jnp.linal.svd → u, s, VT

собств. вектора/числа

сингулярные вектора/числа

Intuition

$U^TU = I$

$V^TV = I$

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

n

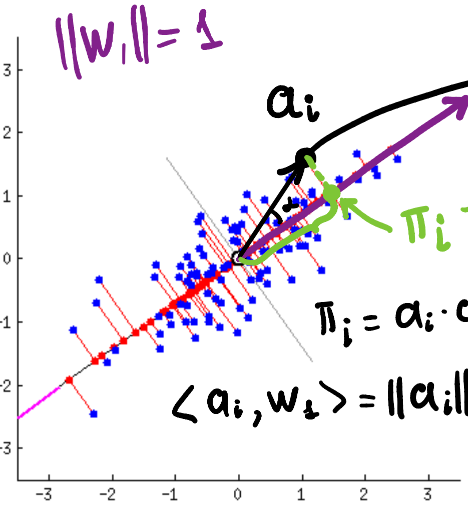
m

Imagine, that you have a dataset of points. Your goal is to choose orthogonal axes, that describe your data the most informative way. To be precise, we choose first axis in such a way, that maximize the variance (expressiveness) of the projected data. All the following axes have to be orthogonal to the previously chosen ones, while satisfy largest possible variance of the projections.

Let's take a look at the simple 2d data. We have a set of blue points on the plane. We can easily see that the projections on the first axis (red dots) have maximum variance at the final position of the animation. The second (and the last) axis should be orthogonal to the previous one.

есть m точек $\in \mathbb{R}^n$

задача: понизить размерность (в некотором смысле оптимально) → выбрать такие оси, проекция на которые MAX дисперсию (вариативность)



$\|w_1\| = 1$

a_i

w_1

π_i - число

n

A

m

данные

1

m

проекции

$\pi_i = a_i \cdot \cos\alpha$

$\langle a_i, w_1\rangle = \|a_i\|\cdot\underbrace{\|w_1\|}_{=1}\cdot\cos\alpha$

$\pi_i = \langle a_i, w_1\rangle$

source

This idea could be used in a variety of ways. For example, it might happen, that projection of complex data on the principal plane (only 2 components) bring you enough intuition for clustering. The picture below plots projection of the labeled dataset onto the first to principal components (PCs), we can clearly see, that only two vectors (these PCs) would be enogh to differ Finnish people from Italian in particular dataset (celiac disease (Dubois et al. 2010)) source

нул вектор

0 Нормировка: Убедиться, что $\boxed{\sum_i a_i = 0}$

если это не так

$A_{std} \to A - \mu$ $\quad \boxed{\mu = \frac{1}{n}\sum_{i=1}^{n} a_i} \in \mathbb{R}^n$

Problem

The first component should be defined in order to maximize variance. Suppose, we've already normalized the data, i.e. $\sum\limits_i a_i = 0$, then sample variance will become the sum т.к. данные нормированы, то $\mathbb{E}\,\Pi_i = 0$

of all squared projections of data points to our vector $\mathbf{w}_{(1)}$, which implies the following optimization problem:

Выб. дисп. $\Pi_i = \mathbb{E}(\Pi_i)^2 - \cancel{(\mathbb{E}\Pi_i)^2} \to 0$

$a_i^\top \cdot w = (Aw)_i$

$\underbrace{m\times n \cdot n\times 1}_{m\times 1}$

R - СВ

$$\boxed{Var(R) = \mathbb{E}R^2 - (\mathbb{E}R)^2}$$

$$\boxed{\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i \left(\underbrace{\mathbf{a}_{(i)}^\top \cdot \mathbf{w}}_{\Pi_i}\right)^2 \right\}}$$

or

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \{\|\mathbf{Aw}\|^2\} = \boxed{\arg\max_{\|\mathbf{w}\|=1} \{\mathbf{w}^\top \mathbf{A}^\top \mathbf{Aw}\}}$$

$\|Aw\|^2 = \langle Aw, Aw \rangle = (Aw)^\top \cdot Aw = w^\top \underbrace{A^\top A}_{n\times m\ m\times n} w$

n×n

X

since we are looking for the unit vector, we can reformulate the problem:

ОТВЕТ:
w_1 = собств. вектор $A^\top A$

$$\boxed{\mathbf{w}_{(1)} = \arg\max \left\{ \frac{\mathbf{w}^\top \mathbf{A}^\top \mathbf{Aw}}{\mathbf{w}^\top \mathbf{w}} \right\}}$$

$$\boxed{\lambda_{min} \le \frac{w^\top X w}{w^\top w} \le \lambda_{max}}$$

It is known, that for positive semidefinite matrix $A^\top A$ such vector is nothing else, but eigenvector of $A^\top A$, which corresponds to the largest eigenvalue. The following components will give you the same results (eigenvectors).

собств. век.

$Xw = \lambda w \mid w^\top$

$w^\top X w = \lambda \cdot w^\top w$

$\lambda = \frac{w^\top X w}{w^\top w}$

So, we can conclude, that the following mapping:

данные

$\lambda(A^\top A) = \sigma^2(A)$

проекции

$$\boxed{\underset{m\times r}{\Pi} = \underset{m\times n}{A} \cdot \underset{n\times r}{W}}$$

PC

describes the projection of data onto the k principal components, where W contains first (by the size of eigenvalues) k eigenvectors of $A^\top A$.

Now we'll briefly derive how SVD decomposition could lead us to the PCA.

Firstly, we write down SVD decomposition of our matrix:

$$\boxed{A = U\Sigma W^\top}$$

$\underset{m\times n}{A}$ = $\underset{m\times m}{U}\ \underset{m\times n}{\Sigma}\ \underset{n\times n}{W^\top}$

данные   фичи

$A^\top A = W\Sigma \underbrace{U^\top \cdot U}_{I} \Sigma W^\top = W \Sigma^2 W^\top$

собств.

собств. вектора

спектральное разложение

and to its transpose:

$$
\begin{aligned}
A^\top &= (U\Sigma W^\top)^\top \\
&= (W^\top)^\top \Sigma^\top U^\top \\
&= W\Sigma^\top U^\top \\
&= W\Sigma U^\top
\end{aligned}
$$

Then, consider matrix $AA^\top$:

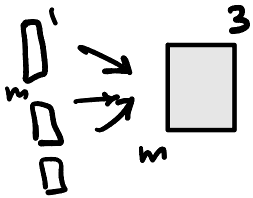
PCA: собств. векторы матрицы $A^\top A$ — оптимальные главные компоненты

w_1, w_2, w_3

$A \cdot w_1$

$A \cdot w_2$

$A \cdot w_3$



$$
\begin{aligned}
A^\top A &= (W\Sigma U^\top)(U\Sigma V^\top) \\
&= W\Sigma I \Sigma W^\top \\
&= W\Sigma\Sigma W^\top \\
&= W\Sigma^2 W^\top
\end{aligned}
$$

Which corresponds to the eigendecomposition of matrix $A^\top A$, where W stands for the matrix of eigenvectors of $A^\top A$, while Σ^2 contains eigenvalues of $A^\top A$.

① $A = U\Sigma W^T$
② проекции
$\Pi_r = U_r \cdot \Sigma_r$
③

At the end:

$U\Sigma W^T$

$$
\begin{aligned}
\underset{m\times n}{\Pi} &= \underset{m\times n}{A} \cdot \underset{n\times n}{W} = \\
&= \underline{U\Sigma W^\top} W = U\Sigma
\end{aligned}
$$

The latter formula provide us with easy way to compute PCA via SVD with any number of principal components:

↓ проекции на первые r PC

$$
\Pi_r = U_r \Sigma_r
$$

Examples

пусть $A \in \mathbb{R}^{10000\times 10}$

хочу взять $r = 2$

PC 1, PC 2

w_1, w_2

W_{2} (n; w_1, w_2)

