附 录

矩阵

矩阵 $\mathbf{A} \in \mathbb{R}^{m\times n}$ 第 i 行第 j 列的元素为 $(\mathbf{A})_{ij} = A_{ij}$.

对于 n 阶方阵 $\mathbf{A}$, 它的**迹**(trace) 是主对角线上的元素之和, 即 $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} A_{ii}$. 迹有如下性质:

$$
\begin{aligned}
\mathrm{tr}(\mathbf{A}^{\mathrm{T}}) &= \mathrm{tr}(\mathbf{A}) \\
\mathrm{tr}(\mathbf{A}+\mathbf{B}) &= \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B}) \\
\mathrm{tr}(\mathbf{AB}) &= \mathrm{tr}(\mathbf{BA}) \\
\mathrm{tr}(\mathbf{ABC}) &= \mathrm{tr}(\mathbf{BCA}) = \mathrm{tr}(\mathbf{CAB}).
\end{aligned}
$$

矩阵 $\mathbf{A} \in \mathbb{R}^{m\times n}$ 的 Frobenius **范数** (norm)定义为

$$
\| \mathbf{A} \|_F = \left(\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{A})\right)^{1/2} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2\right)^{1/2}
$$

容易看出, 矩阵的 Frobenius 范数就是将矩阵张成向量后的 L_2 范数.

导数

向量和矩阵的导数满足乘法法则 (product rule)

$\boldsymbol{a}$ 相对于 $\boldsymbol{x}$ 为常向量

$$\frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{a}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}$$

$$\frac{\partial \mathbf{AB}}{\partial \boldsymbol{x}} = \frac{\partial \mathbf{A}}{\partial \boldsymbol{x}}\mathbf{B} + \mathbf{A}\frac{\partial \mathbf{B}}{\partial \boldsymbol{x}}.$$

类比标量、向量的导数

一些例子:

$$\frac{\partial \parallel \mathbf{A} \parallel_F^2}{\partial \mathbf{A}} = \frac{\partial \mathrm{tr}\left(\mathbf{A}\mathbf{A}^{\mathrm{T}}\right)}{\partial \mathbf{A}} = 2\mathbf{A}.$$

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{x}}(\mathbf{A}\boldsymbol{x}-\boldsymbol{b})^{\mathrm{T}}\mathbf{W}(\mathbf{A}\boldsymbol{x}-\boldsymbol{b}) \quad &= \frac{\partial(\mathbf{A}\boldsymbol{x}-\boldsymbol{b})}{\partial \boldsymbol{x}} \cdot 2\mathbf{W}(\mathbf{A}\boldsymbol{x}-\boldsymbol{b}) \\ &= 2\mathbf{A}^{\mathrm{T}}\mathbf{W}(\mathbf{A}\boldsymbol{x}-\boldsymbol{b}).\end{aligned}$$

奇异值分解

任意实矩阵 $\mathbf{A} \in \mathbb{R}^{m\times n}$ 都可分解为

$$\mathbf{A} = \mathbf{U\Sigma V}^{\mathrm{T}}$$

$\mathbf{U} \in \mathbb{R}^{m\times m}$ 是满足 $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ 的 m 阶酉矩阵 (unitary matrix);

$\mathbf{V} \in \mathbb{R}^{n\times n}$是满足 $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$ 的 n 阶酉矩阵;

$\mathbf{\Sigma} \in \mathbb{R}^{m\times n}$ 是 $m \times n$ 的矩阵，其中 $(\mathbf{\Sigma})_{ii} = \sigma_i$ 且其他位置的元素均为 0,

σ_i 为非负实数且满足 $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant 0$.

当 $\mathbf{A}$ 为对称正定矩阵时，奇异值分解与特征值分解结果相同.

