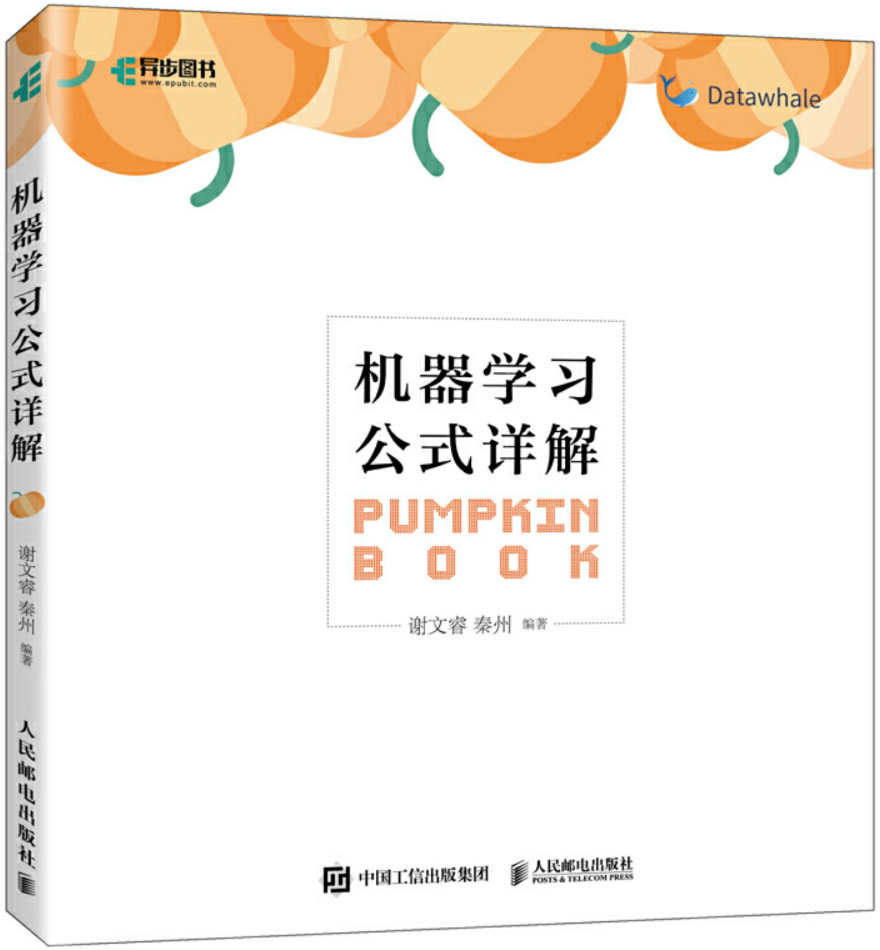


# 《机器学习公式详解》（南瓜书）

## 第3章 二分类线性判别分析

本节主讲：谢文睿

# 本节大纲

Datawhale 异步社区

西瓜书对应章节：3.4

1. 算法原理（模型）
2. 损失函数推导（策略）
3. 拉格朗日乘子法
4. 求解$\boldsymbol{w}$（算法）
5. 广义特征值和广义瑞利商

# 算法原理



从几何的角度，让全体训练样本经过投影后：

- 异类样本的中心尽可能远
- 同类样本的方差尽可能小

# 损失函数推导



经过投影后，异类样本的中心尽可能远（并非严格投影）：

$$\max \left\|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0 - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1\right\|_2^2$$

$$\max \left\||\boldsymbol{w}| \cdot |\boldsymbol{\mu}_0| \cdot \cos\theta_0 - |\boldsymbol{w}| \cdot |\boldsymbol{\mu}_1| \cdot \cos\theta_1\right\|_2^2$$

经过投影后，同类样本的方差尽可能小（并非严格方差）：

$$\min \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w}$$

$$\begin{aligned}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w} &= \boldsymbol{w}^{\mathrm{T}}\left(\sum_{\boldsymbol{x}\in X_0}(\boldsymbol{x}-\boldsymbol{\mu}_0)(\boldsymbol{x}-\boldsymbol{\mu}_0)^{\mathrm{T}}\right)\boldsymbol{w} \\ &= \sum_{\boldsymbol{x}\in X_0}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0)(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}-\boldsymbol{\mu}_0^{\mathrm{T}}\boldsymbol{w})\end{aligned}$$

