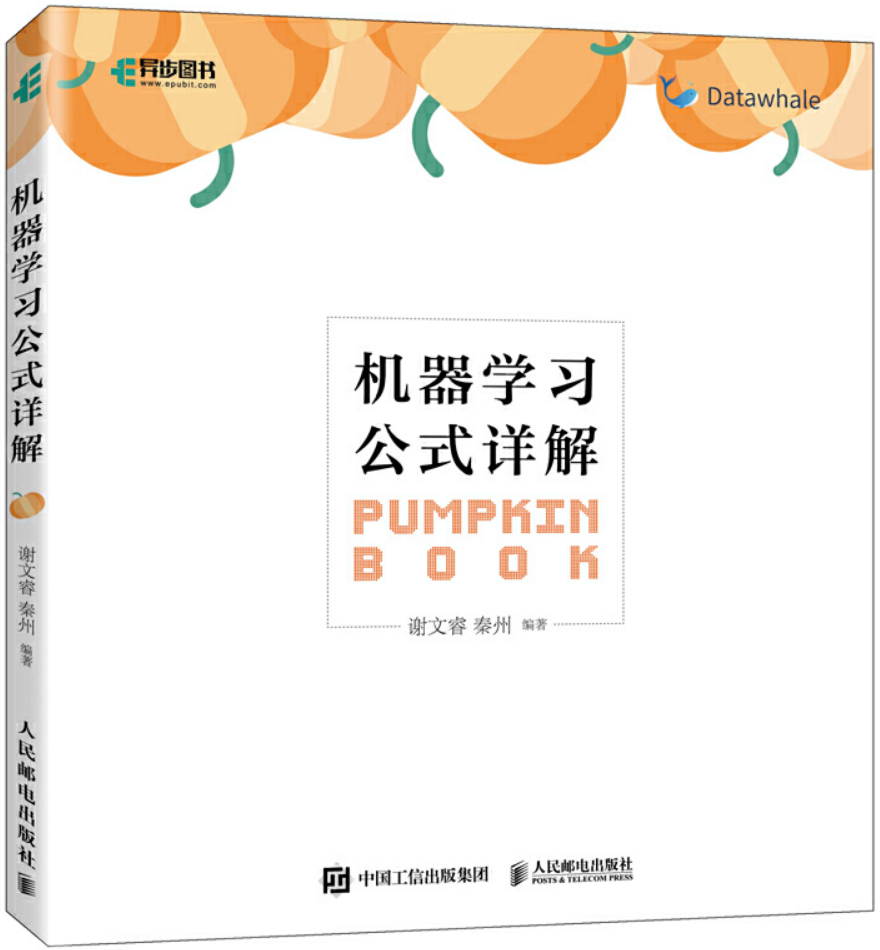

《机器学习公式详解》（南瓜书）

第6章 支持向量机

本节主讲：谢文睿

本节大纲

西瓜书对应章节：6.1、6.2

1. 算法原理
2. 超平面
3. 几何间隔
4. 支持向量机

算法原理

从几何角度，对于线性可分数据集，支持向量机就是找距离正负样本都最远的超平面，相比于感知机，其解是唯一的，且不偏不倚，泛化性能更好。

超平面

n维空间的超平面（$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=0$，其中$\boldsymbol{w},\boldsymbol{x}\in\mathbb{R}^n$）：

- 超平面方程不唯一
- 法向量$\boldsymbol{w}$和位移项b确定一个唯一超平面
- 法向量$\boldsymbol{w}$垂直于超平面（缩放$\boldsymbol{w},b$时，若缩放倍数为负数会改变法向量方向）
- 法向量$\boldsymbol{w}$指向的那一半空间为正空间，另一半为负空间
- 任意点$\boldsymbol{x}$到超平面的距离公式为

$$r=\frac{\left|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b\right|}{\|\boldsymbol{w}\|}$$

