第 4 章 超越梯度下降

第 3 章介绍了梯度下降法。梯度下降法基于函数局部一阶特性。一阶近似是粗糙的，这种粗糙带来了一些问题。本章将介绍函数在局部的二阶特性。基于二阶特性分析函数在局部的性质。

本章首先回顾一些矩阵的相关知识，之后介绍如何在局部对函数进行二阶近似。有了函数的二阶近似就可以确定驻点的类型：极小点、极大点或者鞍点。之后本章介绍对原始梯度下降法的一些改进，这些改进有助于提高收敛速度，防止震荡或发散，规避局部极小。

最后，本章介绍两个基于函数二阶特性的优化算法：牛顿法和共轭方向法。然后介绍用牛顿法训练逻辑回归模型。二阶算法虽然不常用在神经网络和深度学习的训练中。阅读完本章，读者应该对函数的局部形态有更深刻的理解。

4.1　矩阵

首先回顾一下矩阵。这不是一个关于矩阵的全面介绍，例如行列式这个概念就没有出现。本节只介绍一下后文讨论中用得上的相关知识。

4.1.1　矩阵基础

矩阵是实数构成的 2 维阵列。以一个3×3矩阵为例：

$$\boldsymbol{A}=\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}=\begin{pmatrix} \boldsymbol{a}_{*1} & \boldsymbol{a}_{*2} & \boldsymbol{a}_{*3} \end{pmatrix}=\begin{pmatrix} \boldsymbol{a}_{1*}^{\mathrm{T}} \\ \boldsymbol{a}_{2*}^{\mathrm{T}} \\ \boldsymbol{a}_{3*}^{\mathrm{T}} \end{pmatrix} \qquad (4.1)$$

式（4.1）囊括了本书用到的对矩阵的各种表示。本书用大写粗斜体字母表示矩阵，例如$\boldsymbol{A}$。a_{ij}是实数，是矩阵$\boldsymbol{A}$的第i行、第j列元素。$\boldsymbol{a}_{*j}$是矩阵的第j列，它是一个列向量：

$$\boldsymbol{a}_{*j}=\begin{pmatrix} a_{1j} \\ a_{2j} \\ a_{3j} \end{pmatrix} \qquad (4.2)$$

$\boldsymbol{a}_{i*}$是矩阵的第i行，它是一个列向量：

$$\boldsymbol{a}_{i*}=\begin{pmatrix}a_{i1}\\a_{i2}\\a_{i3}\end{pmatrix} \qquad （4.3）$$

式（4.1）中对$\boldsymbol{a}_{i*}$进行了转置，以表示一行。矩阵的行数和列数不一定相等，可以是$m\times n$，$m\neq n$。表示成：

$$\boldsymbol{A}_{m\times n}=\begin{pmatrix}a_{11}&\cdots&a_{1n}\\\vdots&\ddots&\vdots\\a_{m1}&\cdots&a_{mn}\end{pmatrix} \qquad （4.4）$$

一般可省略下标$m\times n$。两个相同形状的矩阵可以相加：

$$\boldsymbol{A}+\boldsymbol{B}=\begin{pmatrix}a_{11}+b_{11}&\cdots&a_{1n}+b_{1n}\\\vdots&\ddots&\vdots\\a_{m1}+b_{m1}&\cdots&a_{mn}+b_{mn}\end{pmatrix} \qquad （4.5）$$

矩阵相加就是把相应元素相加。可以用实数（标量）乘一个矩阵：

$$k\boldsymbol{A}=\begin{pmatrix}ka_{11}&\cdots&ka_{1n}\\\vdots&\ddots&\vdots\\ka_{m1}&\cdots&ka_{mn}\end{pmatrix} \qquad （4.6）$$

$-\boldsymbol{A}$就是$(-1)\boldsymbol{A}$。显然有$\boldsymbol{A}-\boldsymbol{A}=\boldsymbol{A}+(-\boldsymbol{A})=\boldsymbol{O}$。$\boldsymbol{O}$是所有元素都为 0 的矩阵——零矩阵。矩阵$\boldsymbol{A}$的转置定义为：

$$\boldsymbol{A}^{\mathrm{T}}=\begin{pmatrix}a_{11}&\cdots&a_{m1}\\\vdots&\ddots&\vdots\\a_{1n}&\cdots&a_{mn}\end{pmatrix}=(\boldsymbol{a}_{1*}\quad\boldsymbol{a}_{2*}\quad\boldsymbol{a}_{3*})=\begin{pmatrix}\boldsymbol{a}_{*1}^{\mathrm{T}}\\\boldsymbol{a}_{*2}^{\mathrm{T}}\\\boldsymbol{a}_{*3}^{\mathrm{T}}\end{pmatrix} \qquad （4.7）$$

