机器学习常用数学公式汇总（一）

李新春 lxcnju@163.com

2017 年 5 月 7 日

本文主要根据周志华老师的《机器学习》、李航老师的《统计学习方法》、《模式识别》、《矩阵论》、《数理统计》、《优化方法》以及网上博客对机器学习中常用的数学公式进行总结。主要分为矩阵、优化、概率统计三个部分。本次以矩阵为主进行总结其中的公式，然后列出优化里面常用的牛顿法和拟牛顿法公式。概率统计主要涉及常见分布族及其数值特征和各种检验方法等，概率统计将单独总结为一篇文章。

一、矩阵

1、矩阵秩

$$R(\mathbf{A}) = R(\mathbf{A^T}) = R(\mathbf{A^T A}) = R(\mathbf{A A^T}) \tag{0.1}$$

$$max\{R(\mathbf{A}), R(\mathbf{B})\} \leq R(\mathbf{A}, \mathbf{B}) \leq R(\mathbf{A}) + R(\mathbf{B}) \tag{0.2}$$

$$R(\mathbf{A} \pm \mathbf{B}) \leq R(\mathbf{A}) + R(\mathbf{B}) \tag{0.3}$$

$$R(\mathbf{AB}) \leq min\{R(\mathbf{A}), R(\mathbf{B})\} \tag{0.4}$$

2、矩阵迹（**a b x**为列向量）

$$tr(\mathbf{AB}) = tr(\mathbf{BA}) \tag{0.5}$$

$$tr(\mathbf{ABC}) = tr(\mathbf{BCA}) = tr(\mathbf{CAB}) \tag{0.6}$$

$$\frac{\partial tr(\mathbf{AB})}{\partial \mathbf{A}} = \mathbf{B^T},\ \frac{\partial tr(\mathbf{AB})}{\partial \mathbf{B}} = \mathbf{A^T} \tag{0.7}$$

$$\frac{\partial tr(\mathbf{A^T B})}{\partial \mathbf{A}} = \mathbf{B},\ \frac{\partial tr(\mathbf{A^T B})}{\partial \mathbf{B}} = \mathbf{A} \tag{0.8}$$

$$\frac{\partial tr(\mathbf{A})}{\partial \mathbf{A}} = \mathbf{I} \tag{0.9}$$

$$tr(\mathbf{ab^T}) = \mathbf{a^T b} \tag{0.10}$$

$$tr(\mathbf{xx^T A^T}) = \mathbf{x^T A x} \tag{0.11}$$

$$\frac{\partial \mathbf{x^T A x}}{\partial \mathbf{A}} = \mathbf{xx^T} \tag{0.12}$$

$$\frac{\partial \mathbf{x^T A x}}{\partial \mathbf{x}} = (\mathbf{A^T} + \mathbf{A})\mathbf{x} \tag{0.13}$$

$$\frac{\partial tr(\mathbf{ABA^T})}{\partial \mathbf{A}} = \mathbf{A}(\mathbf{B} + \mathbf{B^T}) \tag{0.14}$$

$$\frac{\partial tr(\mathbf{ABA^T C})}{\partial \mathbf{A}} = \mathbf{CAB} + \mathbf{C^T A B^T} \tag{0.15}$$

证明: 这里仅给出最后一个式子的证明：$\mathbf{A} \in \mathbf{R^{m\times n}}, \mathbf{B} \in \mathbf{R^{n\times n}}, \mathbf{C} \in \mathbf{R^{m\times m}}$,记$\mathbf{A}$按列分块结果为$(\mathbf{a_1}, \mathbf{a_2}, \cdots, \mathbf{a_n})$,记$\mathbf{B}_{ij}$为$\mathbf{B}$的第i行第j列元素，其余符号含义类推。