超平面

【证明】：对于任意一点$\boldsymbol{x}_0 = (x_1^0, x_2^0, ..., x_n^0)^{\mathrm{T}}$，设其在超平面$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0$上的投影点为$\boldsymbol{x}_1 = (x_1^1, x_2^1, ..., x_n^1)^{\mathrm{T}}$，则$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_1 + b = 0$，且向量$\overrightarrow{\boldsymbol{x}_1\boldsymbol{x}_0}$与法向量$\boldsymbol{w}$平行，因此

$$|\boldsymbol{w} \cdot \overrightarrow{\boldsymbol{x}_1\boldsymbol{x}_0}| = |\|\boldsymbol{w}\| \cdot \cos\pi \cdot \|\overrightarrow{\boldsymbol{x}_1\boldsymbol{x}_0}\|| = \|\boldsymbol{w}\| \cdot \|\overrightarrow{\boldsymbol{x}_1\boldsymbol{x}_0}\| = \|\boldsymbol{w}\| \cdot r$$

$$\begin{aligned}
\boldsymbol{w} \cdot \overrightarrow{\boldsymbol{x}_1\boldsymbol{x}_0} &= w_1(x_1^0 - x_1^1) + w_2(x_2^0 - x_2^1) + ... + w_n(x_n^0 - x_n^1) \\
&= w_1x_1^0 + w_2x_2^0 + ... + w_nx_n^0 - (w_1x_1^1 + w_2x_2^1 + ... + w_nx_n^1) \\
&= \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_0 - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_1 \\
&= \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_0 + b
\end{aligned}$$

$$|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_0 + b| = \|\boldsymbol{w}\| \cdot r \Rightarrow r = \frac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b|}{\|\boldsymbol{w}\|}$$

几何间隔

对于给定的数据集X和超平面$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=0$，定义数据集$X$中的任意一个样本点$(\boldsymbol{x}_i,y_i),y_i\in\{-1,1\},i=1,2,...,m$关于超平面的几何间隔为

$$\gamma_i=\frac{y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)}{\|\boldsymbol{w}\|}$$

正确分类时：$\gamma_i>0$，几何间隔此时也等价于点到超平面的距离
没有正确分类时：$\gamma_i<0$

对于给定的数据集X和超平面$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=0$，定义数据集$X$关于超平面的几何间隔为：数据集$X$中所有样本点的几何间隔最小值

$$\gamma=\min_{i=1,2,...,m}\gamma_i$$

支持向量机

模型：给定线性可分数据集X，支持向量机模型希望求得数据集X关于超平面的几何间隔γ达到最大的那个超平面，然后套上一个sign函数实现分类功能

$$y = \operatorname{sign}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b) = \begin{cases} 1, & \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b > 0 \\ -1, & \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b < 0 \end{cases}$$

所以其本质和感知机一样，仍然是在求一个超平面。那么几何间隔最大的超平面就一定是我们前面所说的那个“距离正负样本都最远的超平面”吗?

答：是的，原因有以下两点：

- 当超平面没有正确划分正负样本时：几何间隔最小的为误分类点，因此$\gamma < 0$
- 当超平面正确划分超平面时：$\gamma \geqslant 0$，且越靠近中央γ越大

支持向量机

策略：给定线性可分数据集X，设X中几何间隔最小的样本为$(\boldsymbol{x}_{min}, y_{min})$，那么支持向量机找超平面的过程可以转化为以下带约束条件的优化问题

$$\begin{aligned} \max \quad & \gamma \\ \text{s.t.} \quad & \gamma_i \geqslant \gamma, \quad i = 1, 2, \ldots, m \end{aligned}$$

$$\begin{aligned} \max_{\boldsymbol{w}, b} \quad & \frac{y_{min}(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_{min} + b)}{\|\boldsymbol{w}\|} \\ \text{s.t.} \quad & \frac{y_i(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b)}{\|\boldsymbol{w}\|} \geqslant \frac{y_{min}(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_{min} + b)}{\|\boldsymbol{w}\|}, \quad i = 1, 2, \ldots, m \end{aligned}$$

$$\begin{aligned} \max_{\boldsymbol{w}, b} \quad & \frac{y_{min}(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_{min} + b)}{\|\boldsymbol{w}\|} \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b) \geqslant y_{min}(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_{min} + b), \quad i = 1, 2, \ldots, m \end{aligned}$$