$\boldsymbol{A}^{\mathrm{T}}$把$\boldsymbol{A}$的行当做列，列当做行。如果$\boldsymbol{A}$是$m\times n$的，那么$\boldsymbol{A}^{\mathrm{T}}$就是$n\times m$的。

如果矩阵$\boldsymbol{A}$是$m\times n$的，它可以与一个n维向量$\boldsymbol{x}$相乘：

$$\boldsymbol{Ax} = \sum_{j=1}^{n} x_j \, \boldsymbol{a}_{*j} \qquad (4.8)$$

矩阵$\boldsymbol{A}$乘向量$\boldsymbol{x}$，使用$\boldsymbol{x}$的元素对矩阵的列进行线性组合。所以$\boldsymbol{A}$的列数和$\boldsymbol{x}$的维数必须相同。得到的结果是一个m维向量。容易看出$\boldsymbol{Ax}$的第i个元素是$\sum_{j=1}^{n} x_j a_{ij} = \boldsymbol{a}_{i*}^{\mathrm{T}}\boldsymbol{x}$，即$\boldsymbol{A}$的第$i$行与$\boldsymbol{x}$的内积。

有了矩阵和向量相乘的定义，就可以定义矩阵与矩阵相乘：

$$\boldsymbol{AB} = (\boldsymbol{Ab}_{*1} \quad \boldsymbol{Ab}_{*2} \quad \cdots \quad \boldsymbol{Ab}_{*k}) \qquad (4.9)$$

$\boldsymbol{A}$与$\boldsymbol{B}$的乘积是矩阵$\boldsymbol{AB}$。$\boldsymbol{AB}$的第j列是$\boldsymbol{A}$与$\boldsymbol{B}$的第j列$\boldsymbol{b}_{*j}$的乘积。如果$\boldsymbol{A}$是$m \times n$的，那么$\boldsymbol{b}_{*j}$必须是n维向量，即$\boldsymbol{B}$必须为n行。$\boldsymbol{B}$的列数任意，例如k。所以要能够与$m \times n$的$\boldsymbol{A}$相乘，$\boldsymbol{B}$的形状必须是$n \times k$，k任意。结果$\boldsymbol{AB}$的形状是$m \times k$。$\boldsymbol{AB}$的第i行、第j列元素是：

$$\boldsymbol{a}_{i*}^{\mathrm{T}}\boldsymbol{b}_{*j} = \sum_{\boldsymbol{s=1}}^{\boldsymbol{n}} a_{is} b_{sj} \qquad (4.10)$$

仅从形状上看$\boldsymbol{B}$与$\boldsymbol{A}$不一定能够相乘，因为k不一定等于m。就算$k = m$，$\boldsymbol{BA}$也不一定等于$\boldsymbol{AB}$。即矩阵乘法不满足交换律。一个反例就可以证明这一点。这里不再赘述。

矩阵的乘法满足结合率：

$$(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC}) \qquad (4.11)$$

矩阵乘法对加法满足结合律：

$$\boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{AB}+\boldsymbol{AC}, \quad (\boldsymbol{A}+\boldsymbol{B})\boldsymbol{C} = \boldsymbol{AC}+\boldsymbol{BC} \qquad (4.12)$$

矩阵乘法对数乘有：

$$\boldsymbol{A}(k\boldsymbol{B}) = (k\boldsymbol{A})\boldsymbol{B} = k(\boldsymbol{AB}) \qquad (4.13)$$

矩阵的数乘满足分配率：

$$k(\boldsymbol{A}+\boldsymbol{B})=k\boldsymbol{A}+k\boldsymbol{B},\quad (k+h)\boldsymbol{A}=k\boldsymbol{A}+h\boldsymbol{A} \tag{4.14}$$

