机器学习笔记

Notes on Machine Learning

J.R.Tsien
jade.ray.tsien@gmail.com

目　　录

第1章　绪论

§ 1.1 梯度下降法（gradient descent）

1.1.1 批处理梯度下降法（batch gradient descent）

1.1.2 随机梯度下降法（stochastic gradient descent）

§ 1.2 矩阵分析（matrix analysis）

1.2.1 矩阵导数(matrix derivatives)

令$f:\mathbb{R}^{m\times n}\mapsto\mathbb{R}$表示将$m\times n$（$m$-by-$n$）矩阵映射为实数的函数。定义$f$对矩阵$\boldsymbol{A}$的导数

$$\nabla_{\boldsymbol{A}}f(\boldsymbol{A})=\begin{pmatrix}\frac{\partial f(\boldsymbol{A})}{\partial a_{11}} & \cdots & \frac{\partial f(\boldsymbol{A})}{\partial a_{1n}}\\ \vdots & \ddots & \vdots\\ \frac{\partial f(\boldsymbol{A})}{\partial a_{n1}} & \cdots & \frac{\partial f(\boldsymbol{A})}{\partial a_{nn}}\end{pmatrix}$$

矩阵的迹（trace）表示的是矩阵的对角元素的和，

$$tr\boldsymbol{A}=\sum_{i=1}^{n}a_{ii}$$

假设A，B，C，D均是方阵

$$trABCD=trDABC=trCDAB=trBCDA \tag{1.1}$$

循环将最右边矩阵放到最左边。假设a是实数

$$trA = trA^T \tag{1.2}$$

$$tr(A+B) = trA + trB \tag{1.3}$$

$$traA = atrA \tag{1.4}$$

下面的一些公式出自Andrew Ng的机器学习讲义，这里证明一下。

$$\nabla_A trAB = B^T \tag{1.5}$$

$$\nabla_{A^T} f(A) = (\nabla_A f(A))^T \tag{1.6}$$

$$\nabla_A trABA^TC = CAB + C^TAB^T \tag{1.7}$$

$$\nabla_A |A| = |A|(A^{-1})^T \tag{1.8}$$

证明 (1) $(\nabla_A trAB)_{ij} = \frac{\partial trAB}{\partial a_{ij}} = \frac{\partial \sum_m \sum_k a_{mk}b_{km}}{\partial a_{ij}}$，只有当$m=i,k=j$时才有$a_{ij}$的系数，所以$(\nabla_A trAB)_{ij} = b_{ji}$，即证。

(2) $(\nabla_{A^T} f(A))_{ij} = \frac{\partial f(A)}{\partial a_{ji}}$，即证。

(3) $trABA^TC = \sum_m \sum_k \sum_t \sum_s a_{mk}b_{kt}a_{st}c_{sm}$，所以

$$(\nabla_A trABA^TC)_{ij} = \frac{\partial \sum_m \sum_k \sum_t \sum_s a_{mk}b_{kt}a_{st}c_{sm}}{\partial a_{ij}}$$

$$= \sum_m \sum_k \sum_t \sum_s \frac{\partial a_{mk}}{\partial a_{ij}} b_{kt}a_{st}c_{sm} + \sum_m \sum_k \sum_t \sum_s a_{mk}b_{kt}\frac{\partial a_{st}}{\partial a_{ij}} c_{sm}$$

左边，令$m=i,k=j$，右边，令$s=i,t=j$，

$$(\nabla_A trABA^TC)_{ij} = \sum_t \sum_s b_{jt}a_{st}c_{si} + \sum_m \sum_k a_{mk}b_{kj}c_{im}$$

$$= \sum_t \sum_s b_{jt}a_{st}c_{si} + \sum_m \sum_k c_{im}a_{mk}b_{kj}$$

$$= (BA^TC)_{ji} + (CAB)_{ij}$$

$$= (C^TAB^T + CAB)_{ij}$$

§ 1.3 非线性规划

§ 1.4 泛函分析

§ 1.5 常用不等式

1.5.1 柯西不等式（Cauchy Inequality）

柯西不等式，又称柯西-施瓦茨不等式（Cauchy-Schwarz inequality）。对于一个内积空间所有向量$\boldsymbol{x}$和$\boldsymbol{y}$，

$$|\langle\boldsymbol{x},\boldsymbol{y}\rangle|^2 \leq \langle\boldsymbol{x},\boldsymbol{x}\rangle\cdot\langle\boldsymbol{y},\boldsymbol{y}\rangle$$

其中$\langle\cdot,\cdot\rangle$表示内积（点积），当且仅当$\boldsymbol{x}$与$\boldsymbol{y}$线性相关时等式成立。

对于欧几里得空间$\mathbb{R}^2$，

$$(\sum_{i=1}^{n} x_i y_i)^2 \leq (\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i^2)$$

当且仅当$\frac{x_1}{y_1}=\frac{x_2}{y_2}=\cdots=\frac{x_n}{y_n}$时等式成立。

1.5.2 赫尔德不等式（Hölder Inequality）

赫尔德不等式揭示了L^p空间的相互关系。设S为测度空间，$1\leq p,q\leq \inf$，且$\frac{1}{p}+\frac{1}{q}=1$，若$f\in L^p(S)$，$g\in L^q(S)$，则$fg\in L^1(S)$，且

$$\| fg \|_1 \leq \| f \|_p \| g \|_q$$

写成序列或向量的形式

$$\sum_{i=1}^{n}|a_i b_i| \leq (\sum_{i=1}^{n}|a_i|^p)^{\frac{1}{p}}(\sum_{i=1}^{n}|b_i|^q)^{\frac{1}{q}}$$

1.5.3 闵可夫斯基不等式（Minkowski Inequality）

闵可夫斯基不等式表明L^p空间是一个赋范向量空间。设S是一个度量空间，$f,g\in L^p(S), 1\leq p\leq \inf$，那么$f+g\in L^p(S)$，有

$$\| f+g \|_p \leq \| f \|_p + \| g \|_p$$

写成序列或向量的形式

$$(\sum_{k=1}^{n}|x_k+y_k|^p)^{\frac{1}{p}} \le (\sum_{k=1}^{n}|x_k|^p)^{\frac{1}{p}}(\sum_{k=1}^{n}|y_k|^p)^{\frac{1}{p}}$$

第2章　感知机

§ 2.1　模型

§ 2.2　策略

§ 2.3　算法

第3章　K近邻

第4章　朴素贝叶斯法

第5章　决策树

第6章　逻辑斯谛回归和最大熵

第7章　支持向量机

第8章　提升方法

第9章　EM方法

§ 9.1 Jensen不等式

第10章　隐马尔可夫模型

参考文献

[1] 李航著．《统计学习方法》．北京:清华大学出版社，2012，3

[2] Jiawei Han, Micheline Kamber, Jian Pei 著.范明, 孟小峰译. 《数据挖掘：概念与技术》. 机械工业出版社, 2012, 8

[3] Tom M. Mitchell 著.曾华军等译. 《机器学习》.机械工业出版社, 2003, 1