« kernel methods » @ AMMI, Rwanda

$\boxed{\text{TP 1}}$

YUNLONG JIAO 21/01/2019

PART I : Basics of Linear Algebra and Matrix Calculus

only real matrices unless o.w. s.

Def. (Matrix): Let $a_{ij} \in \mathbb{R}$ $i=1,\dots,m$, $j=1,\dots,n$.

$${}_m A_n = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Rmk: ① ${}_n A_n$ square matrix

② ${}_n A_1 := a \in \mathbb{R}^n$ vector

Conv. (bold print) A, B, X, U, V matrices

(bold print) $\vec{a}, \vec{b}, \vec{x}, \vec{y}$, vectors

$\lambda, \alpha, a_i, x_i, a_{ij}$ scalars (reals)

(bold print) $O_n := [0]$ null matrix

(bold print) $I_n := \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}$ identity matrix

Def. (Matrix Multiplication)

$${}_m C_p = {}_m A_n B_p \Leftrightarrow c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

compatible for mult.

$$i\begin{pmatrix} & j & \\ & \cdot & \\ & & \end{pmatrix} = i\begin{pmatrix} \quad \text{———} \quad \end{pmatrix}\begin{pmatrix} & \Big| & \\ & j & \end{pmatrix}$$

Rmk: ① ${}_m z_1 = {}_m A_n x_1 \Leftrightarrow z_i = \sum_{j=1}^{n} a_{ij} x_j$

② $\alpha = {}_1 y^T {}_m A_n x_1 \Leftrightarrow \alpha = \sum_{j=1}^{n} \sum_{k=1}^{m} a_{kj} x_j y_k$

Prop. (Multiply partitioned matrices)

$$
{}_{m}A_{n} = \begin{bmatrix} {}_{m_1}A_{11\,n_1} & {}_{m_1}A_{12\,n_2} \\ {}_{m_2}A_{21\,n_1} & {}_{m_2}A_{22\,n_2} \end{bmatrix} \qquad {}_{n}B_{p} = \begin{bmatrix} {}_{n_1}B_{11\,p_1} & {}_{n_1}B_{12\,p_2} \\ {}_{n_2}B_{21\,p_1} & {}_{n_2}B_{22\,p_2} \end{bmatrix}
$$

$$
{}_{m}AB_{p} = \begin{bmatrix} {}_{m_1}(A_{11}B_{11} + A_{12}B_{21})_{p_1} & {}_{m_1}(A_{11}B_{12} + A_{12}B_{22})_{p_2} \\ {}_{m_2}(A_{21}B_{11} + A_{22}B_{21})_{p_1} & {}_{m_2}(A_{21}B_{12} + A_{22}B_{22})_{p_2} \end{bmatrix}
$$

Def. (Basic matrix operations)

Transpose ${}_{n}A^{T}_{m} = [a_{ji}]$

For square matrix, $\left\{ \begin{array}{l} \text{Inverse } {}_{n}A^{-1}_{n} \text{ s.t. } A^{-1}A = AA^{-1} = I \\ \text{Trace } \operatorname{tr}({}_{n}A_{n}) = \sum_{i=1}^{n} a_{ii} \end{array} \right.$

Prop. (Invertible matrices)

${}_{n}A_{n}$ invertible (A^{-1} exists)

$\Leftrightarrow$ $Ax=0$ has only the trivial solution $x=0$.

i.e. $\ker(A) = 0$.

i.e. Columns of A are linearly independent.

Ex. ① $(AB)^T = B^T A^T$

always assumed invertible if -1 exists.

$(AB)^{-1} = B^{-1}A^{-1}$

always assumed compatible for mult.