# 损失函数推导



$$
\begin{aligned}
\max J &= \frac{\left\|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0 - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1\right\|_2^2}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w} + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}} \\
&= \frac{\left\|\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0 - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1\right)^{\mathrm{T}}\right\|_2^2}{\boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1\right)\boldsymbol{w}} \\
&= \frac{\left\|\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right)^{\mathrm{T}}\boldsymbol{w}\right\|_2^2}{\boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1\right)\boldsymbol{w}} \\
&= \frac{\left[\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right)^{\mathrm{T}}\boldsymbol{w}\right]^{\mathrm{T}}\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right)^{\mathrm{T}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1\right)\boldsymbol{w}} \\
&= \frac{\boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right)\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right)^{\mathrm{T}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1\right)\boldsymbol{w}}
\end{aligned}
$$

# 损失函数推导

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

$$\Downarrow$$

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

$$\Downarrow$$

$$\begin{aligned}\min_{\boldsymbol{w}} \quad & -\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w}\\ \text{s.t.} \quad & \boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}=1\end{aligned}$$

# 拉格朗日乘子法



对于仅含等式约束的优化问题（解释下为啥是min）：

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & h_i(\boldsymbol{x}) = 0 \quad i = 1, 2, ..., n \end{aligned}$$

其中自变量$\boldsymbol{x} \in \mathbb{R}^n$，$f(\boldsymbol{x})$和$h_i(\boldsymbol{x})$均有连续的一阶偏导数。首先列出其拉格朗日函数：

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

其中$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_n)^{\mathrm{T}}$为拉格朗日乘子。然后对拉格朗日函数关于$\boldsymbol{x}$求偏导，并令导数等于**0**再搭配约束条件$h_i(\boldsymbol{x}) = 0$解出$\boldsymbol{x}$，求解出的所有$\boldsymbol{x}$即为上述优化问题的所有可能【极值点】

## 求解$\boldsymbol{w}$

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & -\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w} \\ \text{s.t.} \quad & \boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w} = 1 \Leftrightarrow \boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w} - 1 = 0 \end{aligned}$$

由拉格朗日乘子法可得拉格朗日函数为

$$L(\boldsymbol{w}, \lambda) = -\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w} + \lambda(\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w} - 1)$$

对$\boldsymbol{w}$求偏导可得

$$\begin{aligned} \frac{\partial L(\boldsymbol{w}, \lambda)}{\partial \boldsymbol{w}} &= -\frac{\partial(\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w})}{\partial \boldsymbol{w}} + \lambda\frac{\partial(\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w} - 1)}{\partial \boldsymbol{w}} \\ &= -(\mathbf{S}_b + \mathbf{S}_b^{\mathrm{T}})\boldsymbol{w} + \lambda(\mathbf{S}_w + \mathbf{S}_w^{\mathrm{T}})\boldsymbol{w} \end{aligned}$$

由于$\mathbf{S}_b = \mathbf{S}_b^{\mathrm{T}}, \mathbf{S}_w = \mathbf{S}_w^{\mathrm{T}}$，所以

# 求解$\boldsymbol{w}$



$$\frac{\partial L(\boldsymbol{w},\lambda)}{\partial \boldsymbol{w}}=-2\mathbf{S}_b\boldsymbol{w}+2\lambda\mathbf{S}_w\boldsymbol{w}$$

令上式等于0即可得

$$-2\mathbf{S}_b\boldsymbol{w}+2\lambda\mathbf{S}_w\boldsymbol{w}=0$$

$$\mathbf{S}_b\boldsymbol{w}=\lambda\mathbf{S}_w\boldsymbol{w}$$

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

若令$(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}=\gamma$，则

# 求解$\boldsymbol{w}$



$$\gamma(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)=\lambda\mathbf{S}_w\boldsymbol{w}$$

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

由于最终要求解的$\boldsymbol{w}$不关心其大小，只关心其方向，所以$\frac{\gamma}{\lambda}$这个常数项可以任意取值，西瓜书中所说的“不妨令$\mathbf{S}_b\boldsymbol{w}=\lambda(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)$”就等价于令$\gamma=\lambda$，进而使得$\frac{\gamma}{\lambda}=1$，此时求解出的$\boldsymbol{w}$即为公式(3.39)，另外，此处并未严格按照上述拉格朗日乘子法再刻意考虑等式约束条件$\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}-1=0$也是因为不在乎$\boldsymbol{w}$的大小。

# 求解$\boldsymbol{w}$



此时用拉格朗日乘子法求出来的极值点$\boldsymbol{w}$一定是最小值点吗?

