MBML

FCN application

Experience (sensor) → Expertise

(x_i, y_i)

$x_i \in \mathbb{R}^d$

$y_i \in \{\pm 1\}$ → ok or not

$y = f(x)$

$f: \mathbb{R}^d \to \{\pm 1\}$

Experience —ML→ Expertise

f is designed such that $y_i = f(x_i)$ $\forall\ i = 1:N$

course is all about finding f to get $y = f(x)$

ML algos: issues — scalability, stability, complexity

- scalability: $O(d^3)$ for finding evalues of $A_{d \times d}$ — Not scalable
- stability: Spam account & account or account — should still recognize
- complexity: There may be time critical applications

Learning — sup — labels available; Unsup — unavailable labels

Learning algos: input ⇒ training data

Issues in ML

One major issue is datum size.

can be addressed through

- Dimensionality reduction (DR)
- feature extraction or efficient representation of data
- Several DR methods exist: SVD/PCA, manifold learning

Other issues deal w/ designing learning algos, validating such algos etc

Another example

$x_i \in \mathbb{R}^d$ → parameters on loan seeker $i = 1:N$

$y_i \in \{-1, +1\}$ → whether loan seeker is likely to default or not.

$x_i \in \mathbb{R}^d \longrightarrow z_i \in \mathbb{R}^n$ $n < d$ — projection to lower dim space

When projecting to lower dim space, we want discriminating features are retained,

Ideally, want $\|x_i - x_j\|_2 = \|z_i - z_j\|_2$

Since, equality is not possible

at least $\|x_i - x_j\|_2 \approx \|z_i - z_j\|_2$

$z_i = g(x_i)$

- z_i: vector in lower dim space
- higher dim vector
- g has to be built

We may have to reproject back to higher dim.

$x_i \in \mathbb{R}^d \to z_i \in \mathbb{R}^n$ $(n < d) \to \tilde{x}_i$

Patient in remote region → Compression (lower dim for transmission) → Reconstruction (Expert needs higher dim again)

Expectation → $x_i \simeq \tilde{x}_i$

If too much info is lost, we don't face issues

If we want small value for n, good reconstruction. } strike a balance

So in $z_i = g(x_i)$, g can't be arbitrary

- PCA: method of random projection
- Linear (simple)
- non-linear (can get small n) — manifold learning

Linear vs non linear

- Simple
- can't get small n (by Lemma)
- can get small n

Creators of manifold learning

↳ projected faces to get moods on 2D plane

Summary

$(x_i, y_i)_{i=1}^N$, $x_i \in \mathbb{R}^d$

↓ if d is big

Dimensionality reduction

linear	nonlinear
• PCA • random projection	• ~~Fractals~~ fractals • manifold learning

↓

$(x_i, y_i)_{i=1}^N$ small dim.

Learning.

x_i

- supervised (x_i, y_i) (y_i = label)
  - SVM
  - Neural nets
- unsupervised
  - K means
  - Sparsely based clustering

SVD

$A_{m \times n}$ rank $(A) = r \leq \min(m, n)$

SVD: $\exists$ ~~U_1~~ $U_{m \times m}$, $V_{n \times n}$ (both being unitary) and a "diagonal" matrix $D_{m \times n}$ such that $A = U D V^T$

Proof of $A = UDV^T$

- $(A^TA)_{n \times n}$ ~~real~~ → symmetric
- Its e.values must be ≥ 0

$(A^TA) v_i = d_i^2 v_i$

λ_i ~~$\neq$~~ $= 0$ $i > r$ (where r is rank)

Define $u_i \in \mathbb{R}^m$ s.t.