$(A^T)^{-1} = (A^{-1})^T$

② $\operatorname{tr}(A+B) = \operatorname{tr}(A) + \operatorname{tr}(B)$

$\operatorname{tr}(\lambda A) = \lambda \operatorname{tr}(A)$

$\operatorname{tr}(AB) = \operatorname{tr}(BA)$

$\operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \operatorname{tr}(CAB)$

③ $\operatorname{tr}({}_{n}A A^{T}_{n}) = \sum_{i,j} a_{ij}^2 := \|A\|_F^2$ Frobenius norm of matrices

Ex: (Matrix Inversion Lemma)

① $(A + {}_{n}U_{m}C_{m}V_{n})^{-1} = A^{-1} - A^{-1}U(C^{-1} + {}_{m}V_{n}A^{-1}{}_{n}U_{m})^{-1}VA^{-1}$

② $(I + \vec{u}\vec{v}^T)^{-1} = I - \dfrac{uv^T}{1 + v^Tu}$

Ex: (Schur complement, Invert partitioned matrix)

Suppose $M = \left[\begin{array}{c|c} {}_{p}A_{p} & B \\ \hline C & {}_{q}D_{q} \end{array}\right]$, A, D, M invertible.

Denote Schur complements by

$$M/D := A - BD^{-1}C$$

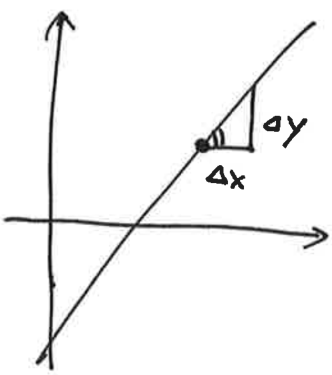
$$M/A := D - CA^{-1}B$$

Show that

$$M^{-1} = \left[\begin{array}{c|c} (M/D)^{-1} & -(M/D)^{-1}BD^{-1} \\ \hline -D^{-1}C(M/D)^{-1} & (M/A)^{-1} \end{array}\right]$$

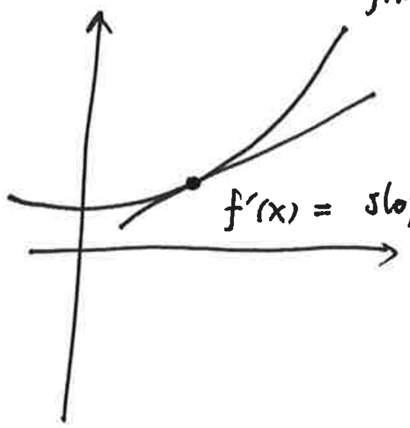
Interlude: Derivative (of a function of a real variable)

$y = f(x)$ $\qquad \dfrac{dy}{dx} = \lim_{h\to 0} \dfrac{f(x+h) - f(x)}{h} \quad (= f'(x))$



$f(x) = kx$.

$f'(x) = k$



$f(x)$

$f'(x) =$ slope of the tangent line at x.

Def. (Matrix Derivatives, Jacobian, & Hessian)

① $\underset{m\ 1}{\vec{y}} = \vec{f}(\underset{n\ 1}{\vec{x}})$ vector-valued function of vectors

Jacobian of $\vec{f}$ (matrix): $\dfrac{\partial y}{\partial x} = \begin{bmatrix} \dfrac{\partial y_1}{\partial x_1} & \cdots & \dfrac{\partial y_n}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial y_m}{\partial x_1} & \cdots & \dfrac{\partial y_m}{\partial x_n} \end{bmatrix}$

② $\alpha = f(\underset{n\ 1}{\vec{x}})$ real-valued function of vectors

Jacobian of f (vector): $\dfrac{\partial \alpha}{\partial x} = \begin{pmatrix} \dfrac{\partial \alpha}{\partial x_1} \\ \vdots \\ \dfrac{\partial \alpha}{\partial x_n} \end{pmatrix} := J(\alpha)$

☆ Hessian of f (matrix): $\dfrac{\partial^2 \alpha}{\partial x \partial x^T} = \begin{bmatrix} \dfrac{\partial^2 \alpha}{\partial x_1^2} & \cdots & \dfrac{\partial^2 \alpha}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 \alpha}{\partial x_n \partial x_1} & \cdots & \dfrac{\partial^2 \alpha}{\partial x_n^2} \end{bmatrix} := H(\alpha)$