矩阵乘积的转置是：

$$(\boldsymbol{AB})^T=\boldsymbol{B}^T\boldsymbol{A}^T \tag{4.15}$$

上述几个结论的证明很简单，只需要检查一下矩阵元素的表达式。向量$\boldsymbol{x}$的转置$\boldsymbol{x}^T$可以乘一个矩阵$\boldsymbol{A}$：

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}=(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x})^{\mathrm{T}} \tag{4.16}$$

行数和列数相同的矩阵是方阵。方阵$\boldsymbol{A}$的对角线元素之和称为它的迹（trace）：

$$tr(\boldsymbol{A})=\sum_{i=1}^{n}a_{ii} \tag{4.17}$$

方阵$\boldsymbol{A}$和$\boldsymbol{B}$的乘积$\boldsymbol{AB}$的迹等于$\boldsymbol{BA}$的迹：

$$tr(\boldsymbol{AB})=\sum_{i=1}^{n}\boldsymbol{a}_{i*}\boldsymbol{b}_{*i}=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}b_{ji}=\sum_{j=1}^{n}\sum_{i=1}^{n}b_{ji}a_{ij}=\sum_{j=1}^{n}\boldsymbol{b}_{j*}\boldsymbol{a}_{*j}=tr(\boldsymbol{BA}) \tag{4.18}$$

如果一个$n\times n$的方阵的对角线元素为 1，其余元素都是 0，那么它是单位阵：

$$\boldsymbol{I}_{n\times n}=\begin{pmatrix}1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 1\end{pmatrix} \tag{4.19}$$

容易验证对于任何矩阵$\boldsymbol{A}_{m\times n}$，$\boldsymbol{I}_{m\times m}\boldsymbol{A}_{m\times n}=\boldsymbol{A}_{m\times n}\boldsymbol{I}_{n\times n}=\boldsymbol{A}_{m\times n}$。在上下文很清晰时一般省略$\boldsymbol{I}$的下标。

4.1.2　矩阵的逆

令$\boldsymbol{A}$是$n\times n$方阵，如果存在$n\times n$方阵$\boldsymbol{A}^{-1}$满足：

$$AA^{-1} = A^{-1}A = I \qquad (4.20)$$

则称$\boldsymbol{A}$是可逆的。$\boldsymbol{A}^{-1}$是$\boldsymbol{A}$的逆矩阵。$\boldsymbol{A}$的逆矩阵是唯一的。因为假如任何一个矩阵$\boldsymbol{B}$是$\boldsymbol{A}$的逆矩阵，根据定义有：

$$\boldsymbol{B} = \boldsymbol{IB} = \boldsymbol{A}^{-1}\boldsymbol{AB} = \boldsymbol{A}^{-1} \qquad (4.21)$$

如果$\boldsymbol{A}$可逆则$\boldsymbol{A}$的列线性独立。因为假如$\boldsymbol{a}_{*j=1\dots n}$线性相关，则存在一组不全为 0 的系数$w_1, w_2, \cdots, w_n$，使得$\sum_{i=1}^{n} w_i \boldsymbol{a}_{*i} = \boldsymbol{0}$。即存在向量$\boldsymbol{w} = (w_1 \quad \cdots \quad w_n)^{\mathrm{T}} \neq \boldsymbol{0}$使：

$$\boldsymbol{Aw} = \boldsymbol{0} \qquad (4.22)$$

因为$\boldsymbol{A}$可逆，存在$\boldsymbol{A}^{-1}$：

$$\boldsymbol{w} = \boldsymbol{A}^{-1}\boldsymbol{Aw} = \boldsymbol{0} \qquad (4.23)$$

这与$\boldsymbol{w} \neq \boldsymbol{0}$矛盾。所以可逆矩阵$\boldsymbol{A}$的列$\boldsymbol{a}_{*j=1\dots n}$一定线性独立。如果方阵$\boldsymbol{A}$的逆矩阵是$\boldsymbol{A}^{\mathrm{T}}$，则称$\boldsymbol{A}$为正交矩阵：

$$\boldsymbol{AA}^{\mathrm{T}} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \boldsymbol{I} \qquad (4.24)$$

从（4.24）可以看出$\boldsymbol{A}$的列$\boldsymbol{a}_{*j=1\dots n}$是单位向量且两两正交：

$$\begin{cases} \boldsymbol{a}_{*i}{}^{T}\boldsymbol{a}_{*j} = 0,\ i \neq j \\ \boldsymbol{a}_{*i}{}^{T}\boldsymbol{a}_{*j} = 1,\ i = j \end{cases} \qquad (4.25)$$