支持向量机

$$\begin{aligned} \max_{\boldsymbol{w},b} \quad & \frac{y_{min}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_{min}+b)}{\|\boldsymbol{w}\|} \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b) \geqslant y_{min}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_{min}+b), \quad i=1,2,\ldots,m \end{aligned}$$

假设该问题的最优解为$(\boldsymbol{w}^*,b^*)$，那么$(\alpha\boldsymbol{w}^*,\alpha b^*),\alpha\in\mathbb{R}^+$也是最优解，且超平面也不变，因此还需要对$\boldsymbol{w},b$做一定限制才能使得上述优化问题有可解的唯一解。不妨令$y_{min}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_{min}+b)=1$，因为对于特定的$(\boldsymbol{x}_{min},y_{min})$来说，能使得$y_{min}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_{min}+b)=1$的$\alpha$有且仅有一个。因此上述优化问题进一步转化为

$$\begin{aligned} \max_{\boldsymbol{w},b} \quad & \frac{1}{\|\boldsymbol{w}\|} \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b) \geqslant 1, \quad i=1,2,\ldots,m \end{aligned}$$

为了方便后续计算，再进一步进行恒等变换

支持向量机

$$
\begin{aligned}
\min_{\boldsymbol{w},b} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 \\
\text{s.t.} \quad & 1-y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\leqslant 0, \quad i=1,2,\ldots,m
\end{aligned}
$$

此优化问题为含不等式约束的优化问题，且为凸优化问题，因此可以直接用很多专门求解凸优化问题的方法求解该问题，在这里，支持向量机通常采用拉格朗日对偶来求解，具体原因待求解完后解释，下面先给出拉格朗日对偶相关知识。

推荐阅读：王书宁 译.《凸优化》、王燕军.《最优化基础理论与方法（第二版）》

支持向量机

对于一般地约束优化问题：

$$\begin{aligned} \min \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & g_i(\boldsymbol{x}) \leqslant 0 \quad i=1,2,...,m \\ & h_j(\boldsymbol{x}) = 0 \quad j=1,2,...,n \end{aligned}$$

若目标函数$f(\boldsymbol{x})$是凸函数，约束集合是凸集，则称上述优化问题为凸优化问题，特别地，$g_i(\boldsymbol{x})$是凸函数，$h_j(\boldsymbol{x})$是线性函数时，约束集合为凸集，该优化问题为凸优化问题。显然，支持向量机的目标函数$\frac{1}{2}\|\boldsymbol{w}\|^2$是关于$\boldsymbol{w}$的凸函数，不等式约束$1-y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)$是也是关于$\boldsymbol{w}$的凸函数，因此支持向量机是一个凸优化问题。

支持向量机

对于一般地约束优化问题（不一定是凸优化问题）：

$$\begin{aligned} \min \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & g_i(\boldsymbol{x}) \leqslant 0 \quad i=1,2,...,m \\ & h_j(\boldsymbol{x}) = 0 \quad j=1,2,...,n \end{aligned}$$

设上述优化问题的定义域为$D = \boldsymbol{dom}\ f \cap \bigcap\limits_{i=1}^{m} \boldsymbol{dom}\ g_i \cap \bigcap\limits_{j=1}^{n} \boldsymbol{dom}\ h_j$，可行集为$\tilde{D} = \{\boldsymbol{x} | \boldsymbol{x} \in D, g_i(\boldsymbol{x}) \leqslant 0, h_j(\boldsymbol{x}) = 0\}$，显然$\tilde{D}$是$D$的子集，最优值为$p^* = \min\{f(\tilde{\boldsymbol{x}})\}$。由拉格朗日函数的定义可知上述优化问题的拉格朗日函数为

$$L(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \mu_i g_i(\boldsymbol{x}) + \sum_{j=1}^{n} \lambda_j h_j(\boldsymbol{x})$$

其中$\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_m)^T, \boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_n)^T$为拉格朗日乘子向量。