这一分解称为**奇异值分解** (Singular Value Decomposition, 简称 SVD),

其中$\mathbf{U}$的列向量 $\boldsymbol{u}_i \in \mathbb{R}^m$ 称为$\mathbf{A}$的左奇异向量(left-singular vector);

$\mathbf{V}$的列向量 $\boldsymbol{v}_i \in \mathbb{R}^n$称为$\mathbf{A}$的右奇异向量(right-singular vector);

σ_i称为奇异值(singular value);

矩阵 $\mathbf{A}$ 的秩(rank)就等于非零奇异值的个数.

奇异值分解的应用

对于低秩矩阵近似 (low-rank matrix approximation) 问题, 给定一个秩为 r 的矩阵 $\mathbf{A}$, 欲求其最优 k 秩近似矩阵 $\widetilde{\mathbf{A}}$, $k \leqslant r$, 该问题可形式化为

$$\begin{aligned} \min_{\widetilde{\mathbf{A}} \in \mathbb{R}^{m \times n}} \quad & \| \mathbf{A} - \tilde{\mathbf{A}} \|_F \\ \text{s.t.} \quad & \operatorname{rank}(\widetilde{\mathbf{A}}) = k. \end{aligned}$$

奇异值分解提供了上述问题的**解析解**: 对矩阵 $\mathbf{A}$ 进行奇异值分解后, 将矩阵 $\mathbf{\Sigma}$ 中的 $r-k$ 个最小的奇异值置零获得矩阵 $\mathbf{\Sigma}_k$, 即仅保留最大的 k 个奇异值, 则

$$\mathbf{A}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^{\mathrm{T}}$$

其中 $\mathbf{U}_k$ 和 $\mathbf{V}_k$ 分别是$\mathbf{A} = \mathbf{U\Sigma V}^{\mathrm{T}}$中的前 k 列组成的矩阵.

这个结果称为 Eckart-Young-Mirsky 定理.

这一结果和主成分分析（PCA）有密切联系 → 第10章

二次规划

二次规划 (Quadratic Programming, QP) 中，目标函数是变量的**二次**函数, 而约束条件是变量的**线性**不等式.

假定变量个数为 d, 约束条件的个数为 m, 则标准的二次规划问题形如

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\mathbf{Q}\boldsymbol{x} + \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{s.t.} \quad & \mathbf{A}\boldsymbol{x} \leqslant \boldsymbol{b} \end{aligned}$$

其中 $\boldsymbol{x}$ 为 d 维向量, $\mathbf{Q} \in \mathbb{R}^{d\times d}$ 为实对称矩阵, $\mathbf{A} \in \mathbb{R}^{m\times d}$ 为实矩阵, $\boldsymbol{b} \in \mathbb{R}^{m}$ 和 $\boldsymbol{c} \in \mathbb{R}^{d}$ 为实向量, $\mathbf{A}\boldsymbol{x} \leqslant \boldsymbol{b}$ 的每一行对应一个约束.

若 $\mathbf{Q}$ 为**半正定矩阵**, 则目标函数是凸函数, 相应的二次规划是**凸**二次优化问题; 此时若约束条件 $\mathbf{A}\boldsymbol{x} \leqslant \boldsymbol{b}$ 定义的**可行域**不为空, 且目标函数在此可行域有下界, 则该问题将有**全局最小值**.

若$\boldsymbol{Q}$非半正定，则难以有效求解

半正定规划

二次规划 (Quadratic Programming, QP)

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\mathbf{Q}\boldsymbol{x} + \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{s.t.} \quad & \mathbf{A}\boldsymbol{x} \leqslant \boldsymbol{b} \end{aligned}$$

半正定规划(Semi-Definite Programming, 简称SDP)中的变量可组织成**半正定对称矩阵**形式, 且优化问题的目标函数和约束都是这 些变量的线性函数.

$$\begin{aligned} \min_{\mathbf{X}} \quad & \mathbf{C} \cdot \mathbf{X} \\ \text{s.t.} \quad & \mathbf{A}_i \cdot \mathbf{X} = b_i, i = 1,2,\ldots,m \\ & \mathbf{X} \succeq 0. \end{aligned}$$