也就是说，正交矩阵$\boldsymbol{A}$的列都是单位向量，$\left\|\boldsymbol{a}_{*j=1\dots n}\right\| = 1$,。任意两列是正交的（夹角为$\pi/2$）。因为$\boldsymbol{A}$可逆，所以$\boldsymbol{a}_{*j=1\dots n}$线性独立，是$n$维线性空间$\mathbb{R}^n$的一组基。因为$\boldsymbol{a}_{*j=1\dots n}$两两正交，还是单位向量，所以它们被称为$\mathbb{R}^n$的一组标准正交基。

如果一组向量$\boldsymbol{x}_{i=1\dots n}$是线性独立的，可以通过施密特正交化过程构造一组正交的向量$\boldsymbol{x'}_{i=1\dots n}$。构造过程是：

首先令$\boldsymbol{x'}_1 = \boldsymbol{x}_1$。然后令：

$$\boldsymbol{x'}_2 = \boldsymbol{x}_2 - \frac{{\boldsymbol{x}_2}^{\mathrm{T}}\boldsymbol{x'}_1}{{\boldsymbol{x'}_1}^{\mathrm{T}}\boldsymbol{x'}_1}\boldsymbol{x'}_1 = \boldsymbol{x}_2 - \frac{\|\boldsymbol{x}_2\| \cdot \cos\theta_{21}}{\|\boldsymbol{x'}_1\|}\boldsymbol{x'}_1 \qquad (4.26)$$

θ_{21}是$\boldsymbol{x}_2$与$\boldsymbol{x'}_1$的夹角。$\boldsymbol{x'}_2$是$\boldsymbol{x}_2$减去$\boldsymbol{x}_2$向$\boldsymbol{x'}_1$的投影。$\boldsymbol{x}_2$与$\boldsymbol{x'}_1 = \boldsymbol{x}_1$线性独立，它不是$\boldsymbol{x'}_1$的数乘，所以$\boldsymbol{x'}_2$不是零向量。而且容易验证$\boldsymbol{x'}_2$与$\boldsymbol{x'}_1$正交。再令：

$$\boldsymbol{x'}_3 = \boldsymbol{x}_3 - \frac{{\boldsymbol{x}_3}^{\mathrm{T}}\boldsymbol{x'}_1}{{\boldsymbol{x'}_1}^{\mathrm{T}}\boldsymbol{x'}_1}\boldsymbol{x'}_1 - \frac{{\boldsymbol{x}_3}^{\mathrm{T}}\boldsymbol{x'}_2}{{\boldsymbol{x'}_2}^{\mathrm{T}}\boldsymbol{x'}_2}\boldsymbol{x'}_2 = \boldsymbol{x}_3 - \frac{\|\boldsymbol{x}_3\| \cdot \cos\theta_{31}}{\|\boldsymbol{x'}_1\|}\boldsymbol{x'}_1 - \frac{\|\boldsymbol{x}_3\| \cdot \cos\theta_{32}}{\|\boldsymbol{x'}_2\|}\boldsymbol{x'}_2 \qquad (4.27)$$

θ_{31}是$\boldsymbol{x}_3$与$\boldsymbol{x'}_1$的夹角，θ_{32}是$\boldsymbol{x}_3$与$\boldsymbol{x'}_2$的夹角。$\boldsymbol{x'}_3$是$\boldsymbol{x}_3$减去$\boldsymbol{x}_3$向$\boldsymbol{x'}_1, \boldsymbol{x'}_2$张成空间的投影。如果$\boldsymbol{x'}_3 = 0$，那么$\boldsymbol{x}_3$可以被$\boldsymbol{x'}_1, \boldsymbol{x'}_2$线性表出，也就可以被$\boldsymbol{x}_1, \boldsymbol{x}_2$线性表出，这与$\boldsymbol{x}_{i=1\dots n}$线性独立矛盾。故$\boldsymbol{x'}_3 \neq 0$。容易验证$\boldsymbol{x'}_3$正交于$\boldsymbol{x'}_1, \boldsymbol{x'}_2$。此过程继续下去，最终可构造一组正交向量$\boldsymbol{x'}_{i=1\dots n}$。这就是施密特正交化过程。将$\boldsymbol{x'}_{i=1\dots n}$的每一个向量除以各自的模，缩放到长度为 1，就得到了一组正交的单位向量。