定义上述优化问题的拉格朗日对偶函数$\Gamma(\boldsymbol{\mu}, \boldsymbol{\lambda})$（注意其自变量不包含$\boldsymbol{x}$）为$L(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda})$关于$\boldsymbol{x}$的下确界，也即

$$\Gamma(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \inf_{\boldsymbol{x} \in D} L(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \inf_{\boldsymbol{x} \in D} \left(f(\boldsymbol{x}) + \sum_{i=1}^{m} \mu_i g_i(\boldsymbol{x}) + \sum_{j=1}^{n} \lambda_j h_j(\boldsymbol{x}) \right)$$

对偶函数$\Gamma(\boldsymbol{\mu}, \boldsymbol{\lambda})$有如下重要性质：

- 无论上述优化问题是否是凸优化问题，其对偶函数$\Gamma(\boldsymbol{\mu}, \boldsymbol{\lambda})$恒为**凹函数**（证明参见《凸优化》§ 3.2.3）；
- 当$\boldsymbol{\mu} \succeq 0$时，$\Gamma(\boldsymbol{\mu}, \boldsymbol{\lambda})$构成了上述优化问题最优值$p^*$的下界，也即

$$\Gamma(\boldsymbol{\mu}, \boldsymbol{\lambda}) \leqslant p^*$$

【证明】：设$\tilde{\boldsymbol{x}} \in \tilde{D}$是优化问题的可行点，那么当$\boldsymbol{\mu} \succeq 0$时

$$\sum_{i=1}^{n} \mu_i g_i(\tilde{\boldsymbol{x}}) + \sum_{j=1}^{m} \lambda_j h_j(\tilde{\boldsymbol{x}}) \leqslant 0$$

这是因为左边第一项非正而第二项恒为0。根据此不等式可以进一步推得

$$\Gamma(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \inf_{\boldsymbol{x} \in D} L(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \leqslant L(\tilde{\boldsymbol{x}}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \leqslant f(\tilde{\boldsymbol{x}})$$

$$\Gamma(\boldsymbol{\mu}, \boldsymbol{\lambda}) \leqslant \min\{f(\tilde{\boldsymbol{x}})\} = p^*$$

所以，当$\boldsymbol{\mu} \succeq 0$时，$\Gamma(\boldsymbol{\mu}, \boldsymbol{\lambda}) \leqslant p^*$恒成立，证毕。

支持向量机

定义在满足$\boldsymbol{\mu} \succeq 0$这个约束条件下求对偶函数最大值的优化问题为拉格朗日对偶问题（原优化问题称为主问题）

$$\begin{aligned} \max \quad & \Gamma(\boldsymbol{\mu}, \boldsymbol{\lambda}) \\ \text{s.t.} \quad & \boldsymbol{\mu} \succeq 0 \end{aligned}$$

设该优化问题的最优值为d^*，显然$d^* \leqslant p^*$，此时称为“弱对偶性”成立，若$d^* = p^*$，则称为“强对偶性”成立。曲线救国，找到了求p^*的方法~

- 当主问题满足某些**充分条件**时，强对偶性成立。常见的充分条件有Slater条件：“若主问题是凸优化问题，且可行集$\tilde{D}$中存在一点能使得**所有**不等式约束的不等号成立，则强对偶性成立”（证明参见《凸优化》§5.3.2）。显然，支持向量机满足Slater条件。
- 无论主问题是否为凸优化问题，对偶问题恒为**凸优化问题**，因为对偶函数$\Gamma(\boldsymbol{\mu}, \boldsymbol{\lambda})$恒为**凹函数**（加个负号即可转为凸函数），约束条件$\boldsymbol{\mu} \succeq 0$恒为凸集。

支持向量机

- 设$f(\boldsymbol{x}), g_i(\boldsymbol{x}), h_j(\boldsymbol{x})$一阶偏导连续，$\boldsymbol{x}^*, (\boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$分别为主问题和对偶问题的最优解，若强对偶性成立，则$\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*$一定满足如下5个条件（证明参见《凸优化》）：

$$
\begin{cases}
\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) = \nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(\boldsymbol{x}^*) + \sum_{j=1}^{n} \lambda_j^* \nabla h_j(\boldsymbol{x}^*) = 0 & (1) \\
h_j(\boldsymbol{x}^*) = 0 & (2) \\
g_i(\boldsymbol{x}^*) \leqslant 0 & (3) \\
\mu_i^* \geqslant 0 & (4) \\
\mu_i^* g_i(\boldsymbol{x}^*) = 0 & (5)
\end{cases}
$$

