Mathematical Preliminaries

Chia-Hui Chang (張嘉惠)
National Central University
2008.9.17

Outline

- Linear Algebra
- Concepts from Geometry
- Elements of Differential Calculus

Data Fitting

- Examples
  - Find the least-squares solution $\vec{x}^*$ of the system $A\vec{x} = \vec{b}$, where

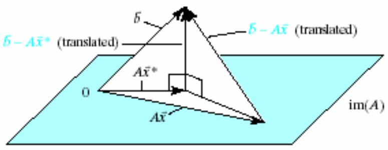
$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \text{ and } \vec{b} = \begin{bmatrix} 0 \\ 0 \\ 6 \end{bmatrix}$$

  - Fit a quadratic function to the four data points $(a_1, b_1) = (-1, 8)$, $(a_2, b_2) = (0, 8)$, $(a_3, b_3) = (1, 4)$, and $(a_4, b_4) = (2, 16)$.

Data Fitting

- Given data points $(a_1, b_1), (a_2, b_2), \ldots, (a_n, b_n)$
- L_2 regression (QP): $\arg\max_{\mathbf{x}} \sum_j (b_j - \mathbf{x}a_j)^2$
- L_1 regression: $\arg\max_{\mathbf{x}} \sum_j \| b_j - \mathbf{x}a_j \|$
- Least-square regression (L_2)

$$\| \vec{b} - A\vec{x}^* \| \leq \| \vec{b} - A\vec{x} \|$$



Quadratic Form

- Example
  - Consider the quadratic form
    $$q(x_1, x_2, x_3) = 9x_1^2 + 7x_2^2 + 3x_3^2 - 2x_1x_2 + 4x_1x_3 - 6x_2x_3.$$
    Find a symmetric matrix *A* such that $q(\vec{x}) = \vec{x} \cdot A\vec{x}$ for all $\vec{x}$ in R^3.

- Matrix form $q(\mathbf{x}) = \mathbf{x}^T Q\mathbf{x}$

Hyperplanes

- Hyperplane
  - $H= \{x \in R^n \mid u_1x_1+u_2x_2+\ldots+u_nx_n =u^Tx=v\}$
  - $a \in H$, $H= \{x \in R^n \mid u^T(x-a)=0\}$

- Halfspace
  - $H_+=\{x \in R^n \mid u^Tx \geq v\}$
  - $H_-=\{x \in R^n \mid u^Tx \leq v\}$

Convex Set

- Line segment
  - $x, y \in R^n$, $\{ v \mid v=\alpha x+(1-\alpha)y, \alpha\in[0,1] \}$
- Convex set
  - A set Θ is convex if, for all $x, y \in \Theta$, the line segment between x and y lies in Θ.

Polytopes and Polyhedra

- Convex polytope
  - A set that can be expressed as the intersection of a finite number of half-spaces
- Polyhedron
  - A nonempty bounded polytope

Differential calculus

- Consider a function f: $R^n \rightarrow R^m$

$$f(x) = \begin{bmatrix} f_1 \\ \vdots \\ f_m \end{bmatrix}$$

- Affine function
  - A function A: $R^n \rightarrow R^m$ is affine if there exists a linear function L: $R^n \rightarrow R^m$ and a vector $y \in R^m$ such that $A(x) = L(x) + y$.
  - In $R \rightarrow R$, an affine function has the form $A(x) = ax + b$, with $a, b \in R$.
- Idea: Approximating an arbitrary function f: $R^n \rightarrow R^m$ near point x_0 by an affine function L.
  - $A(x_0) = f(x_0)$
  - $A(x) = L(x - x_0) + f(x_0)$
  - $\lim_{x \to x_0} \frac{\| f(x) - A(x) \|}{\| x - x_0 \|} = 0$

$$f(x_0) = A(x_0) = L(x_0) + y$$
$$y = f(x_0) - L(x_0)$$
$$A(x) = L(x) + \left(f(x_0) - L(x_0) \right) = L(x - x_0) + f(x_0)$$
$$\lim_{x \to x_0} \frac{\left\| f(x) - \left(L(x - x_0) + f(x_0) \right) \right\|}{\left\| x - x_0 \right\|} = 0$$

Differentiability

- Differentiable
  - A function f is said to be differentiable at x_0 if there is an affine function L that approximates f near x_0; that is, there exists L: $R^n \rightarrow R^m$ such that

$$\lim_{x \to x_0} \frac{\| f(x) - (L(x - x_0) + f(x_0)) \|}{\| x - x_0 \|} = 0$$

  - The linear transformation above is called the derivative of f at x_0.
  - Any linear transformation can be represented by an *mxn* matrix.

$$Df(x_0) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x_0) & \cdots & \frac{\partial f}{\partial x_n}(x_0) \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x_0) & \frac{\partial f_1}{\partial x_2}(x_0) & \cdots & \frac{\partial f_1}{\partial x_n}(x_0) \\ \frac{\partial f_2}{\partial x_1}(x_0) & \frac{\partial f_2}{\partial x_2}(x_0) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x_0) & \frac{\partial f_m}{\partial x_2}(x_0) & \cdots & \frac{\partial f_m}{\partial x_n}(x_0) \end{bmatrix}$$