$\mathbf{X}$、$\mathbf{C}$为$d \times d$ 的对称矩阵,

$$\mathbf{C} \cdot \mathbf{X} = \sum_{i=1}^{d}\sum_{j=1}^{d} C_{ij}X_{ij}$$

若 $\mathbf{A}_i (i = 1,2,\ldots,m)$ 也是 $d \times d$ 的对称矩阵, $b_i (i = 1,2,\ldots,m)$ 为 m 个实数

半正定规划能将线性规划、二次规划等统一起来

梯度下降法

梯度下降法(gradient descent) 求解无约束优化问题:

$$\min_x f(\boldsymbol{x})$$

其中 $f(\boldsymbol{x})$ 为**连续可微**函数. 若能构造一个序列 $\boldsymbol{x}^0, \boldsymbol{x}^1, \boldsymbol{x}^2, \ldots$ 满足

$$f(\boldsymbol{x}^{t+1}) < f(\boldsymbol{x}^t),\ t = 0,1,2,\ldots$$

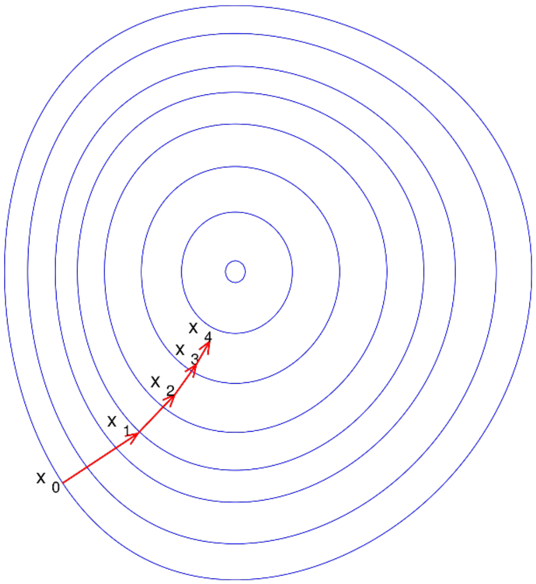
则不断执行该过程即可收敛到**局部极小点**.

根据泰勒展式有

$$f(\boldsymbol{x} + \Delta\boldsymbol{x}) \simeq f(\boldsymbol{x}) + \Delta\boldsymbol{x}^{\mathrm{T}} \nabla f(\boldsymbol{x})$$

于是, 欲满足 $f(\boldsymbol{x} + \Delta\boldsymbol{x}) < f(\boldsymbol{x})$, 可选择

$$\Delta\boldsymbol{x} = -\gamma \nabla f(\boldsymbol{x}),$$



最常见的**无约束**优化方法

拉格朗日乘子法

考虑具有 m 个等式约束和 n 个不等式约束，且可行域 $\mathbb{D} \subset \mathbb{R}^d$ 非空的优化问题

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & h_i(\boldsymbol{x}) = 0 \ (i = 1, \dots, m), \\ & g_j(\boldsymbol{x}) \leqslant 0 \ (j = 1, \dots, n). \end{aligned}$$

拉格朗日乘子法 (Lagrange multipliers) 引入**拉格朗日乘子** $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_m)^{\mathrm{T}}$ 和 $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)^{\mathrm{T}} \geqslant 0$,

拉格朗日函数 $L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ 为

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x}) + \sum_{j=1}^{n} \mu_j g_j(\boldsymbol{x})$$

使用约束增广目标函数（augment the objective function）

构成原始优化问题的下界

拉格朗日乘子法

原问题:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & h_i(\boldsymbol{x}) = 0 \ (i = 1, \dots, m), \\ & g_j(\boldsymbol{x}) \leqslant 0 \ (j = 1, \dots, n). \end{aligned}$$

最优解$\boldsymbol{x}^*$对应的目标函数值为p^*

对偶问题:

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu}} \min_{\boldsymbol{x}} f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x}) + \sum_{j=1}^{n} \mu_j g_j(\boldsymbol{x})$$

$$\text{s.t. } \boldsymbol{\mu} \geq 0$$

最优解$\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*$对应的目标函数值为$d^*$

弱对偶（Weak duality）： $d^* \leq p^*$

强对偶（Strong duality）： $d^* = p^*$

一般的凸问题都具有**强对偶**性质

拉格朗日乘子法

优化问题

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & h_i(\boldsymbol{x}) = 0 \ (i = 1, \dots, m), \\ & g_j(\boldsymbol{x}) \leqslant 0 \ (j = 1, \dots, n). \end{aligned}$$

Karush-Kuhn-Tucker (KKT) 条件 $(j = 1,2, \dots, n)$ 为

$$\begin{cases} h_i(\boldsymbol{x}) = 0 \\ g_j(\boldsymbol{x}) \leqslant 0 \\ \mu_j \geqslant 0 \\ \mu_j g_j(\boldsymbol{x}) = 0 \\ \nabla f(\boldsymbol{x}) + \sum_{i=1}^{m} \mu_j \nabla g_j(\boldsymbol{x}) + \sum_{i=1}^{p} \lambda_i \nabla h_i(\boldsymbol{x}) = 0 \end{cases}$$

$$\begin{aligned} \text{minimize} \quad & (1/2)\boldsymbol{x}^T P \boldsymbol{x} + q^T \boldsymbol{x} + r \\ \text{subject to} \quad & A\boldsymbol{x} = b \end{aligned}$$

假设$P \in \mathbf{S}_+^n$，则KKT条件为

$$A\boldsymbol{x}^\star = b, \ \ P\boldsymbol{x}^\star + q + A^T \boldsymbol{\nu}^\star = 0,$$

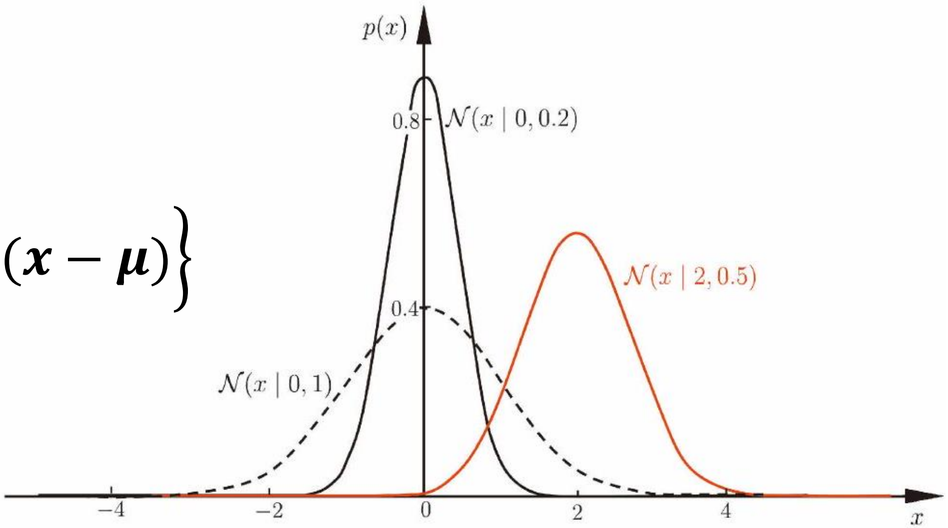
高斯分布

高斯分布(Gaussian distribution)亦称正态分布(normal distribution)，是应用最为广泛的连续概率分布.

