the set of feasible solutions)
- The optimal set is the set of feasible points for which the objective function achieves the optimal value, denoted by $f*$
- A point $x*$ is optimal if it belongs to the optimal set

Convex optimization problem

- A convex optimization problem, has the form

$$\boxed{\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \; i = 1, \ldots, m \\ & a_j^t x = b_j, \; j = 1, \ldots, p \end{array}}$$

  - where $f_0 : \mathbb{R}^d \to \mathbb{R}$, and $f_i : \mathbb{R}^d \to \mathbb{R}, i = 1, \ldots, m$ are convex functions,
  - $a_j = (a_{j1}, \ldots, a_{jd})^t \in \mathbb{R}^d$ is a vector of coefficients and $b_j \in \mathbb{R}$, $j = 1, \ldots, p$.

Convex sets

- **Line segment:** Let x, y be two points in $\mathbb{R}^d$ with $x \neq y$. Points of the form

$$z = \theta x + (1-\theta)y$$

with $\theta \in [0,1]$ form the line segment joining x and y.

- **Convex set:** A set $C \in \mathbb{R}^d$ is convex if

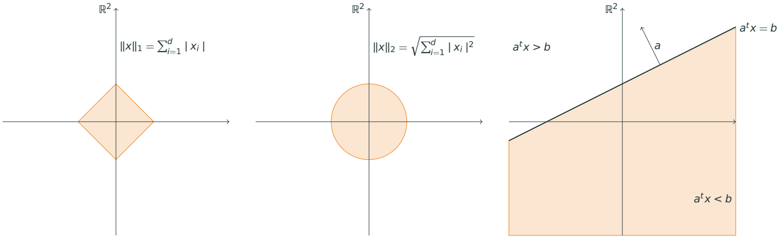
$$x, y \in C \Rightarrow \theta x + (1-\theta)y \in C, \quad \forall \theta \in [0,1]$$

with $\theta \in R$ form the line segment through x and y.

Examples of convex sets

In all the following examples, $x = (x_1, \dots, x_d)^t \in \mathbb{R}^d$

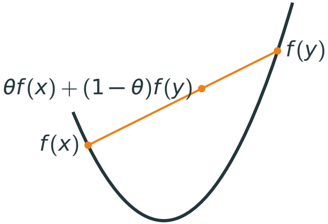
- $\{x : \|x\| \leq r\}$, for a given norm $\|\cdot\|$, $r > 0$
- $\{x : a^t x = b\}$, where $a = (a_1, \dots, a_d)^t \in \mathbb{R}^d$ and $b \in \mathbb{R}$.
- $\{x : a^t x \leq b\}$, where $a = (a_1, \dots, a_d)^t \in \mathbb{R}^d$ and $b \in \mathbb{R}$.
- $\{x : Ax \leq b\}$, where A is a $m \times d$ matrix and $b \in \mathbb{R}^m$.
- ...



Convex functions

- **Convex function:** A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $\mathrm{dom}(f)$ is convex and for all $x, y \in \mathrm{dom}(f)$ and $\theta \in [0, 1]$

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$



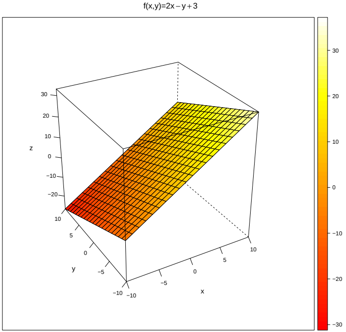
- **Strictly convex function:** A function $f : \mathbb{R}^d \to \mathbb{R}$ is strictly convex if $\mathrm{dom}(f)$ is convex and for all $x, y \in \mathrm{dom}(f)$ with $x \neq y$ and $\theta \in (0, 1)$

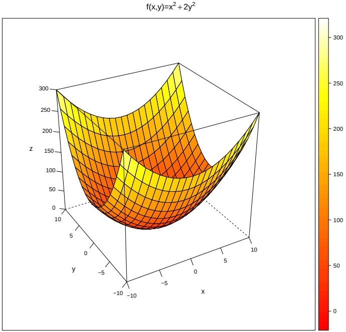
$$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y)$$

Examples of convex functions

In all the following examples, $x = (x_1, \dots, x_d)^t \in \mathbb{R}^d$

- Affine functions $f(x) = a^t x + b$, where , $a = (a_1, \dots, a_d)^t \in \mathbb{R}^d$ and $b \in \mathbb{R}$.
- Quadratic forms $f(x) = x^t A x$, where A is semidefinite positive $d \times d$ matrix.
- Least squares loss $f(x) = \|y - Ax\|_2^2$, where $y = (y_1, \dots, y_p)^t \in \mathbb{R}^p$ and A is a $p \times d$ matrix
- Norm function $f(x) = \|x\|$ for any norm
- ...