$$
\begin{aligned}
tr(\mathbf{A^T C A B}) &= tr\left(\begin{bmatrix} \mathbf{a_1^T} \\ \mathbf{a_2^T} \\ \vdots \\ \mathbf{a_n^T} \end{bmatrix} \mathbf{C} \begin{bmatrix} \mathbf{a_1} \ \mathbf{a_2} \ \ldots \ \mathbf{a_n} \end{bmatrix} \mathbf{B}\right) \\
&= tr\left(\begin{bmatrix} \mathbf{a_1^T} \\ \mathbf{a_2^T} \\ \vdots \\ \mathbf{a_n^T} \end{bmatrix} \begin{bmatrix} \mathbf{Ca_1} \ \mathbf{Ca_2} \ \ldots \ \mathbf{Ca_n} \end{bmatrix} \mathbf{B}\right) \\
&= tr\left(\begin{bmatrix} \mathbf{a_1^T Ca_1} & \mathbf{a_1^T Ca_2} & \ldots & \mathbf{a_1^T Ca_n} \\ \mathbf{a_2^T Ca_1} & \mathbf{a_2^T Ca_2} & \ldots & \mathbf{a_2^T Ca_n} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{a_n^T Ca_1} & \mathbf{a_n^T Ca_2} & \ldots & \mathbf{a_n^T Ca_n} \end{bmatrix} \mathbf{B}\right) \\
&= \sum_{i,j}^{n} \mathbf{a_i^T C a_j B_{ij}} \\
\frac{\partial \sum_{i,j}^{n} \mathbf{a_i^T C a_j B_{ij}}}{\partial \mathbf{a}_i} &= \sum_{j}^{n} \mathbf{C a_j B_{ij}} + \sum_{j}^{n} \mathbf{C^T a_j B_{ji}} \\
&\triangleq \mathbf{pa_i} \\
\frac{\partial tr(\mathbf{A B A^T C})}{\partial \mathbf{A}} &= \frac{\partial tr(\mathbf{A^T C A B})}{\partial \mathbf{A}} \\
&= \frac{\partial \sum_{i,j}^{n} \mathbf{a_i^T C a_j B_{ij}}}{\partial \mathbf{A}} \\
&= \begin{bmatrix} \mathbf{pa_1} \ \mathbf{pa_2} \ \ldots \ \mathbf{pa_n} \end{bmatrix} \\
&= \mathbf{CAB} + \mathbf{C^T A B^T}
\end{aligned}
$$

3、矩阵导数

$$\left(\frac{\partial \mathbf{a}}{\partial x}\right)_i = \frac{\partial \mathbf{a_i}}{\partial x} \quad (\mathbf{a} \in \mathbf{R}^n, x \in \mathbf{R}) \tag{0.16}$$

$$\left(\frac{\partial \mathbf{A}}{\partial x}\right)_{ij} = \frac{\partial \mathbf{A}_{ij}}{\partial x} \quad (\mathbf{A} \in \mathbf{R}^{m\times n}, x \in \mathbf{R}) \tag{0.17}$$

$$\left(\frac{\partial x}{\partial \mathbf{a}}\right)_i = \frac{\partial x}{\partial \mathbf{a_i}} \quad (\mathbf{a} \in \mathbf{R}^n, x \in \mathbf{R}) \tag{0.18}$$

$$\left(\frac{\partial x}{\partial \mathbf{A}}\right)_{ij} = \frac{\partial x}{\partial \mathbf{A}_{ij}} \quad (\mathbf{A} \in \mathbf{R}^{m\times n}, x \in \mathbf{R}) \tag{0.19}$$

$$\frac{\partial f}{\partial \mathbf{A}} = \begin{bmatrix} \frac{\partial f}{\partial \mathbf{A}_{11}} & \frac{\partial f}{\partial \mathbf{A}_{12}} & \cdots & \frac{\partial f}{\partial \mathbf{A}_{1n}} \\ \frac{\partial f}{\partial \mathbf{A}_{21}} & \frac{\partial f}{\partial \mathbf{A}_{22}} & \cdots & \frac{\partial f}{\partial \mathbf{A}_{2n}} \\ \vdots & \ddots & \ddots & \vdots \\ \frac{\partial f}{\partial \mathbf{A}_{m1}} & \frac{\partial f}{\partial \mathbf{A}_{m2}} & \cdots & \frac{\partial f}{\partial \mathbf{A}_{mn}} \end{bmatrix} \quad (\mathbf{X} \in \mathbf{R}^{m\times n}) \tag{0.20}$$