π_1, π_2 для каждого a_i

🌼 Iris dataset

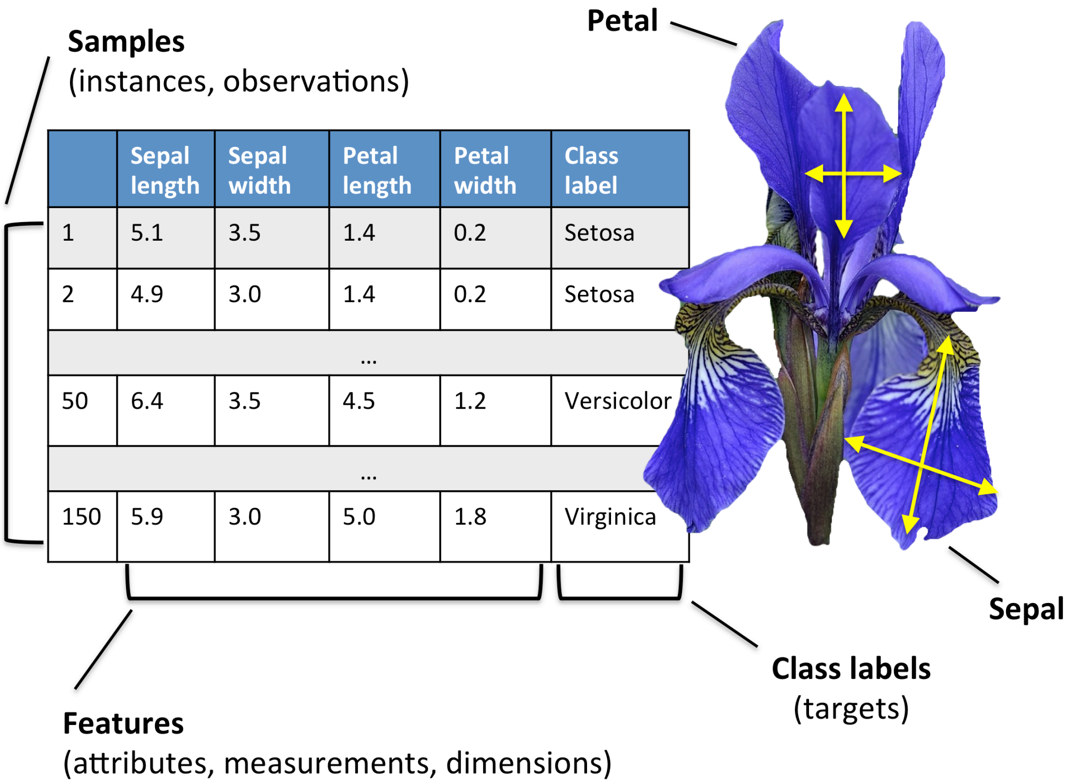
Consider the classical Iris dataset

сжатое представление (π_1, π_2; m; 2)

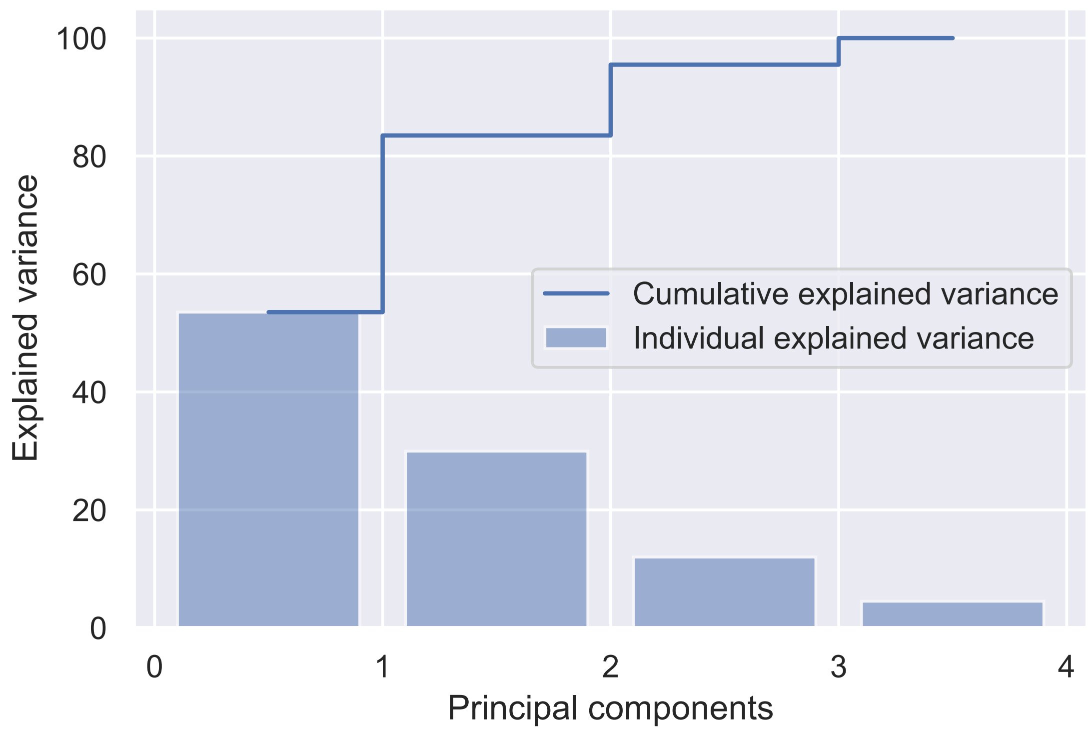
$\tilde{A} = \underset{m\times 2}{\Pi} \cdot \underset{2\times n}{W}$