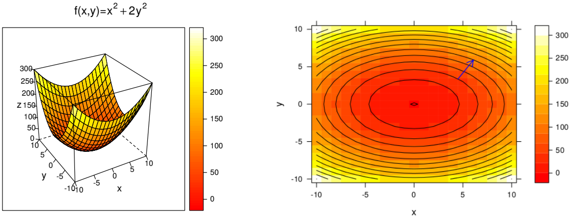


First and second order characterizations

- Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable. The gradient of f at $x \in \mathbb{R}^d$ is given by

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_d} \end{pmatrix}$$

- Recall that gradient vector give us the direction of greatest increase of f



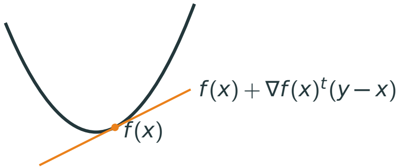
- First-order Taylor approximation: Given $f : \mathbb{R}^d \to \mathbb{R}$ differentiable

$$f(y) \approx f(x) + \nabla f(x)^t (y - x)$$

First and second order characterizations

- First order characterization: Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable. Then f convex if and only if $\text{dom}(f)$ is convex and for all $x, y \in \text{dom}(f)$

$$f(y) \geq f(x) + \nabla f(x)^t (y - x)$$



First-order Taylor approximation is a global underestimator of f

- Note that, if $\nabla f(x) = 0$ then for all $y \in \text{dom}(f)$ we have $f(y) \geq f(x)$, that is, x is a global minimizer of f

First and second order characterizations

- Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable. The Hessian of f at $x \in \mathbb{R}^d$ is given by

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_d} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_d \partial x_1} & \frac{\partial^2 f(x)}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_d^2} \end{pmatrix}$$

- Second-order Taylor approximation: Given $f : \mathbb{R}^d \to \mathbb{R}$ differentiable

$$f(y) \approx f(x) + \nabla f(x)^t (y - x) + \frac{1}{2}(y - x)^t \nabla^2 f(x)(y - x)$$

First and second order characterizations

- Second order characterization: Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable. Then f convex if and only if $\mathrm{dom}(f)$ is convex and for all $x \in \mathrm{dom}(f)$

$$\nabla^2 f(x) \succeq 0$$

- Geometrically, this characterization requires that the graph of the function have positive curvature at x

Convex optimization problem

- A convex optimization problem, has the form

$$\boxed{\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \ i = 1, \ldots, m \\ & a_j^t x = b_j, \ j = 1, \ldots, p \end{array}}$$

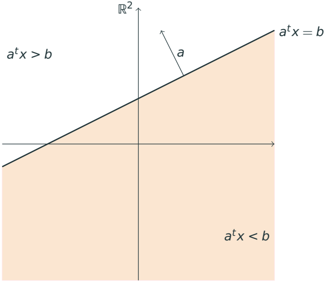
  - where $f_0 : \mathbb{R}^d \to \mathbb{R}$, and $f_i : \mathbb{R}^d \to \mathbb{R}, i = 1, \ldots, m$ are convex functions,
  - $a_j = (a_{j1}, \ldots, a_{jd})^t \in \mathbb{R}^d$ is a vector of coefficients and $b_j \in \mathbb{R}$, $j = 1, \ldots, p$.
- The feasible set of a convex optimization problem is convex
- We minimize a convex objective function over a convex set

Convex optimization problem

- For example, linear programs are convex problems with affine objective and constraint functions.

$$\boxed{\begin{array}{ll} \text{minimize} & c^t x \\ \text{subject to} & d_i^t x \leq e_i, \ i = 1, \ldots, m \\ & a_j^t x = b_j, \ j = 1, \ldots, p \end{array}}$$

  - where $x \in \mathbb{R}^d$, $c \in \mathbb{R}^d$,
  - $d_i \in \mathbb{R}^d$, $e_i \in \mathbb{R}$, $i = 1, \ldots, m$
  - $a_j \in \mathbb{R}^d$, $b_j \in \mathbb{R}$, $j = 1, \ldots, p$