$u_i = \frac{1}{\lambda_i} A v_i$ ~~$i = 1 : r$~~

$u_i^T u_j = \frac{1}{\lambda_i \lambda_j} (Av_i)^T (A v_j)$

$= \frac{1}{\lambda_i \lambda_j} v_i^T (A^TA) v_j$

$= \frac{1}{\lambda_i \lambda_j} v_i^T \lambda_j^2 v_j$

$u_i^T u_j = \frac{v_i^T v_j}{\lambda_i} \lambda_j$

$= \frac{\lambda_j}{\lambda_i} (v_i^T v_j) \rightarrow \delta_{ij} = 1 \; i = j$, $= 0 \; i \neq j$

$u_1, u_2, \ldots, u_r$ are ~~LD~~ LI

$\{u_1, u_2, \ldots, u_r\}$ can be extended to a basis say $\{u_1, u_2, \ldots, u_m\}$

AV

$= [Av_1 \; Av_2 \; \ldots \; Av_n]$

$= [\lambda_1 u_1 \; \lambda_2 u_2 \; \ldots \; 0 \cdot u_{r+1} \; \ldots]$ (0s)

$= \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_r \\ & & & \end{bmatrix} \begin{bmatrix} u_1 & u_2 & \ldots & u_m \end{bmatrix} \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_r \end{bmatrix}$

$AV = UD$ or ~~A~~ $A = UDV^T$

Aug 04 MBML

Basic optimization theory

The following are some opt. problems

- $\min_{U,W} \sum_{i=1}^{N} \|x_i - UWx_i\|_2^2$ (reconstruction, compression)
- $\min_{\mu_i \in S} \sum_{i=1}^{k} \|x_i - \mu_i\|$

} unconstrained optimization problem

- $\min \left(\frac{1}{2} \|w\|_2^2 \text{ such that } y_i(\langle w, x_i \rangle + b) \geq 1 \right)$

} constrained optimization problem

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

Directional derivative

Suppose $d \in \mathbb{R}^n$

$f'(x,d) = \lim_{t \to 0} \frac{f(x+td) - f(x)}{t}$

Suppose $d = e_1 = (1, 0, 0, \ldots 0)$

then $f'(x, e_1) = \frac{\partial f}{\partial x_1}$

directional derivative along d

Suppose all partial derivatives exist for f, then gradient of f is defined by

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

$$f'(x^*, d) = \nabla f(x^*) \cdot d \quad \text{(dot)}$$

Suppose double partial derivatives exist for f (continuous), at all points, Hessian is defined by

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

Basic opt theory

$f: \mathbb{R}^n \to \mathbb{R}$ and $0 \neq d \in \mathbb{R}^n$

if $\lim_{t \to 0} \frac{f(x_0 + d) - f(x_0)}{t}$ exists

$= f'(x_0, d)$ (notation) → directional derivative of f at x_0 along direction d

If $d = e_1 = (1, 0, 0, \ldots 0)$, then

$f(x_0, d) = \frac{\partial f}{\partial x_1}(x_0)$

If $f'(x_0, e_i)$ exists $\forall\ i = 1:n$,

$$\nabla f(x_0) = \begin{pmatrix} \frac{\partial f(x_0)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x_0)}{\partial x_n} \end{pmatrix}$$ → Gradient of f at x_0

$f'(x_0, d) = \nabla' f(x_0) \cdot d$

Suppose f has continuous partial derivatives of second order

$$H = [\nabla^2 f(x_0)] = \begin{pmatrix} \frac{\partial^2 f(x_0)}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x_0) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x_0) \\ \vdots & & & \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x_0) & - & - & \frac{\partial^2 f}{\partial x_n^2}(x_0) \end{pmatrix}_{n \times n}$$

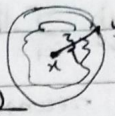
(notation)

Hessian of f at x_0.

First order approximation theorem

Suppose $f: \mathbb{R}^n \to \mathbb{R}$ has all continuous derivative of 2^{nd} order and $x, y \in \mathbb{R}^n$.