③ $A = f(\alpha)$ matrix-valued function of reals

Jacobian of f (matrix): $\dfrac{\partial A}{\partial \alpha} = \left[\dfrac{\partial A_{ij}}{\partial \alpha} \right]$

④ $\alpha = f(A)$ real-valued function of matrices

Jacobian of f: $\dfrac{\partial \alpha}{\partial A} = \left[\dfrac{\partial \alpha}{\partial A_{ij}} \right]$

Rmk: ① All defined by element-by-element derivatives of real functions

② The order of Jacobian may depend on A, or A^T by convention.

Ex:

① $\vec{y} = A\vec{x}$, $\quad \frac{\partial y}{\partial x} = A$

$\quad m\ 1 \quad m\ n\ 1$

proof. $y_i = \sum_{j=1}^{n} a_{ij} x_j$, $\quad \frac{\partial y_i}{\partial x_j} = a_{ij}$

② $\alpha = y^T A x$

$\quad 1\ m\ n\ 1$

$$\frac{\partial \alpha}{\partial x} = A^T y, \quad \frac{\partial \alpha}{\partial y} = Ax$$

③ $\alpha = x^T A x$

$$J_x(\alpha) = (A + A^T)\, x$$

$$H_x(\alpha) = A + A^T$$

④ $\varepsilon = \|y - Ax\|_2^2 = \sum_i (y_i - \sum_j a_{ij} x_j)^2$

$$\frac{\partial \varepsilon}{\partial x} = -2A^T(y - Ax)$$

~~Prop~~. Ex (Chain rule):

① $\vec{y} = A\vec{x}$, $\vec{x}$ is a function of $\vec{z}$

(A: Not a function of $\vec{z}$, unless specified!)

$$\frac{\partial y}{\partial z} = A\frac{\partial x}{\partial z}$$

proof. LHS: by def. $\frac{\partial y}{\partial z} = \left[\frac{\partial y_i}{\partial z_j}\right]$, $\quad y_i = \sum_{k=1}^{n} a_{ik} x_k$

$\Rightarrow \frac{\partial y_i}{\partial z_j} = \sum_{k=1}^{n} a_{ik} \frac{\partial x_k}{\partial z_j}$ by chain rule of diff. of real functions

RHS: by def. $\frac{\partial x}{\partial z} = \left[\frac{\partial x_i}{\partial z_j}\right]$

by matrix mult. $\left[A\frac{\partial x}{\partial z}\right]_{ij} = \sum_{k=1}^{n} a_{ik}\frac{\partial x_k}{\partial z_j}$

② $\alpha = \vec{y}^T A \vec{x}$, $\vec{x}, \vec{y}$ are both functions of $\vec{z}$

$$\frac{\partial \alpha}{\partial z} = x^T A^T \frac{\partial y}{\partial z} + y^T A \frac{\partial x}{\partial z}$$

Ex: ① Suppose A is a (matrix-valued) function of α

$$\frac{\partial (A^{-1})}{\partial \alpha} = - A^{-1} \frac{\partial A}{\partial \alpha} A^{-1}$$

② $$\frac{\partial}{\partial A} tr(\underset{m}{A}\underset{n}{B}_{p}) = B^T$$

$$\frac{\partial}{\partial A} tr(A) = I$$

compare
$$\left\{\begin{array}{l} \dfrac{\partial}{\partial A} tr(ABA^T) = A(B+B^T) \\ \dfrac{\partial}{\partial B} tr(ABA^T) = A^T A \end{array}\right.$$

③ $$\mathcal{E} = \|X - WH\|_F^2 = \sum_i \sum_k \left(x_{ik} - \sum_j w_{ij} h_{jk}\right)^2$$

$$\frac{\partial \mathcal{E}}{\partial W} = -2XH^T + 2WHH^T$$

P A U S E "BREAK"

Def. (Eigenvector Equation)

${}_nA_n$. $\quad A \underline{u_i} = \underline{\lambda_i}\, \underline{u_i}$ $\quad i = 1, \ldots, n$.

eigenvalue

eigenvector

Rmk. rank(A) = number of non-zero eigenvalues.