4.1.3 特征值与特征向量

特征值和特征向量的概念不局限于方阵，但本书主要关注方阵。用方阵$\boldsymbol{A}$乘向量$\boldsymbol{x}$是在$\mathbb{R}^n$中进行一个变换，将$\boldsymbol{x}$变换成$\boldsymbol{A}\boldsymbol{x}$。例如矩阵：

$$\boldsymbol{R} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \qquad (4.28)$$

用$\boldsymbol{R}$乘向量$\boldsymbol{x}$等于将$\boldsymbol{x}$逆时针旋转θ度。这可以自行验证。任何方阵$\boldsymbol{A}$也改变不了零向量$\boldsymbol{0}$，因为$\boldsymbol{A}\boldsymbol{0} \equiv \boldsymbol{0}$。如果对于某非零向量$\boldsymbol{v} \neq \boldsymbol{0}$，$\boldsymbol{A}$只能改变$\boldsymbol{v}$的长度而不能改变其方向，即存在某个标量（可以为 0）λ，有：

$$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v} \qquad (4.29)$$

则称λ是$\boldsymbol{A}$的特征值，$\boldsymbol{v}$是$\boldsymbol{A}$的对于λ的特征向量。同一个特征向量不可能对应两个特征值。假如$\lambda_1 \neq \lambda_2$都有$\boldsymbol{v}$是其特征向量：

$$(\lambda_1-\lambda_2)\boldsymbol{v}=\lambda_1\boldsymbol{v}-\lambda_2\boldsymbol{v}=\boldsymbol{A}\boldsymbol{v}-\boldsymbol{A}\boldsymbol{v}=\boldsymbol{0} \quad (4.30)$$

$\lambda_1-\lambda_2\neq 0$且$\boldsymbol{v}\neq\boldsymbol{0}$，所以式（4.30）是不可能的。但是同一个特征值可以对应多个特征向量。如果$\boldsymbol{v}$是λ对应的特征向量，容易验证$k\boldsymbol{v}$也是λ对应的特征向量。线性独立的两个向量也有可能是同一个特征值对应的特征向量。

假如λ对应的特征向量$\boldsymbol{v}$和$\boldsymbol{w}$是线性独立的，即谁也不是另一个的数乘。那么$k\boldsymbol{v}+l\boldsymbol{w}$也是$\lambda$对应的特征向量。这也很容易验证。如果特征值$\lambda$共有$k$个线性独立的特征向量，由它们线性组合而得的向量也是λ的特征向量。这k个线性独立的特征向量张成的k维线性空间称为λ对应的特征空间，其中所有向量都是λ的特征向量。

将式（4.29）变形。如果λ是$\boldsymbol{A}$的特征值，它必须满足对某个$\boldsymbol{v}\neq\boldsymbol{0}$，有：

$$(\boldsymbol{A}-\lambda\boldsymbol{I})\boldsymbol{v}=\boldsymbol{0} \quad (4.31)$$

因为$\boldsymbol{v}\neq\boldsymbol{0}$，所以$\boldsymbol{A}-\lambda\boldsymbol{I}$的列线性相关。求$\boldsymbol{A}$的特征值和特征向量，就是求满足方程（4.31）的$\lambda$和$\boldsymbol{v}$。若要$\boldsymbol{A}-\lambda\boldsymbol{I}$的列线性相关，则$\boldsymbol{A}-\lambda\boldsymbol{I}$的行列式$|\boldsymbol{A}-\lambda\boldsymbol{I}|$等于 0。本书没有涉及行列式，因为行列式与本书主线关系不大，加进来会影响流畅性。读者可以查阅任何一种线性代数教材。$|\boldsymbol{A}-\lambda\boldsymbol{I}|=0$是$\lambda$的$n$次方程。它有$n$个根（包括重根和复数根）。即$\boldsymbol{A}$有$n$个特征值（包括重复的以及复特征值）。求得了$\lambda$，就可以再求它对应的特征向量。