答：是的，因为$-\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w}=-\|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1\|_2^2\leqslant 0$，所以目标函数最大值为0，且一定存在最小值，所以求出来的极值点只要代入目标函数不为0则一定是最小值点。

# 广义特征值



设$\mathbf{A},\mathbf{B}$为n阶方阵，若存在数λ，使得方程$\mathbf{A}\boldsymbol{x}=\lambda\mathbf{B}\boldsymbol{x}$存在非零解，则称λ为$\mathbf{A}$相对于$\mathbf{B}$的广义特征值，$\boldsymbol{x}$为$\mathbf{A}$相对于$\mathbf{B}$的属于广义特征值λ的特征向量。特别地，当$\mathbf{B}=\mathbf{I}$（单位矩阵）时，广义特征值问题退化为标准特征值问题。

# 广义瑞利商



设$\mathbf{A},\mathbf{B}$为n阶厄米（Hermitian）矩阵，且$\mathbf{B}$正定，称$R(\boldsymbol{x})=\dfrac{\boldsymbol{x}^{\mathrm{H}}\mathbf{A}\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{H}}\mathbf{B}\boldsymbol{x}}(\boldsymbol{x}\neq\mathbf{0})$为$\mathbf{A}$相对于$\mathbf{B}$的广义瑞利商。特别地，当$\mathbf{B}=\mathbf{I}$（单位矩阵）时，广义瑞利商退化为瑞利商。

# 广义瑞利商



广义瑞利商的性质：设$\lambda_i, \boldsymbol{x}_i (i=1,2,...,n)$为$\mathbf{A}$相对于$\mathbf{B}$的广义特征值和特征向量，且$\lambda_1 \leqslant \lambda_2 \leqslant ... \leqslant \lambda_n$。

$$\min_{\boldsymbol{x} \neq \mathbf{0}} R(\boldsymbol{x}) = \frac{\boldsymbol{x}^{\mathrm{H}} \mathbf{A} \boldsymbol{x}}{\boldsymbol{x}^{\mathrm{H}} \mathbf{B} \boldsymbol{x}} = \lambda_1, \boldsymbol{x}^* = \boldsymbol{x}_1$$

$$\max_{\boldsymbol{x} \neq \mathbf{0}} R(\boldsymbol{x}) = \frac{\boldsymbol{x}^{\mathrm{H}} \mathbf{A} \boldsymbol{x}}{\boldsymbol{x}^{\mathrm{H}} \mathbf{B} \boldsymbol{x}} = \lambda_n, \boldsymbol{x}^* = \boldsymbol{x}_n$$

【证明】：当固定$\boldsymbol{x}^{\mathrm{H}} \mathbf{B} \boldsymbol{x} = 1$时，使用拉格朗日乘子法可推得$\mathbf{A}\boldsymbol{x} = \lambda \mathbf{B} \boldsymbol{x}$这样一个广义特征值问题，因此$\boldsymbol{x}$所有可能的解即为$\boldsymbol{x}_i (i=1,2,...,n)$这n个广义特征向量，将其分别代入$R(\boldsymbol{x})$即可推得上述结论。

# 预告

下一节：决策树

西瓜书对应章节：4.1、4.2

# 结束语

欢迎加入【南瓜书读者交流群】，我们将在群里进行答疑、勘误、本次直播回放、本次直播PPT发放、下次直播通知等最新资源发放和活动通知。

加入步骤：

1. 关注公众号【Datawhale】，发送【南瓜书】三个字获取机器人“小豚”的微信二维码
2. 添加“小豚”为微信好友，然后对“小豚”发送【南瓜书】三个字即可自动邀请进群、