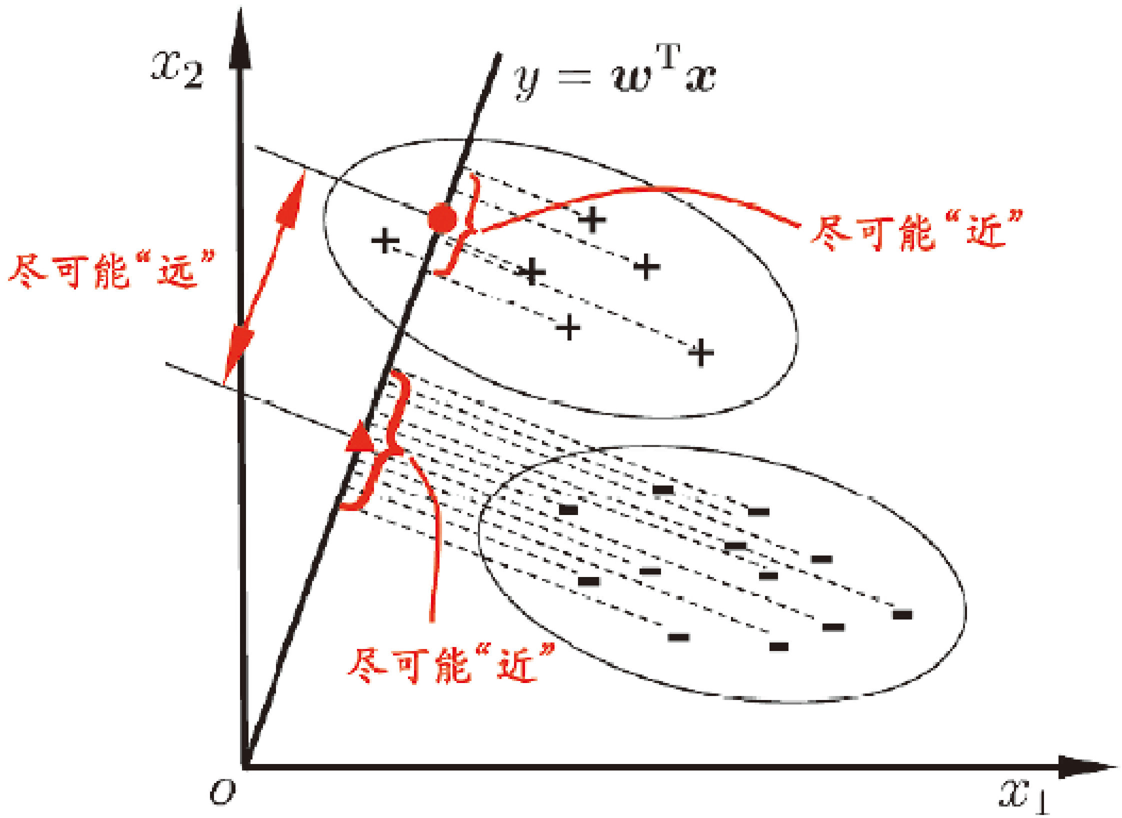
对于单变量 $x \in (-\infty, \infty)$, 高斯分布的参数为均值 $\mu \in (-\infty, \infty)$ 和方差 $\sigma^2 > 0$.

对于 d 维向量 $\boldsymbol{x}$, 多元高斯分布的参数为 d 维均值向量 $\boldsymbol{\mu}$ 和 $d \times d$ 的对称 正定协方差矩阵 $\boldsymbol{\Sigma}$. 表示为$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$



线性判别分析 (Linear Discriminant Analysis)



由于将样例投影到一条直线（低维空间），因此也被视为一种 "监督降维" 技术　降维 ⟶ 第10章

LDA的目标

给定数据集 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$

- 第 i 类示例的集合 X_i
- 第 i 类示例的均值向量 $\boldsymbol{\mu}_i$
- 第 i 类示例的协方差矩阵 $\boldsymbol{\Sigma}_i$
- 两类样本的中心在直线上的投影：$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0$ 和 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1$
- 两类样本的协方差：$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w}$ 和 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}$

同类样例的投影点尽可能接近 → $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w} + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}$ 尽可能小