Def. (Symmetric Matrix)

${}_nA_n$ Symmetric iff $A = A^T$

Def. (Orthogonal Matrix)

${}_nU_n$ orthogonal iff $U^{-1} = U^T$ $\left(\begin{array}{l} \Leftrightarrow UU^T = U^T U = I \\ \Leftrightarrow u_i^T u_j = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases} := \delta_{ij} \end{array} \right)$ why?

6

Thm. (Eigendecomposition)

Any (real) symmetric matrix ${}_nA_n$ can be decomposed:

$$ {}_nA_n = {}_nU_n \Lambda_n U^T_n $$

where U is an orthogonal matrix,
whose columns are eigenvectors of A.
$\Lambda := \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ is a diagonal matrix,
whose diagonal are eigenvalues of A. why?

Cor. ① Symmetric matrix A invertible $\Leftrightarrow$ $\lambda_i \neq 0 \quad \forall i$

② $A^{-1} = U \Lambda^{-1} U^T$, $\quad \Lambda^{-1} = \mathrm{diag}(\lambda_1^{-1}, \dots, \lambda_n^{-1})$

③ why? $tr(A) = \sum_i \lambda_i$, $\quad tr(A^{-1}) = \sum_i \lambda_i^{-1}$

Def. (Positive (semi-) definite matrix)

${}_nA_n \succeq 0$ p.s.d. iff $x^TAx \geq 0, \ \forall x \in \mathbb{R}^n$.

${}_nA_n \succ 0$ p.d. iff $x^TAx > 0, \ \forall x \in \mathbb{R}^n$.

Prop. (Eigenvalues of Symmetric p.(s.)d. matrices)

① ${}_nA_n \succeq 0$ p.s.d. and Symm. $\Leftrightarrow$ e.v. $\lambda_i \geq 0$, $i = 1, \dots, n$

② ${}_nA_n \succ 0$ p.d. and symm. $\Leftrightarrow$ e.v. $\lambda_i > 0$, $i = 1, \dots, n$ → full-rank matrix

Thm. (Singular Value Decomposition, SVD)

Any (real) matrix $\underset{m\ n}{A}$ can be decomposed as

$$\underset{m\ n}{A} = \underset{m\ m}{U}\ \underset{m\ n}{\Sigma}\ \underset{n\ n}{V^T}$$

where $\underset{m\ m}{U}, \underset{n\ n}{V}$ are orthogonal matrices

whose columns are left- and right- singular vectors of A

$\underset{m\ n}{\Sigma}$ is a "diagonal" matrix $\begin{pmatrix} \ddots & \vdots & 0 \end{pmatrix}_{m<n}$ OR $\begin{pmatrix} \ddots \\ 0 \end{pmatrix}_{m>n}$ with non-negative

real numbers on diagonal, called Singular values of A.

denoted by $\sigma_i \geq 0$, $i = 1, \ldots, \min(m,n)$.

Ex: For any general $\underset{m\ n}{X}$. show that

① X^TX and XX^T are symmetric, p.s.d.

② The non-zero eigenvalues of X^TX and XX^T are the same, that are $\{\sigma_i^2 \mid \sigma_i \neq 0,\ i=1,\ldots,\min(m,n)\}$ where σ_i's are singular values of X.