Then $\exists$ "$z \in [x,y]$"



~~$f(y) = f(x) + f(y-x)^T \nabla f(x) + \frac{1}{2}(y-x)$~~

$$f(y) = f(x) + f(y-x)^T \nabla f(x) + \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x)$$

Second order approximation theorem

Suppose $f: \mathbb{R} \to \mathbb{R}$ has continuous partial derivatives of 2^{nd} order

Then ~~$f(x) =$~~

$$f(y) = f(x) + (y-x)^T \nabla f(x) + \frac{1}{2}(y-x)^T (\nabla^2 f(x))(y-x) + o(\|y-x\|^2)$$

↳ accounting for residual

Little-o notation

$f(x) = o(g(x))$

$g(a) \neq 0$, $\lim_{x \to a} \frac{f(x)}{g(x)} = 0$

$f_n = o(g_n)$ $\lim_{n \to \infty} \frac{f_n}{g_n} = 0$

Supp $f: \mathbb{R}^n \to \mathbb{R}$ is "sufficiently nice" (all partial derivatives should exist)

~~Thm~~: Theorem

If x^* is a local minimal (or max) point of f, then $\nabla f(x^*) = 0$

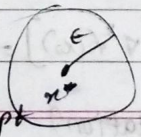
↳ vector.

proof: $g(t) = ~~f(x^*)~~ f(x^* + te_i)$

for some $i \in \{1, \dots, n\}$

$g(t) = f(x^* + te_i) \geq f(x^*) = g(0)$

$\forall t$ such that $x^* + te_i$ lies in the nbd of x^*



$\Rightarrow 0$ is a local min pt of f or g

$\therefore g'(0) = 0$

$g(t) = \frac{\partial f}{\partial x_i}(x^* + te_i)$

$g'(0) = 0 \Rightarrow \frac{\partial f}{\partial x_i}(x^*) = 0$

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

$\therefore$ If x^* is an optimal point then $\nabla f(x^*) = 0$

How to know if a point x^* is maximum or minimum

↓

we get from 2^{nd} order ~~approx~~ approx theorem.

Results

1) If x^* is a local minimum pt for f, then $(\nabla^2 f(x^*)) \geq 0$

positive semi definite (PSD)

A is PSD if $x^T A x \geq 0 \ \forall \ x \in \mathbb{R}^n$

proof

from 2^{nd} order approx theorem

$$f(y) = f(x) + (y-x)^T \nabla f(x) + \frac{1}{2}(y-x)^T (\nabla^2 f(x))(y-x) + o(\|y-x\|^2)$$

$$0 \leq f(y) - f(x^*) = \frac{1}{2}(y-x^*)^T \nabla^2 f(z)(y-x^*)$$

Aug 07

Recall:

$\min_{x \in \Omega} f(x)$ $\quad \Omega = \mathbb{R}^n$ $\quad S \subseteq \mathbb{R}^n$

unconstrained optimization problem

If f has a local $\min_{\max}$ at x^* then $\nabla f(x^*) = 0$

1st order approximation theorem

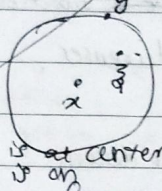
f is sufficiently nice

$x, y \in \mathbb{R}^n$

$\exists$ "$z \in [x, y]$" such that

$f(y) = f(x) + (y-x)^T \nabla f(x) + \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x)$

$\|x - z\|_2 < \|x - y\|_2$



x is ~~at~~ center
y is on circumference
inside

vector $\begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \end{bmatrix}_{n \times 1}$ $\quad \begin{bmatrix} \frac{\partial^2 f}{\partial x_i \partial x_j} \end{bmatrix}_{n \times n}$

2nd order approximation theorem

$f(y) = f(x) + (y-x)^T \nabla f(x) + \frac{1}{2}(y-x)^T \nabla^2 f(x)(y-x) + o(\|y-x\|^2)$

If f has a ~~min~~ local min pt at x^*, then $[\nabla^2 f(x^*)] \geq 0$

positive semidefinite (all eigen values ≥ 0)

$f(y) - f(x^*) = \frac{1}{2}(y-x^*)^T \nabla^2 f(z)(y-x^*)$

$y \in B(x^*, r)$ ← ball, small

$\Rightarrow$ all y ~~big~~ lies close to x^*.

for all y close to x^*,

If It is PSD matrix, so

$f(y) - f(x^*) \geq 0 \quad \forall y$

So x^* is local ~~optima~~ minima

If f has a local max at x^*

then $[\nabla^2 f(x)] \leq 0$.