异类样例的投影点尽可能远离 → $\|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0 - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1\|_2^2$ 尽可能大

于是，最大化

$$J = \frac{\|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0 - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1\|_2^2}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w} + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}} = \frac{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)\boldsymbol{w}}$$

LDA的目标

类内散度矩阵 (within-class scatter matrix)

$$\begin{aligned}\mathbf{S}_w &= \mathbf{\Sigma}_0 + \mathbf{\Sigma}_1 \\ &= \sum_{\boldsymbol{x} \in X_0} (\boldsymbol{x} - \boldsymbol{\mu}_0)(\boldsymbol{x} - \boldsymbol{\mu}_0)^{\mathrm{T}} + \sum_{\boldsymbol{x} \in X_1} (\boldsymbol{x} - \boldsymbol{\mu}_1)(\boldsymbol{x} - \boldsymbol{\mu}_1)^{\mathrm{T}}\end{aligned}$$

类间散度矩阵 (between-class scatter matrix)

$$\mathbf{S}_b = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}}$$

LDA的目标：最大化广义瑞利商 (generalized Rayleigh quotient)

$$\boxed{J = \frac{\boldsymbol{w}^{\mathrm{T}} \mathbf{S}_b \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} \mathbf{S}_w \boldsymbol{w}}}$$

$\boldsymbol{w}$ 成倍缩放不影响 J 值
仅考虑方向

求解思路

令 $\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}=1$，最大化广义瑞利商等价形式为

$$\min_{\boldsymbol{w}} \ -\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w}$$

$$\text{s.t. } \boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}=1$$

运用拉格朗日乘子法，有 $\mathbf{S}_b\boldsymbol{w}=\lambda\mathbf{S}_w\boldsymbol{w}$

由 $\mathbf{S}_b$ 定义，有 $\mathbf{S}_b\boldsymbol{w}=(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}$

注意到 $(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}$ 是标量，令其等于 λ

于是 $\boldsymbol{w}=\mathbf{S}_w^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)$

实践中通常是进行奇异值分解 $\mathbf{S}_w=\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$ → 附录A

然后 $\mathbf{S}_w^{-1}=\mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}}$

推广到多类

假定有 N 个类

- 全局散度矩阵 $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w = \sum_{i=1}^{m} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$
- 类内散度矩阵 $\mathbf{S}_w = \sum_{i=1}^{N} \mathbf{S}_{w_i}$ $\quad \mathbf{S}_{w_i} = \sum_{\boldsymbol{x} \in X_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^T$
- 类间散度矩阵 $\mathbf{S}_b = \mathbf{S}_t - \mathbf{S}_w = \sum_{i=1}^{N} m_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$

多分类LDA有多种实现方法：采用 $\mathbf{S}_b$, $\mathbf{S}_w$, $\mathbf{S}_t$ 中的任何两个

例如，$\max_{\mathbf{W}} \frac{\operatorname{tr}(\mathbf{W}^{\mathrm{T}} \mathbf{S}_b \mathbf{W})}{\operatorname{tr}(\mathbf{W}^{\mathrm{T}} \mathbf{S}_w \mathbf{W})} \Longrightarrow \mathbf{S}_b \mathbf{W} = \lambda \mathbf{S}_w \mathbf{W}$

$\mathbf{W} \in \mathbb{R}^{d \times (N-1)}$

$\mathbf{W}$的闭式解是 $\mathbf{S}_w^{-1}\mathbf{S}_b$的 d' ($\leq N$-1) 个最大非零广义特征值对应的特征向量组成的矩阵

教材附录 A 介绍了基本的矩阵计算. 设向量 $x \sim \mathcal{N}(0, \Sigma)$, 给定半正定矩阵 M, 定义 $\|x\|_M^2 = x^\top M x$, 试证明 $\mathbb{E}\left[\|x\|_M\right] \le \sqrt{\operatorname{tr}(\Sigma M)}$. 可利用如下性质