$$
\begin{aligned}
F(\mathbf{X}) &= \begin{bmatrix} f_{11}(\mathbf{X}) & \dots & f_{1s}(\mathbf{X}) \\ \vdots & \ddots & \vdots \\ f_{r1}(\mathbf{X}) & \dots & f_{rs}(\mathbf{X}) \end{bmatrix} \quad (\mathbf{X} \in \mathbf{R}^{m \times n}) \\
\frac{\partial F(\mathbf{X})}{\partial \mathbf{X}} &= \begin{bmatrix} \frac{\partial F}{\partial \mathbf{X}_{11}} & \cdots & \frac{\partial F}{\partial \mathbf{X}_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial F}{\partial \mathbf{X}_{m1}} & \cdots & \frac{\partial F}{\partial \mathbf{X}_{mn}} \end{bmatrix} \\
\frac{\partial F}{\partial \mathbf{X}_{ij}} &= \begin{bmatrix} \frac{\partial f_{11}}{\partial \mathbf{X}_{ij}} & \cdots & \frac{\partial f_{1n}}{\partial \mathbf{X}_{ij}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_{m1}}{\partial \mathbf{X}_{ij}} & \cdots & \frac{\partial f_{mn}}{\partial \mathbf{X}_{ij}} \end{bmatrix}
\end{aligned} \tag{0.21}
$$

$$
\begin{aligned}
\mathbf{f}(\mathbf{x}) &= \begin{pmatrix} f_1(\mathbf{x}) & \dots & f_m(\mathbf{x}) \end{pmatrix}^{\mathbf{T}} \quad \mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n) \\
J(\mathbf{x}) &= \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial \mathbf{x}_1} & \cdots & \frac{\partial f_1}{\partial \mathbf{x}_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial \mathbf{x}_1} & \dots & \frac{\partial f_m}{\partial \mathbf{x}_n} \end{bmatrix}
\end{aligned} \tag{0.22}
$$

$$
H(\mathbf{x}) = \frac{\partial \bigtriangledown f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial^2 f}{\partial \mathbf{x}_1 \partial \mathbf{x}_1} & \frac{\partial^2 f}{\partial \mathbf{x}_1 \partial \mathbf{x}_2} & \cdots & \frac{\partial^2 f}{\partial \partial \mathbf{x}_1 \mathbf{x}_n} \\ \frac{\partial^2 f}{\partial \mathbf{x}_2 \partial \mathbf{x}_1} & \frac{\partial^2 f}{\partial \mathbf{x}_2 \partial \mathbf{x}_2} & \cdots & \frac{\partial^2 f}{\partial \mathbf{x}_2 \partial \mathbf{x}_n} \\ \vdots & \ddots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial \mathbf{x}_n \partial \mathbf{x}_1} & \frac{\partial^2 f}{\partial \mathbf{x}_n \partial \mathbf{x}_2} & \cdots & \frac{\partial^2 f}{\partial \mathbf{x}_n \partial \mathbf{x}_n} \end{bmatrix} \tag{0.23}
$$

$$
\frac{\partial \mathbf{x}^{\mathbf{T}} \mathbf{a}}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^{\mathbf{T}} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a} \tag{0.24}
$$

$$
\frac{\partial \mathbf{A}\mathbf{B}}{\partial \mathbf{x}} = \frac{\partial \mathbf{A}}{\partial \mathbf{x}} \mathbf{B} + \mathbf{A} \frac{\partial \mathbf{B}}{\partial \mathbf{x}} \tag{0.25}
$$

注：（0.16）是函数向量的导数，（0.17）是函数矩阵的导数，（0.18）是变量对向量的导数，（0.19）是变量对矩阵的导数，（0.20）是函数对矩阵的导数，（0.21）是向量值函数对矩阵的导数，（0.22）是Jocabi矩阵（可由0.21推导出来），（0.23）是Hessian矩阵。