If $[\nabla^2 f(x^*)] > 0$, then f has a strict min pt at x^*.

Saddle point

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

$[\nabla^2 f(x^*)]$ is neither PSD nor PND (NSD?) (positive negative? definite?)

Global min pt

$f : \mathbb{R}^n \to \mathbb{R}$

$[\nabla^2 f(x)] \geq 0 \quad \forall x \in \mathbb{R}^n$

The critical point x^* is a global min pt.

Example

$f(x) = x^T A x + b^T x + c, \quad b \in \mathbb{R}^n, \ c \in \mathbb{R}$

gradient $\nabla f(x) = 2Ax + 2b$

Hessian $\nabla^2 f(x) = A$

$\nabla f(x) = 0 \Rightarrow Ax = -b$ ①

If $\nabla^2 f(x) = A \geq 0$ (PSD) $\forall x \in \mathbb{R}^n$

then each solution of ① is a min pt for $\mathbb{R}^n$

If $\nabla^2 f(x) = A \leq 0$ (PND) $\forall x \in \mathbb{R}^n$

then each soln of ① is a global max pt for $\mathbb{R}^n$

If A is PD (strictly P.D) (all e.values are strictly > 0)

then A is invertible and

$x = -A^{-1}b$ is the critical point

If "$A \succ 0$" $x = -A^{-1}b$ is the global min pt

If "$A \prec 0$" $x = -A^{-1}b$ is the global max pt

Theorem

Suppose $f(x) = x^TAx + 2b^Tx + c$, ($b \in \mathbb{R}^n$, $c \in \mathbb{R}$, and A is symmetric)

Then the following are equivalent.

① $f(x) \geq 0 \quad \forall x \in \mathbb{R}^n$

② $\begin{bmatrix} A & b \\ b^T & c \end{bmatrix}_{(n+1)\times(n+1)} \succeq 0$ → (is PSD)

- Suppose ② holds.

$$\begin{pmatrix} x \\ 1 \end{pmatrix}^T \begin{bmatrix} A & b \\ b^T & c \end{bmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix} \geq 0 \Rightarrow f(x) \geq 0 \quad \forall x \in \mathbb{R}^n$$

- Suppose ① holds

consider $\begin{pmatrix} x \\ t \end{pmatrix}^T \begin{bmatrix} A & b \\ b^T & c \end{bmatrix} \begin{pmatrix} x \\ t \end{pmatrix} \quad \forall \begin{pmatrix} x \\ t \end{pmatrix} \in \mathbb{R}^{n+1}$

$$= \begin{bmatrix} x^TA + tb^T & x^Tb + tc \end{bmatrix} \begin{pmatrix} x \\ t \end{pmatrix}$$

$$= x^TAx + tb^Tx + x^Tbt + t^2c$$

$$= x^TAx + t x^Tb + t b^Tx + t^2c$$

$$= t^2\left[\left(\frac{x^T}{t}\right)A\left(\frac{x}{t}\right) + 2b^T\left(\frac{x}{t}\right) + c\right]$$

$$= t^2 f\left(\frac{x}{t}\right) \geq 0$$

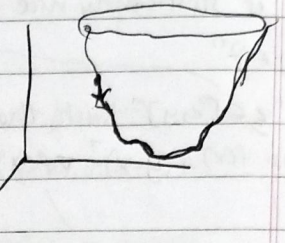
$\min_{x \in \mathbb{R}^n} f(x)$

$\nabla f(x) = 0 \Rightarrow Ax = b$ (b or −b?)