[Jensen 不等式] 对任意凸函数 $f(x)$, 有

$$f(\mathbb{E}(x)) \le \mathbb{E}(f(x))$$

解:

首先对 $\|x\|_M^2$ 进行变换, 通过矩阵迹的性质, 得到

$$\|x\|_M^2 = x^\top M x = \operatorname{tr}\left(x^\top M x\right) = \operatorname{tr}\left(x x^\top M\right).$$

由于 $f(x) = x^2$ 是凸函数, 因此有

$$\left(\mathbb{E}\left[\|x\|_M\right]\right)^2 \le \mathbb{E}\left[\|x\|_M^2\right]$$

因此有

$$\mathbb{E}\left[\|x\|_M\right] = \sqrt{\left(\mathbb{E}\left[\|x\|_M\right]\right)^2} \le \sqrt{\mathbb{E}\left[\|x\|_M^2\right]} = \sqrt{\mathbb{E}\left[x^\top M x\right]} = \sqrt{\operatorname{tr}\left(\mathbb{E}\left[x x^\top\right] M\right)} = \sqrt{\operatorname{tr}(\Sigma M)}.$$

$\square$

教材 2.2.3 节描述了自助法 (bootstrapping), 下面使用自助法估计统计量, 对自助法做进一步分析. 考虑 m 个从分布 $p(x)$ 中独立同分布抽取的（互不相等的）观测值 $x_1, x_2, \ldots, x_m$, $p(x)$ 的均值为 μ, 方差为 σ^2. 通过 m 个样例, 可使用如下方式估计分布的均值

$$\bar{x}_m = \frac{1}{m} \sum_{i=1}^{m} x_i \ , \tag{3}$$

和方差

$$\bar{\sigma}_m^2 = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x}_m)^2 \ . \tag{4}$$

设 $x_1^*, x_2^*, \ldots, x_m^*$ 为通过自助法采样得到的结果, 且

$$\bar{x}_m^* = \frac{1}{m} \sum_{i=1}^{m} x_i^* \ , \tag{5}$$

1. 请证明 $\mathbb{E}[\bar{x}_m] = \mu$ 且 $\mathbb{E}[\bar{\sigma}_m^2] = \sigma^2$;
2. 计算 $\mathrm{var}[\bar{x}_m]$;
3. 计算 $\mathbb{E}[\bar{x}_m^* \mid x_1, \ldots, x_m]$ 和 $\mathrm{var}[\bar{x}_m^* \mid x_1, \ldots, x_m]$;
4. 计算 $\mathbb{E}[\bar{x}_m^*]$ 和 $\mathrm{var}[\bar{x}_m^*]$; 可利用如下定理

   [全方差公式] 设 X, Y 为同一空间中的随机变量, 则

   $$\mathrm{var}[Y] = \mathbb{E}\left[\mathrm{var}\left[Y \mid X\right]\right] + \mathrm{var}\left[\mathbb{E}\left[Y \mid X\right]\right] \ .$$

1. 通过如下方式证明 $\bar{x}_m$ 和 $\bar{\sigma}_m^2$ 是分布均值和方差的无偏估计:

针对 $\bar{x}_m$ 求期望, 基于期望的线性性质, 得到

$$\mathbb{E}[\bar{x}_m] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} x_i\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[x_i] = \frac{1}{m}\sum_{i=1}^{m}\mu = \mu\ .$$

对于 $\bar{\sigma}_m^2$ 的期望计算, 首先将其展开, 得到

$$\begin{aligned}
\mathbb{E}\left[\bar{\sigma}_m^2\right] &= \mathbb{E}\left[\frac{1}{m-1}\sum_{i=1}^{m}\left(x_i^2+\bar{x}_m^2-2x_i\bar{x}_m\right)\right] \\
&= \frac{1}{m-1}\sum_{i=1}^{m}\mathbb{E}\left[x_i^2\right] + \frac{1}{m-1}\sum_{i=1}^{m}\mathbb{E}\left[\bar{x}_m^2-2x_i\bar{x}_m\right] \\
&= \frac{1}{m-1}\sum_{i=1}^{m}\mathbb{E}\left[x_i^2\right] + \frac{m}{m-1}\mathbb{E}\left[\bar{x}_m^2\right] - \frac{1}{m-1}\sum_{i=1}^{m}\mathbb{E}[2x_i\bar{x}_m]\ .
\end{aligned}$$