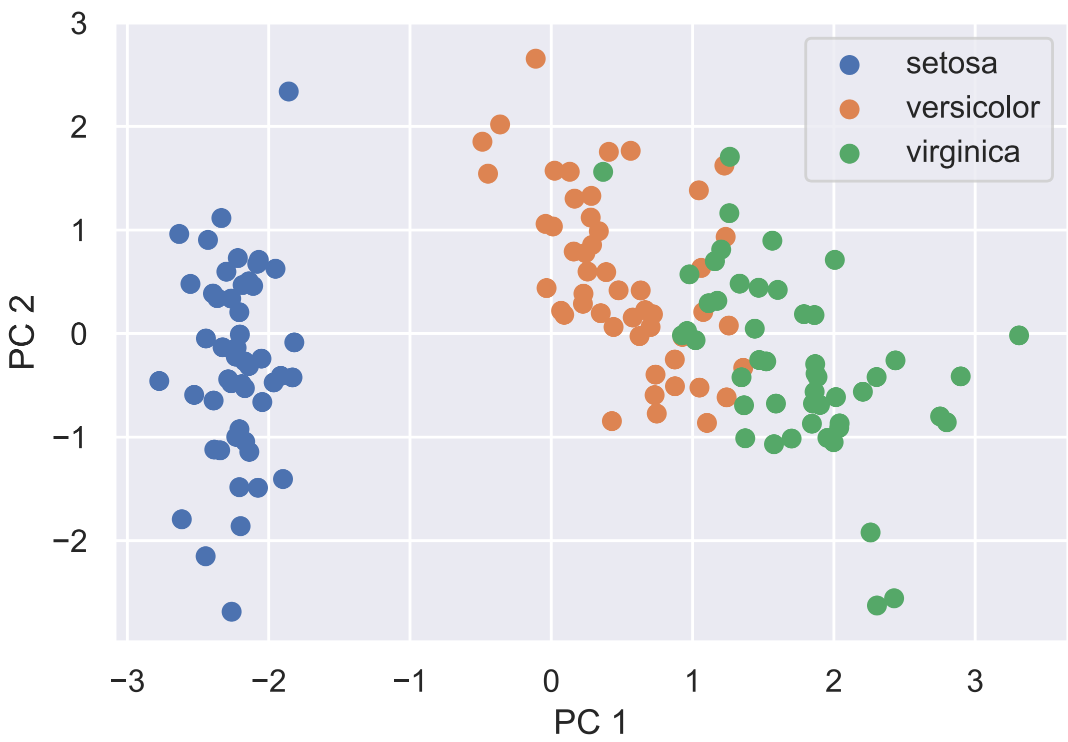
$\tilde{a}_j = \pi_1^j \cdot w_1 + \pi_2^j \cdot w_2$

$\tilde{A} \xrightarrow[r \to \min(m,n)]{} A$



source We have the dataset matrix $A \in \mathbb{R}^{150\times 4}$





Code

Open in Colab

Related materials

- Wikipedia
- Blog post
- Blog post

Useful definitions and notations

We will treat all vectors as column vectors by default. The space of real vectors of length n is denoted by $\mathbb{R}^n$, while the space of real-valued $m \times n$ matrices is denoted by $\mathbb{R}^{m\times n}$.