4、矩阵逆、广义逆（用$\mathbf{A}^+$表示）

$$
\frac{\partial \mathbf{A}^{-1}}{\partial \mathbf{x}} = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial \mathbf{x}} \mathbf{A}^{-1} \tag{0.26}
$$

$$
if \quad \mathbf{AGA} = \mathbf{A}, \mathbf{GAG} = \mathbf{G}, (\mathbf{AG})^{\mathbf{T}} = \mathbf{AG}, (\mathbf{GA})^{\mathbf{T}} = \mathbf{GA}, \quad then \quad \mathbf{A}^+ = \mathbf{G} \tag{0.27}
$$

$$
(\mathbf{A}^+)^{\mathbf{T}} = (\mathbf{A}^{\mathbf{T}})^+, \quad (\mathbf{A}\mathbf{A}^{\mathbf{T}})^+ = (\mathbf{A}^{\mathbf{T}})^+ \mathbf{A}^+ \tag{0.28}
$$

$$
if \quad \mathbf{A} = \mathbf{U\Sigma V^T}, \quad then \quad \mathbf{A}^+ = \mathbf{V\Sigma^{-1}U^T} \tag{0.29}
$$

5、矩阵Frobenius范数

$$
||\mathbf{A}||_{\mathbf{F}} = (tr(\mathbf{A}^{\mathbf{T}}\mathbf{A}))^{1/2} \tag{0.30}
$$

$$
\frac{\partial ||\mathbf{A}||_{\mathbf{F}}^2}{\partial \mathbf{A}} = \frac{\partial \mathbf{A}^{\mathbf{T}}\mathbf{A}}{\partial \mathbf{A}} = 2\mathbf{A} \tag{0.31}
$$

注：关于矩阵广义逆和矩阵范数还有很多知识点，以后会有更详细的总结，这里先省去。（矩阵广义逆有很多种，其和最小二乘法以及矩阵投影之间有着密切的联系；矩阵范数更是重要的一点，矩阵的收敛性、条件数、特征值等与其密切相关。）

二、优化

1、牛顿法（0.32是由泰勒展开式得到；假设$\mathbf{G}_k$正定，利用一阶最优性条件得到0.33）

$$f(\mathbf{x}_k+\mathbf{s})=f(\mathbf{x}_k)+\mathbf{s}^{\mathbf{T}}\mathbf{g}_k+\frac{1}{2}\mathbf{s}^{\mathbf{T}}\mathbf{G}_k\mathbf{s} \tag{0.32}$$

$$\mathbf{G}_k\mathbf{s}=-\mathbf{g}_k \quad \Rightarrow \quad \mathbf{s}_k=-\mathbf{G}_k^{-1}\mathbf{g}_k \tag{0.33}$$

2、拟牛顿法（公式0.34是通过梯度函数$\mathbf{g}(\mathbf{x})$在$\mathbf{x}_{k+1}$点的近似得到；式子0.36是通过利用$\mathbf{B}_{k+1}$和$\mathbf{H}_{k+1}$来分别近似$\mathbf{G}_{k+1}$和$\mathbf{G}_{k+1}^{-1}$得到，称之为拟牛顿方式；下面公式是通过各种校正方法进行求解。）

$$g(\mathbf{x}_k)\approx\mathbf{g}_{k+1}+\mathbf{G}_{k+1}(\mathbf{x}_k-\mathbf{x}_{k+1}) \tag{0.34}$$

$$let \ \ \mathbf{y}_k=\mathbf{g}_{k+1}-\mathbf{g}_k,\ \ \mathbf{s}_k=\mathbf{x}_{k+1}-\mathbf{x}_k \ \ then \ \ \mathbf{y}_k\approx\mathbf{G}_{k+1}\mathbf{s}_k \tag{0.35}$$

$$\mathbf{y}_k=\mathbf{B}_{k+1}\mathbf{s}_k,\ \ \mathbf{s}_k=\mathbf{H}_{k+1}\mathbf{y}_k \tag{0.36}$$