以上5个条件也称为KKT条件，原始定义参见《机器学习公式详解》（南瓜书）第6章-附注，值得一提的是，Slater条件也是第6章-附注中所说的“约束限制条件”。

支持向量机

主问题：

$$\begin{aligned} \min_{\boldsymbol{w},b} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{s.t.} \quad & 1-y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\leqslant 0, \quad i=1,2,\ldots,m \end{aligned}$$

拉格朗日函数：

$$\begin{aligned} L(\boldsymbol{w},b,\boldsymbol{\alpha}) &= \frac{1}{2}||\boldsymbol{w}||^2+\sum_{i=1}^{m}\alpha_i(1-y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)) \\ &= \frac{1}{2}||\boldsymbol{w}||^2+\sum_{i=1}^{m}\alpha_i-\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i-b\sum_{i=1}^{m}\alpha_i y_i \end{aligned}$$

若将$\boldsymbol{w},b$合并为$\hat{\boldsymbol{w}}=(\boldsymbol{w};b)$，显然上式是关于$\hat{\boldsymbol{w}}$的凸函数，直接求一阶导令其等于0，然后带回即可得到最小值，也即拉格朗日对偶函数，下面再给出另一种推导方法。

支持向量机

由于$L(\boldsymbol{w}, b, \boldsymbol{\alpha})$是关于$\boldsymbol{w}$的凸函数，关于b的线性函数，所以当$b$的系数不为0时下确界为$-\infty$，当$b$的系数为0时，下确界就由其他部分来确定，所以$L(\boldsymbol{w}, b, \boldsymbol{\alpha})$的下确界（对偶函数）为

$$\Gamma(\boldsymbol{\alpha}) = \inf_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \begin{cases} \inf\limits_{\boldsymbol{w}} \left\{ \frac{1}{2}||\boldsymbol{w}||^2 + \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i \right\}, & \text{if } \sum_{i=1}^{m} \alpha_i y_i = 0 \\ -\infty, & \text{otherwise} \end{cases}$$

$$\Gamma(\boldsymbol{\alpha}) = \inf_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \begin{cases} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j, & \text{if } \sum_{i=1}^{m} \alpha_i y_i = 0 \\ -\infty, & \text{otherwise} \end{cases}$$

对偶问题：

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \Gamma(\boldsymbol{\alpha}) \\ \text{s.t.} \quad & \boldsymbol{\alpha} \succeq 0 \end{aligned} \quad \Rightarrow \quad \begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j \\ \text{s.t.} \quad & \boldsymbol{\alpha} \succeq 0 \\ & \sum_{i=1}^{m} \alpha_i y_i = 0 \end{aligned}$$

支持向量机

再根据强对偶性成立推得最优解必须满足如下KKT条件

$$\begin{cases} \alpha_i \geqslant 0 \\ y_i f\left(\boldsymbol{x}_i\right)-1 \geqslant 0 \\ \alpha_i\left(y_i f\left(\boldsymbol{x}_i\right)-1\right)=0 \end{cases}$$

为什么支持向量机通常都采用拉格朗日对偶求解呢?

1. 无论主问题是何种优化问题，对偶问题恒为凸优化问题，因此更容易求解（尽管支持向量机的主问题本就是凸优化问题），而且原始问题的时间复杂度和特征维数呈正比（因为未知量是$\boldsymbol{w}$），而对偶问题和数据量成正比（因为未知量是$\boldsymbol{\alpha}$），当特征维数远高于数据量的时候拉格朗日对偶更高效；
2. 对偶问题能很自然地引入核函数，进而推广到非线性分类问题（最主要的原因）

预告

下一节：软间隔与支持向量回归

西瓜书对应章节：6.4、6.5

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、