(*) $n\times n$ $n\times 1$

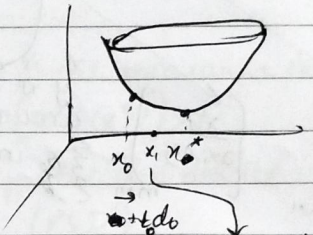
Analytical method is harder to use

Descent method: To solve (*) iteratively



start at some point

There is a direction from the point where the slope of function value decreases.



- find the descent direction d_k
- find the step size t_k

$x_{k+1} = x_k + t_k d_k$ $\left[f(x_{k+1}) < f(x_k)\right.$ $\left| f(x_0) > f(x_1) = f(x_0 + t_0 d_0)\right.$

- Stopping criterion

$\|\nabla f(x_k)\| < \epsilon$

2 things to find

1. t_k
2. d_k

Lemma: (Sufficient decrease lemma)

Suppose $f: \mathbb{R}^n \to \mathbb{R}$ is sufficiently nice and $0 \neq d \in \mathbb{R}^n$ is a descent direction, and $\alpha \in (0,1)$

Then $\exists \epsilon > 0$ s.t.

$$0 \geq f(x) - f(x+td) \geq -t\alpha\left[\nabla f(x) \cdot d\right]$$

$\forall t \in [0, \epsilon]$

↓ Sufficient decrease lemma

Aug 08

d_k is a descent direction

$f'(x_k, d_k) = \nabla f(x_k) \cdot d_k \overset{<0}{\underset{\geq 0}{}}$

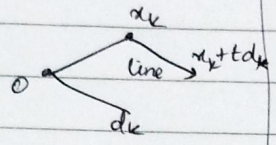
- stopping criterion
  - $\| \nabla f(x_k) \| \leq \epsilon$
  - $\| x_{k+1} - x_k \|_2 < \epsilon$

on finding t_k and d_k :

Finding t_k is referred to as Line search

Define $g(t) = f(x_k + t d_k)$

ways of finding t_k



- Constant step size

  $t_k = \bar{t} \quad \forall k$

  Pros: Simplicity

  Cons: At some k, might be good, at other k, might be bad.

- Exact line search

  $\min_{t \geq 0} g(t)$ (now this in itself is an optimization problem)

  We want $g(t) = f(x_k + t d_k)$ to be a "good" function
  - Not be non-convex
  - Not be ugly looking function that makes everything more complex.

  $t_k = \arg\min_{t \geq 0} g(t)$

- Backtracking

  $\alpha, \beta \in (0,1) \quad s > 0$

  start with $t_k = s$

  while $f(x_k) - f(x_k + t_k d_k) < -t_k \alpha \nabla f(x_k) \cdot d_k$

  $t_k \longrightarrow \beta \cdot t_k$ (we want sufficient decrease lemma to hold)

In other words $t_k = \beta^{i_0}$'s where i_0 is the smallest +ve integer such that

$f(x_k) - f(x_k + t_k d_k) \geq -t_k \alpha \nabla f(x_k) d_k$ holds

adv	disadv	
Simple	Sometimes not useful	Constant step size
Theoretically ideal	Not always practical	Exact line search
Practical		Backtracking

Example:

$f(x) = x^T A x$ where A is P.D. and symmetric

Find expression for t_k

$g(t) = f(x_k + t d_k)$

$f(x) = x^T A x$

$g(t) = (x_k + t d_k)^T A (x_k + t d_k)$

$= (x_k^T + t d_k^T) A (x_k + t d_k)$

$= x_k^T A x_k + t^2 d_k^T A d_k + t x_k^T A d_k + t d_k^T A x_k$

$g'(t) = 2t\, d_k^T A d_k + x_k^T A d_k + d_k^T A x_k$

$g'(t) = 0 \Rightarrow t_k = \dfrac{-(x_k^T A d_k + d_k^T A x_k)}{2 d_k^T A d_k}$