Basic linear algebra background

The standard **inner product** between vectors x and y from $\mathbb{R}^n$ is given by

$$\langle x, y \rangle = x^\top y = \sum_{i=1}^n x_i y_i = y^\top x = \langle y, x \rangle$$

Here x_i and y_i are the scalar i-th components of corresponding vectors.

The standard **inner product** between matrices X and Y from $\mathbb{R}^{m\times n}$ is given by

$$\langle X, Y \rangle = \text{tr}(X^\top Y) = \sum_{i=1}^m \sum_{j=1}^n X_{ij} Y_{ij} = \text{tr}(Y^\top X) = \langle Y, X \rangle$$

The determinant and trace can be expressed in terms of the eigenvalues

$$\text{det} A = \prod_{i=1}^n \lambda_i, \qquad \text{tr} A = \sum_{i=1}^n \lambda_i$$

Don't forget about the cyclic property of a trace for a square matrices A, B, C, D:

$$\text{tr}(ABCD) = \text{tr}(DABC) = \text{tr}(CDAB) = \text{tr}(BCDA)$$

The largest and smallest eigenvalues satisfy

$$\boxed{\lambda_{\text{min}}(A) = \inf_{x\neq 0} \frac{x^\top A x}{x^\top x}, \qquad \lambda_{\text{max}}(A) = \sup_{x\neq 0} \frac{x^\top A x}{x^\top x}}$$

and consequently $\forall x \in \mathbb{R}^n$ (Rayleigh quotient):

$$\lambda_{\text{min}}(A) x^\top x \leq x^\top A x \leq \lambda_{\text{max}}(A) x^\top x$$

A matrix $A \in \mathbb{S}^n$ (set of square symmetric matrices of dimension n) is called **positive (semi)definite** if for all $x \neq 0 (\text{for all } x) : x^\top A x > (\geq) 0$. We denote this as

$A \succ (\succeq) 0.$

$f(x) = x^T A x = 100x_1^2 + x_2^2$

$A_1 = \begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix}$ $\kappa(A_1) = \frac{100}{1}$

$A_2 = \begin{pmatrix} 1.5 & 0 \\ 0 & 1 \end{pmatrix}$ $\kappa(A_2) = 1.5$

The **condition number** of a nonsingular matrix is defined as

$A_3 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ $\kappa(A_3) = 1$

$$\boxed{\kappa(A) = \|A\|\|A^{-1}\|}$$

$K(A) = \dfrac{\sigma_{MAX}(A)}{\sigma_{MIN}(A)}$

$A \succ 0 \Rightarrow K(A) = \dfrac{\lambda_{MAX}(A)}{\lambda_{MIN}(A)}$

Matrix and vector multiplication

Let A be a matrix of size $m \times n$, and B be a matrix of size $n \times p$, and let the product AB be:

$$C = AB$$

then C is a $m \times p$ matrix, with element (i, j) given by:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

Let A be a matrix of shape $m \times n$, and x be $n \times 1$ vector, then the i-th component of the product:

$$z = Ax$$

is given by:

$$z_i = \sum_{k=1}^{n} a_{ik} x_k$$

Finally, just to remind:

- $C = AB \quad C^\top = B^\top A^\top$
- $AB \neq BA$
- $e^A = \sum\limits_{k=0}^{\infty} \frac{1}{k!} A^k$
- $e^{A+B} \neq e^A e^B$ (but if A and B are commuting matrices, which means that $AB = BA$, $e^{A+B} = e^A e^B$)
- $\langle x, Ay \rangle = \langle A^\top x, y \rangle$

Gradient

Let $f(x) : \mathbb{R}^n \to \mathbb{R}$, then vector, which contains all first order partial derivatives:

$$\nabla f(x) = \frac{df}{dx} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

named gradient of $f(x)$. This vector indicates the direction of steepest ascent. Thus, vector $-\nabla f(x)$ means the direction of the steepest descent of the function in the point. Moreover, the gradient vector is always orthogonal to the contour line in the point.