$SR1:$

$$\mathbf{H}_{k+1}=\mathbf{H}_k+\mathbf{v}_k\mathbf{v}_k^T \tag{0.37}$$

$$\mathbf{H}_{k+1}=\mathbf{H}_k+\frac{(\mathbf{s}_k-\mathbf{H}_k\mathbf{y}_k)(\mathbf{s}_k-\mathbf{H}_k\mathbf{y}_k)^T}{(\mathbf{s}_k-\mathbf{H}_k\mathbf{y}_k)^T\mathbf{y}_k} \tag{0.38}$$

$$\mathbf{B}_{k+1}=\mathbf{B}_k+\frac{(\mathbf{y}_k-\mathbf{B}_k\mathbf{s}_k)(\mathbf{y}_k-\mathbf{B}_k\mathbf{s}_k)^T}{(\mathbf{y}_k-\mathbf{B}_k\mathbf{s}_k)^T\mathbf{s}_k} \tag{0.39}$$

$DFP:$

$$\mathbf{H}_{k+1}=\mathbf{H}_k+a\mathbf{v}_k\mathbf{v}_k^T+b\mathbf{u}_k\mathbf{u}_k^T \tag{0.40}$$

$$\mathbf{H}_{k+1}=\mathbf{H}_k+\frac{\mathbf{s}_k\mathbf{s}_k^T}{\mathbf{s}_k^T\mathbf{y}_k}-\frac{\mathbf{H}_k\mathbf{y}_k\mathbf{y}_k^T\mathbf{H}_k}{\mathbf{y}_k^T\mathbf{H}_k\mathbf{y}_k} \tag{0.41}$$

$$\mathbf{B}_{k+1}=\left(\mathbf{I}-\frac{\mathbf{y}_k\mathbf{s}_k^T}{\mathbf{y}_k^T\mathbf{s}_k}\right)\mathbf{B}_k\left(\mathbf{I}-\frac{\mathbf{s}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{s}_k}\right)+\frac{\mathbf{y}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{s}_k} \tag{0.42}$$

$BFGS:$

$$\mathbf{B}_{k+1}=\mathbf{B}_k+a\mathbf{v}_k\mathbf{v}_k^T+b\mathbf{u}_k\mathbf{u}_k^T \tag{0.43}$$

$$\mathbf{B}_{k+1}=\mathbf{B}_k+\frac{\mathbf{y}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{s}_k}-\frac{\mathbf{B}_k\mathbf{s}_k\mathbf{s}_k^T\mathbf{B}_k}{\mathbf{s}_k^T\mathbf{B}_k\mathbf{s}_k} \tag{0.44}$$

$$\mathbf{H}_{k+1}=\left(\mathbf{I}-\frac{\mathbf{s}_k\mathbf{y}_k^T}{\mathbf{s}_k^T\mathbf{y}_k}\right)\mathbf{H}_k\left(\mathbf{I}-\frac{\mathbf{y}_k\mathbf{s}_k^T}{\mathbf{s}_k^T\mathbf{y}_k}\right)+\frac{\mathbf{s}_k\mathbf{s}_k^T}{\mathbf{s}_k^T\mathbf{y}_k} \tag{0.45}$$

$Broyden:$

$$\mathbf{H}_{k+1}^{\alpha}=(1-\alpha)\mathbf{H}_k^{DFP}+\alpha\mathbf{H}_k^{BFGS} \tag{0.46}$$

$Morrison:$

$$(\mathbf{A}+\mathbf{u}\mathbf{v}^{\mathbf{T}})^{-1}=\mathbf{A}^{-1}-\frac{\mathbf{A}^{-1}uv^T\mathbf{A}^{-1}}{1+v^T\mathbf{A}^{-1}u} \tag{0.47}$$

注：关于最优化理论，还有很多优化算法。共轭梯度法、约束优化、二次规划、最小二乘、罚函数等可能会在以后专题详细介绍。因为其中涉及东西太多，这里不全面给出。同时，拟牛顿法只是给出了主要公式，其全局收敛性证明、超线性收敛性证明及条件等异常繁杂，请读者参考最优化理论的书籍。

参考资料（仅列出博客）：

1. 矩阵的迹 求导等公式 $http://blog.sina.com.cn/s/blog_74b69f7101016tg5.html$