Partial Derivative

- Derivative of f: $R^n \to R^m$ at x_0

$$Df(x_0) = \left[\tfrac{\partial f}{\partial x_1}(x_0) \quad \cdots \quad \tfrac{\partial f}{\partial x_n}(x_0) \right]$$

- Partial derivative of f: $R^n \to R^m$ along e_j $\dfrac{\partial f}{\partial x_j}(x_0) = Le_j$

$$x_j = x_0 + te_j, \text{where } e_j = \quad ,$$

$$\Rightarrow \lim_{t \to 0} \frac{\left\| f(x_j) - \left(tLe_j + f(x_0)\right) \right\|}{t} = 0$$

$$\Rightarrow \lim_{t \to 0} \frac{\left\| f(x_j) - f(x_0) \right\|}{t} = Le_j.$$

if $f(x) = \begin{bmatrix} f_1 \\ \vdots \\ f_m \end{bmatrix}$, then $\dfrac{\partial f}{\partial x_j}(x_0) = \begin{bmatrix} \frac{\partial f_1}{\partial x_j}(x_0) \\ \vdots \\ \frac{\partial f_m}{\partial x_j}(x_0) \end{bmatrix}$,

and Df is the Jacbian matrix.

Special Cases

- Special cases
  - For f : $R \to R$, *Df(x)=a.*
  - For f : $R^n \to R$, *Df is a 1xn vector.*
  - For f : $R^n \to R^m$, *Df is a mxn matrix.*

$$Df(x) = \left[\frac{\partial f(x)}{\partial x_1} \quad \cdots \quad \frac{\partial f(x)}{\partial x_n} \right]$$

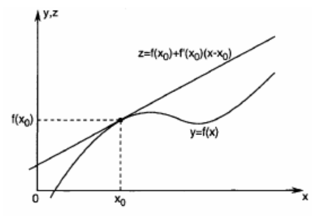


Figure 5.1 Illustration of the notion of the derivative

Gradient & Hessian Matrix

- Gradient
  - If $f: R^n \rightarrow R$ is differentiable at every point of its domain, the the gradient ∇f is defined by

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} = Df(x)^T$$

$\nabla f: R^n \rightarrow R^n$

- Hessian Matrix
  - If ∇f is differentiable,
  - then f is twice differentiable.

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Level Sets and Gradients

- Level Set
  - The level set of a function $f: R^n \rightarrow R$ at level c is the set of points S={x| $f(x)=c$}.

- Theorem
  - The gradient vector ∇f is orthogonal or normal to an arbitrary smooth curve passing through x_0 on the level set S determined by $f(x)=f(x_0)$.

    i.e. $\nabla f(x_0)^T(x-x_0)=0$, if $\nabla f(x_0) \neq 0$.

Theorem

- Orthogonality of the gradient to the level set
- Gradient is the direction of maximum rate of increase of f at x_0.

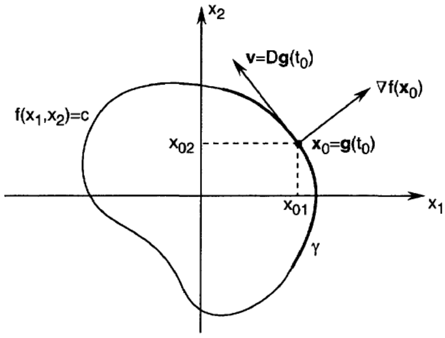


Figure 5.4 Orthogonality of the gradient to the level set

The graph of f: $R^n \to R$

- The graph of f: $R^n \to R$ is the set $\{[x^T, f(x)]^T: x \to R^n\} \subset R^{n+1}$

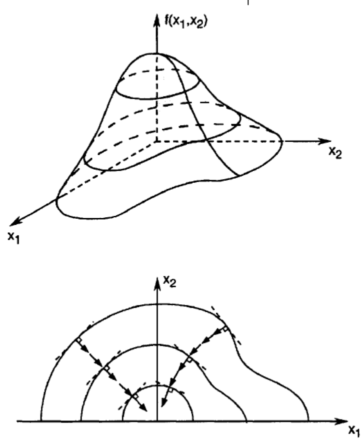


Figure 5.5 Illustration of a path of steepest ascent

Taylor's Series

- Theorem

$$f(x) = f(x_0) + \frac{(x-x_0)}{1!} f^{(1)}(x_0) + \frac{(x-x_0)^2}{2!} f^{(2)}(x_0) + \ldots + \frac{(x-x_0)^{m-1}}{(m-1)!} f^{(m-1)}(x_0) + R_m,$$

where $f^{(i)}$ is the ith derivative of f, and

$$R_m = \frac{(x-x_0)^m}{m!} f^{(m)}(x_0 + \theta(x-x_0)),$$

with $\theta \in (0,1)$.