Hessian

Let $f(x) : \mathbb{R}^n \to \mathbb{R}$, then matrix, containing all the second order partial derivatives:

$$f''(x) = \frac{\partial^2 f}{\partial x_i \partial x_j} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}$$

In fact, Hessian could be a tensor in such a way: $(f(x) : \mathbb{R}^n \to \mathbb{R}^m)$ is just 3d tensor, every slice is just hessian of corresponding scalar function $(H(f_1(x)), H(f_2(x)), \ldots, H(f_m(x)))$.

Jacobian

The extension of the gradient of multidimensional $f(x) : \mathbb{R}^n \to \mathbb{R}^m$ is the following matrix:

$$f'(x) = \frac{df}{dx^T} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

Summary

$$f(x) : X \to Y; \qquad \frac{\partial f(x)}{\partial x} \in G$$

X	Y	G	Name
$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$	$f'(x)$ (derivative)
$\mathbb{R}^n$	$\mathbb{R}$	$\mathbb{R}^n$	$\frac{\partial f}{\partial x_i}$ (gradient)
$\mathbb{R}^n$	$\mathbb{R}^m$	$\mathbb{R}^{m \times n}$	$\frac{\partial f_i}{\partial x_j}$ (jacobian)
$\mathbb{R}^{m \times n}$	$\mathbb{R}$	$\mathbb{R}^{m \times n}$	$\frac{\partial f}{\partial x_{ij}}$

General concept

Naive approach

The basic idea of naive approach is to reduce matrix/vector derivatives to the well-known scalar derivatives.

Matrix notation of a function

$$f(x) = c^\top x$$

Scalar notation of a function

$$f(x) = \sum_{i=1}^n c_i x_i$$

Matrix notation of a gradient

$$\nabla f(x) = c$$

$$\frac{\partial f(x)}{\partial x_k} = c_k$$

Simple derivative

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial \left(\sum_{i=1}^n c_i x_i\right)}{\partial x_k}$$

One of the most important practical tricks here is to separate indices of sum (i) and partial derivatives (k). Ignoring this simple rule tends to produce mistakes.

Differential approach

The guru approach implies formulating a set of simple rules, which allows you to calculate derivatives just like in a scalar case. It might be convenient to use the differential notation here.

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

Differentials

$df = f(x+dx) - f(x)$ $\quad \|dx\| \to 0$

After obtaining the differential notation of df we can retrieve the gradient using following formula:

Задача: посчитать ∇f ?

$\nabla f \xrightarrow{\text{градиент}} \nabla f$

1) Посчитать df

дифференциал (скаляр)

$$\boxed{df(x) = \langle \nabla f(x), dx \rangle}$$

2) Привести к виду

Then, if we have differential of the above form and we need to calculate the second derivative of the matrix/vector function, we treat "old" dx as the constant dx_1, then calculate $d(df) = d^2 f(x)$

$$\boxed{d^2 f(x) = \langle \nabla^2 f(x) dx_1, dx \rangle = \langle H_f(x) dx_1, dx \rangle}$$

$df = \langle \dots, dx \rangle$

3) это ответ

Properties

Let A and B be the constant matrices, while X and Y are the variables (or matrix functions).

- $dA = 0$
- $d(\alpha X) = \alpha (dX)$
- $d(AXB) = A(dX)B$
- $d(X+Y) = dX + dY$
- $d(X^\top) = (dX)^\top$
- $d(XY) = (dX)Y + X(dY)$
- $d\langle X, Y \rangle = \langle dX, Y \rangle + \langle X, dY \rangle$
- $d\left(\dfrac{X}{\phi}\right) = \dfrac{\phi dX - (d\phi)X}{\phi^2}$
- $d\left(\det X\right) = \det X \langle X^{-\top}, dX \rangle$
- $d\left(\text{tr } X\right) = \langle I, dX \rangle$
- $df(g(x)) = \dfrac{df}{dg} \cdot dg(x)$
- $H = (J(\nabla f))^T$

Пример 1 $f(x) = \|x\|^2$ $\quad \nabla f = ?$ $\quad \nabla f \in \mathbb{R}^n$

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

Решение:

1) $df = d(\|x\|^2) = d(\langle x, x\rangle) =$

$$= \langle dx, x\rangle + \langle x, dx\rangle = \langle x, dx\rangle + \langle x, dx\rangle =$$

$$= \langle 2x, \underline{dx}\rangle$$

$$\rightarrow \boxed{\nabla f = 2x} \in \mathbb{R}^n$$

Упр. 1 $f(x) = \|X\|_F^2$, $X \in \mathbb{R}^{m\times n}$ $\quad \nabla f = ?$

$\nabla f \in \mathbb{R}^{m\times n}$

Решение:

1) $df = d(\|X\|_F^2) = d(\langle X, X\rangle) =$

$$= \langle dX, X\rangle + \langle X, dX\rangle = 2\langle X, dX\rangle =$$

$$= \langle \underbrace{2X}_{\nabla f}, dX\rangle \Rightarrow \boxed{\nabla f = 2\cdot X}$$

$\mathbb{R}^{m\times n}$

Пример 2 $f(x) = \frac{1}{2}x^T A x - b^T x + c$ $\quad f: \mathbb{R}^n \to \mathbb{R}$

$\nabla f = ?$ $\quad A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$

Решение:

$x^T y = \langle x, y\rangle$

1) $df = d\left(\frac{1}{2}x^T A x - b^T x + c\right) =$

$$= d\left(\frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle + c\right) =$$

$$= d\left(\frac{1}{2}\langle x, Ax\rangle\right) - d(\langle b, x\rangle) + \cancel{d(c)}^{\,0} =$$

$$= \frac{1}{2}\left(\langle dx, Ax\rangle + \langle x, d(Ax)\rangle\right) - \cancel{\langle db, x\rangle}^{\,0} - \langle dx, b\rangle + 0 =$$

в нужной форме

надо поработать

$$= \frac{1}{2}\left(\langle Ax, dx\rangle + \langle x, Adx\rangle\right) - \langle b, dx\rangle =$$

$$= \frac{1}{2}\left(\langle Ax, dx\rangle + \langle A^T x, dx\rangle\right) - \langle b, dx\rangle =$$

$$= \frac{1}{2}\langle Ax + A^T x, dx\rangle - \langle b, dx\rangle =$$

$$= \langle \underbrace{\tfrac{1}{2}(A+A^T)x - b}_{\nabla f}, dx\rangle$$

$$\Rightarrow \boxed{\underset{n\times 1}{\nabla f} = \frac{1}{2}\underset{n\times n}{(A+A^T)}\underset{n\times 1}{x} - \underset{n\times 1}{b}}$$

если $A \succeq 0 \Rightarrow A = A^T$

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^T \to \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\boxed{\nabla f = Ax - b}$$

$\boxed{\text{Упр 2}}$ задача лин. регрессии с l_2 регуляризацией

$$f(x) = \|Ax - b\|^2 + \frac{\lambda}{2}\|x\|^2 \qquad f: \mathbb{R}^n \to \mathbb{R}$$

$\nabla f = ?$

Решение:

1) $$df = d\left(\|Ax-b\|^2 + \frac{\lambda}{2}\|x\|^2\right) =$$

$$= d\left(\|Ax-b\|^2\right) + d\left(\frac{\lambda}{2}\|x\|^2\right) =$$

$$= d(\langle y, y\rangle) = \qquad y = Ax - b$$

$$= 2\langle y, dy\rangle = \qquad dy = d(Ax-b) = Adx$$

$$\frac{\lambda}{2} d\left(\|x\|^2\right) = \frac{\lambda}{2} d(\langle x, x\rangle) = \frac{\lambda}{2}\cdot 2\langle x, dx\rangle = \langle \lambda x, dx\rangle$$

$$= \langle 2(Ax-b), A\,dx \rangle =$$

$$= \langle 2A^T(Ax-b), dx \rangle$$

$$\boxed{\nabla f = \lambda x + 2A^T(Ax-b)} \in \mathbb{R}^n$$

1×1 $n\times 1$ $n\times m$ $m\times n$ $n\times 1$ $m\times 1$

Упр. $f(X) = tr(X)$ $\quad \nabla f = ? \in \mathbb{R}^{n\times n}$

$$\langle X, Y \rangle = tr(X^T Y) = tr(Y^T X)$$

1) $df = d(tr(X)) =$

$= d(tr(I \cdot X)) =$

$= d(\langle X, \underbrace{I^T}_{I} \rangle) = \langle dX, I \rangle + \langle X, dI \rangle =$