$= -\dfrac{2 x_k^T A d_k}{2 d_k^T A d_k}$

Finding d_k $\rightarrow f'(x_k, d)$

$\min_{d \in \mathbb{R}^n} \nabla f(x_k) \cdot d$

$\|d\|_2 = 1$

Consider

$\nabla f(x_k) \cdot d \geq -\|\nabla f(x_k)\|_2 \underbrace{\|d\|_2}_{=1}$ (Cauchy Schwartz)

$\nabla f(x_k) \cdot d \geq -\|\nabla f(x_k)\| \quad \forall\, 0 \neq d \in \mathbb{R}^n$ where $\|d_2\| = 1$ (*)

lower bound

Set $d = -\dfrac{\nabla f(x_k)}{\|\nabla f(x_k)\|_2}$

$d_k^T A x_k = (-)(|||)(|)$

$x_k = \begin{bmatrix} \\ \\ \end{bmatrix} \quad d_k = \begin{bmatrix} \\ \\ \end{bmatrix}$

$x_k^T A d_k = (-)(|||)(||)$

$= (-)[1]$

$= [\,]$

$x_k A d_k^T = [\,][\,]$

Set $d = \frac{-\nabla f(x_k)}{\|\nabla f(x_k)\|_2}$

$$\nabla f(x_k)^T \left(\frac{-\nabla f(x_k)}{\|\nabla f(x_k)\|_2} \right) = \nabla f(x_k) \cdot \left(\frac{-\nabla f(x_k)}{\|\nabla f(x_k)\|} \right)$$

$$= -\|\nabla f(x_k)\| \quad (**)$$

From (*), (**), optimal direction is

$$d_k = \frac{-\nabla f(x_k)}{\|\nabla f(x_k)\|}$$

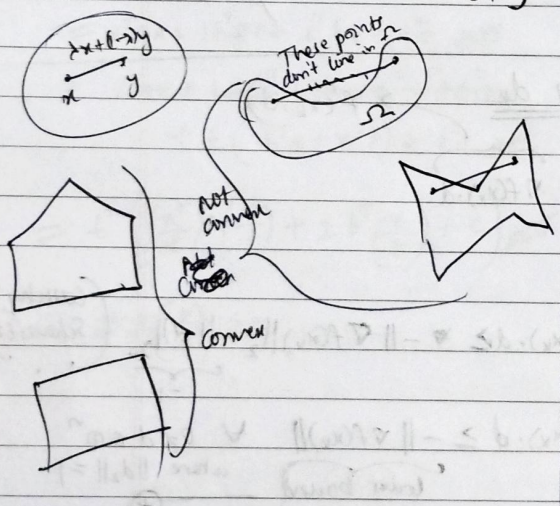
Gradient (or steepest descent) method:

- $x_{k+1} = x_k - t_k \nabla f(x_k)$
- t_k can be obtained via line search

Convex sets

$\Omega \subseteq \mathbb{R}^n$ is convex

if $[\lambda x + (1-\lambda) y] \in \mathbb{R}$ $\forall (x,y) \in \Omega$ and $\forall \lambda \in [0,1]$



Results: Suppose $\{C_i\}_{i \in I}$ is a set of convex sets where $C_i \subseteq \mathbb{R}^n$ and I is finite or infinite.

- $\bigcap_{i \in I} C_i$ is also convex
- $\lambda_1 C_1 + \lambda_2 C_2 + \dots + \lambda_n C_n$ is also convex.
- $C_1 \times C_2 \times C_3 \times \dots \times C_n \times \dots$ is also convex.
- Let $A \in \mathbb{R}^{m \times n}$

  $A(C) = \{Ax \mid x \in C\}$ is also convex (↑ convex).

$$\begin{bmatrix} \alpha_1 & \alpha_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

$$\begin{bmatrix} \alpha_1 a_{11} + \alpha_2 a_{21} & \alpha_1 a_{12} + \alpha_2 a_{22} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

$$\sum \beta_i (\alpha_1 a_{1i} + \alpha_2 a_{2i})$$