其中第一项利用方差的性质 $\mathbb{E}[x_i^2] = \mu^2+\sigma^2$, 得到

$$\frac{1}{m-1}\sum_{i=1}^{m}\mathbb{E}\left[x_i^2\right] = \frac{1}{m-1}\sum_{i=1}^{m}\mathbb{E}\left[x_i^2\right] = \frac{m}{m-1}(\mu^2+\sigma^2)\ .$$

2. 利用方差的性质将 $\mathrm{var}[\bar{x}_m]$ 展开, 并利用随机变量的独立性, 有

$$\begin{aligned}
\mathrm{var}[\bar{x}_m] &= \mathrm{var}\left[\frac{1}{m}\sum_{i=1}^{m} x_i\right] \\
&= \frac{1}{m^2}\,\mathrm{var}\left[\sum_{i=1}^{m} x_i\right] \\
&= \frac{1}{m^2}\cdot m\cdot \mathrm{var}[x] \\
&= \frac{1}{m}\sigma^2 \; .
\end{aligned}$$

3. 计算条件期望时, 将 $\{x_1, ..., x_m\}$ 视为定值. 根据 x_i^* 服从分布 $P(x_i^* = x_i) = \frac{1}{m}$, 得到

$$\mathbb{E}[x_i^* \mid x_1, ..., x_m] = \frac{1}{m}\sum_{i=1}^{m} x_i = \bar{x}_m \ .$$

于是

$$\begin{aligned}\mathbb{E}[\bar{x}_m^* \mid x_1, ..., x_m] &= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} x_i^* \mid x_1, ..., x_m\right] \\ &= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[x_i^* \mid x_1, ..., x_m] \\ &= \bar{x}_m \ .\end{aligned}$$

为了计算 $\operatorname{var}[\bar{x}_m^* \mid x_1, ..., x_m]$, 首先有

$$\operatorname{var}[x_i^* \mid x_1, ..., x_m] = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x}_m)^2 = \frac{m-1}{m}\bar{\sigma}_m^2 \ .$$

于是

$$\operatorname{var}[\bar{x}_m^* \mid x_1, ..., x_m] = \frac{\operatorname{var}[x_i^* \mid x_1, ..., x_m]}{m} = \frac{m-1}{m^2}\bar{\sigma}_m^2 \ .$$

4. 根据全期望公式 $\mathbb{E}Y = \mathbb{E}[\mathbb{E}[Y|X]]$, 我们有,

$$\begin{aligned}
\mathbb{E}[\bar{x}_m^*] &= \mathbb{E}[\mathbb{E}[\bar{x}_m^* \mid x_1, ..., x_m]] \\
&= \mathbb{E}[\bar{x}_m] \\
&= \mu \, .
\end{aligned}$$

根据全方差公式, 有,

$$\begin{aligned}
\mathrm{var}[\bar{x}_m^*] &= \mathbb{E}[\mathrm{var}[\bar{x}_m^* \mid x_1, ..., x_m]] + \mathrm{var}[\mathbb{E}[\bar{x}_m^* \mid x_1, ..., x_m]] \\
&= \mathbb{E}\left[\frac{m-1}{m^2}\bar{\sigma}_m^2\right] + \mathrm{var}[\bar{x}_m] \\
&= \frac{m-1}{m^2}\sigma^2 + \frac{\sigma^2}{m} \\
&= \frac{\sigma^2}{m}\left(2 - \frac{1}{m}\right) .
\end{aligned}$$

前往……