$= \langle I, dX \rangle \Rightarrow \boxed{\nabla f = I}$

Мы научились считать ∇f. Как считать $f'' = ?$

$$\boxed{df = \langle \nabla f, dx \rangle}$$

Как считать f''? $\nabla^2 f$

1) $df = \langle \nabla f, dx \rangle$ $\qquad dx := dx_1$

считаем $dx_1 = const$

2) $d(df) = d^2 f = d(\langle \nabla f, dx_1 \rangle) =$

$= \langle d(\nabla f), dx_1 \rangle$

f''

3) Привести к виду: $d^2 f = \langle (\dots) dx_1, dx \rangle$

Пример: $f(x) = \frac{1}{2}x^T A x - b^T x + c$

1) $$\boxed{df = \langle \tfrac{1}{2}(A+A^T)x - b,\ dx \rangle}$$

$dx := dx_1$

2) Считаем $d^2 f = d\left(\langle \tfrac{1}{2}(A+A^T)x - b, dx_1 \rangle\right) =$

$= \langle d\left(\tfrac{1}{2}(A+A^T)x - b\right), dx_1 \rangle =$ $\quad db = 0$

$= \langle \tfrac{1}{2} d\left((A+A^T)\cdot x\right), dx_1 \rangle =$

$= \langle \tfrac{1}{2}(A+A^T)dx, dx_1 \rangle =$ $\quad \boxed{f'' = \tfrac{1}{2}(A+A^T)}$

МАТР. $\quad dx,\ dx_1$

$(A+A^T)^T = A^T + A = A + A^T$

$= \langle dx, \tfrac{1}{2}(A+A^T)dx_1 \rangle =$

$= \langle \tfrac{1}{2}(A+A^T)dx_1, dx \rangle$

$\boxed{\text{Упр.}}$ $\quad f(x) = \|Ax-b\|^2 + \frac{\lambda}{2}\|x\|^2 \qquad f'' = ?$

Решение: $$\boxed{df = \langle 2A^T(Ax-b) + \lambda x, dx \rangle}$$

меняем dx на dx_1

$d^2 f = \langle \ldots dx, dx_1 \rangle$

матр. f''

1) $d(df) = d^2 f =$

$= \langle d(2A^T(Ax-b) + \lambda x), dx_1 \rangle =$

$=$

$$2A^T \cdot d(Ax - b) =$$
$$= 2A^T \cdot (d(Ax) - 0) =$$
$$= 2A^T \cdot A\,dx$$

$$d(\lambda x) = \lambda \cdot dx$$

$$= \langle 2A^TA\,dx + \lambda dx, dx_1 \rangle =$$
$$= \langle (2A^TA + \lambda \cdot I)\,dx, dx_1 \rangle$$

$$I \cdot v = v$$

$$\lambda dx = \lambda \cdot I \cdot dx$$

$$\cancel{\lambda = \lambda I}$$

$$\boxed{f'' = 2A^TA + \lambda I}$$

$$(A^TA)^T = A^TA^{T^T} = A^TA$$

$$\|dx\| \to 0$$

$$d(dx)$$

$$dx \quad 1e-5$$
$$(dx)^2 = 1e-10$$

$$d(\langle f'(x), dx \rangle) =$$
$$= \langle df', dx \rangle + \langle f', \cancelto{0}{d(dx)} \rangle$$

- $d(X^{-1}) = -X^{-1}(dX)X^{-1}$

References

- Convex Optimization book by S. Boyd and L. Vandenberghe - Appendix A. Mathematical background.
- Numerical Optimization by J. Nocedal and S. J. Wright. - Background Material.
- Matrix decompositions Cheat Sheet.
- Good introduction
- The Matrix Cookbook
- MSU seminars (Rus.)
- Online tool for analytic expression of a derivative.
- Determinant derivative