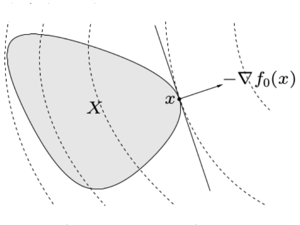
Convex optimization problem

- In a convex optimization problem any locally optimal point is also (globally) optimal
- Suppose f_0 is differentiable. Then, for all $x, y \in \mathrm{dom}(f_0)$

$$f_0(y) \geq f_0(x) + \nabla f_0(x)^t (y - x)$$

  Let X denote the feasible set of the problem. A point x is optimal if and only if $x \in X$ and for all $y \in X$,

$$\nabla f_0(x)^t (y - x) \geq 0$$



Descent methods

$$\boxed{x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)} \quad \text{with } f(x^{(k+1)}) < f(x^{(k)})}$$

- Δx is the search direction
- $t^{(k)} \geq 0$ is the step size
- From Taylor expansion, we have that

$$f(x^{(k+1)}) \approx f(x^{(k)}) + \nabla f(x^{(k)})^t (x^{(k+1)} - x^{(k)}).$$

  Therefore, if we want $f(x^{(k+1)}) < f(x^{(k)})$, we must choose a search direction such that

$$\nabla f(x^{(k)})^t \Delta x^{(k)} < 0.$$

  The vector minimizing that product is $\Delta x^{(k)} = -\nabla f(x^{(k)})$

Descent methods

Algorithm: General descent method

given a start point $x \in \operatorname{dom}(f)$
repeat

1. Determine a descent direction Δx
2. Line search. Choose a step size $t > 0$
3. Update. $x := x + t\Delta x$

until stopping criterion is satisfied

Gradient descent method

- A natural choice for the search direction is the negative gradient $\Delta x = -\nabla f(x)$.
- The resulting algorithm is called the gradient algorithm or gradient descent method

---

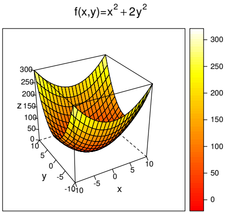
Algorithm: Gradien descent method

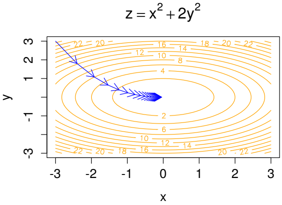
---

given a start point $x \in \operatorname{dom}(f)$
repeat
1. $\Delta x = -\nabla f(x)$
2. Line search. Choose a step size $t > 0$
3. Update. $x := x + t\Delta x$

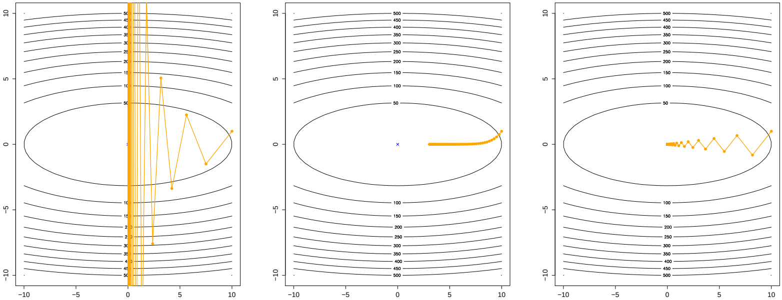
until stopping criterion is satisfied

- The stopping criterion is usually of the form $\|\nabla f(x)\|^2 \leq \eta$, for η small





Gradient descent methods: How do we choose the stepsize?



Gradient descent methods: How do we choose the stepsize?

- Exact line search
$$t = \arg\min_{s \geq 0} f(x + s\Delta x)$$
- Backtracking line search. One inexact line search method that is very simple and quite effective is called backtracking line search.

Algorithm: Backtracking line search

given a descent direction Δx for f at $x \in \text{dom}(f)$, $\alpha \in (0, 0.5]$ and $\beta \in (0, 1)$
$t := 1$
while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^t \Delta x, \quad t := \beta t$

References

Boyd, S. and Vandenberghe, L. (2004).
Convex optimization.

Tibshirani, R.
Course on Convex Optimization. Carnegie Mellon University - CMU

Duchi, J.
Course on Introduction to Convex Optimization for Machine Learning. University of California, Berkeley