PART II: Intro to Constrained Convex optimization

Linear Regression:

$\mathcal{D} = \{(\vec{x}_i, y_i)\}_{i=1}^{m}$, $\quad \underbrace{y_i}_{\text{observations}} = \underbrace{\vec{w}^T \vec{x}_i}_{\text{model fit}} + \underbrace{\varepsilon_i}_{\text{error}}$, $\quad \vec{w} = ?$

$\vec{x}_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$

Squared loss: $\quad \|\varepsilon\|^2 = \sum_{i=1}^{m} \varepsilon_i^2 = \sum_{i=1}^{m} (y_i - w^T x_i)^2$

$$=: \|\vec{y} - X\vec{w}\|_2^2$$

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_m^T \end{pmatrix} \in \mathbb{R}^{m\times n} \qquad \vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^m$$

Empirical Risk Minimization (ERM):

$$\min_{\vec{w}\in\mathbb{R}^n} \|\vec{y} - X\vec{w}\|_2^2 := \ell(\vec{w})$$

$$\frac{\partial \ell}{\partial w} = -2X^T(y - Xw) = 0$$ why? Fermat's!

$$(X^TX)w = X^Ty$$

If (X^TX) invertible, $\quad \hat{w}^{OLS} = (X^TX)^{-1}X^Ty$

Rmk: ① (Ordinary) least-squares solution to linear regression

$\hat{w}^{OLS}$: Best Linear Unbiased Estimate (Gauss–Markov)

② (X^TX) may not be invertible, → pseudo-inverse ...

or invertible but ill-conditioned. → condition number $\kappa = \frac{\lambda_{max}(X^TX)}{\lambda_{min}(X^TX)}$

$\lambda_{min}(X^TX) \approx 0$

$\hat{w}^{OLS}$ is unbiased but can have very large variance $\frac{1}{\lambda_{min}(X^TX)}$

Regularization

to alleviate the problem!

"Ridge"

$$\min_{w} \|y - Xw\|_2^2 \quad \text{s.t.} \quad \|w\|_2^2 \le t$$

$\overset{?}{\Leftrightarrow}$ why?

Guaranteed by strong duality!

$\exists \lambda > 0$ (dual optimal)

$$\min_{w} \underbrace{\|y - Xw\|_2^2 + \lambda \|w\|_2^2}_{\text{Lagrangian}} = \ell(w)$$

$$\frac{\partial \ell}{\partial w} = -2X^T(y - Xw) + 2\lambda w = 0$$

$$\hat{w}^{\text{Ridge}} = \underbrace{(X^TX + \lambda I)^{-1}}\, X^T y$$

Rmk: / Σ_X: $(X^TX + \lambda I)$ is always invertible.

and $\lambda_{\min}(X^TX + \lambda I) \ge \lambda$

Rmk: Constrained opt $\longrightarrow$ unconstrained opt $\to$ Fermat's

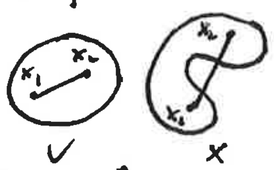
In general, what to do ???

Move to slide presentation / BREAK

Def. (Convex set) $\mathcal{X}$ convex set (in a vector space) iff
$\forall x_1, x_2 \in \mathcal{X}$, $\forall t \in [0,1]$, $\quad t x_1 + (1-t) x_2 \in \mathcal{X}$

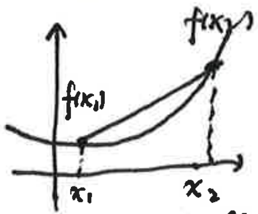
Def. (Convex function) $f: \mathcal{X} \to \mathbb{R}$, $\mathcal{X}$ strictly convex set iff
$\forall x_1, x_2 \in \mathcal{X}$, $\forall t \in [0,1]$, $\quad f(t x_1 + (1-t) x_2) \le t f(x_1) + (1-t) f(x_2)$

Def. (Concave function) g is concave iff $-g$ is convex.

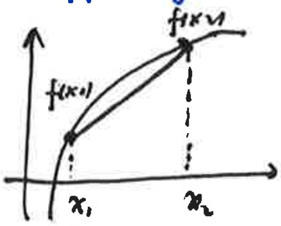


eg: $\mathbb{R}^n$

ball $B = \{x : \|x\|^2 \le t\}$



eg: Convex (e^x)



Concave ($\log x$)