矩阵$\boldsymbol{A}$属于不同特征值的特征向量是线性独立的。现在证明这一点。$\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_k$是$k$个不同特征值$\lambda_1,\lambda_2,\cdots,\lambda_k$对应的特征向量。如果它们线性相关，则其中某一个$\boldsymbol{v}_s$可以被其他$\boldsymbol{v}_{i\neq s}$线性表出：

$$\boldsymbol{v}_s=\sum_{i\neq s}a_i\,\boldsymbol{v}_i \quad (4.32)$$

因为$\boldsymbol{v}_s$是$\boldsymbol{A}$的特征向量，所以它不是零向量。那么$a_{i\neq s}$一定不全为 0。另外根据特征值和特征向量的定义：

$$\lambda_s\boldsymbol{v}_s=\boldsymbol{A}\boldsymbol{v}_s=\sum_{i\neq s}a_i\,\boldsymbol{A}\boldsymbol{v}_i=\sum_{i\neq s}a_i\,\lambda_i\boldsymbol{v}_i \quad (4.33)$$

如果$\lambda_s=0$，那么$\lambda_{i\neq s}\neq 0$。再加上$a_{i\neq s}$不全为 0，说明$\boldsymbol{v}_{i\neq s}$线性相关。在$\lambda_s=0$情况下我们将问题规模减小了 1。如果$\lambda_s\neq 0$，有：

$$\boldsymbol{v}_s = \sum_{i \neq s} a_i \frac{\lambda_i}{\lambda_s} \boldsymbol{v}_i \qquad (4.34)$$

于是式（4.34）等于式（4.32），所以有：

$$\sum_{i \neq s} \left(a_i \frac{\lambda_i}{\lambda_s} \boldsymbol{v}_i - a_i \boldsymbol{v}_i \right) = \sum_{i \neq s} a_i \left(\frac{\lambda_i}{\lambda_s} - 1 \right) \boldsymbol{v}_i = \mathbf{0} \qquad (4.35)$$

因为都是不同的特征值，所以$\lambda_{i \neq s} / \lambda_s - 1 \neq 0$。再加上$a_{i \neq s}$不全为 0，说明$\boldsymbol{v}_{i \neq s}$线性相关。在$\lambda_s \neq 0$情况下我们也将问题规模减小了 1。这个过程持续下去，最终将只剩下两个向量$\boldsymbol{v}_i$和$\boldsymbol{v}_j$。他们分属不同的特征值λ_i和λ_j。且$\boldsymbol{v}_i$和$\boldsymbol{v}_j$线性相关，其中一个是另一个的数乘。不妨假设$\boldsymbol{v}_i = k\boldsymbol{v}_j$，则$\boldsymbol{v}_i$也是$\lambda_j$的特征向量。之前已经证明，一个向量不可能同时属于两个不同特征值。这就推翻了最早的假设，证明了$\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_k$线性独立。

4.1.4　对称矩阵的谱分解

如果方阵$\boldsymbol{A}$满足:

$$\boldsymbol{A} = \boldsymbol{A}^{\mathrm{T}} \qquad (4.36)$$

如果$\boldsymbol{A}$的元素都是实数，则它是一个实矩阵。实矩阵的特征值都是实数，特征向量是实向量。为了证明这个结论，我们需要暂时离开实数域。

复数$\lambda = a + bi$的共轭是$\bar{\lambda} = a - bi$。

$$\lambda\bar{\lambda} = (a + bi)(a - bi) = a^2 + b^2 \geq 0 \qquad (4.37)$$

只有当$a = b = 0$，即$\lambda = 0$时，才有$\lambda\bar{\lambda} = 0$。否则$\lambda\bar{\lambda} > 0$。对于两个复数$\lambda = a + bi$和$\xi = c + di$，有：

$$\lambda\xi = (ac - bd) + (bc + ad)i = \bar{\lambda}\bar{\xi} \qquad (4.38)$$

把复矩阵$\boldsymbol{A}$的元素全都取共轭就得到$\boldsymbol{A}$的共轭$\bar{\boldsymbol{A}}$。如果（复数）λ和（复向量）$\boldsymbol{v}$是$\boldsymbol{A}$的特征值及对应特征向量，由式（4.38）容易看出：$\bar{\lambda}$和$\bar{\boldsymbol{v}}$是$\bar{\boldsymbol{A}}$的特征值及对应特征向量。

$$\bar{\boldsymbol{A}}\bar{\boldsymbol{v}} = \boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v} = \bar{\lambda}\bar{\boldsymbol{v}} \tag{4.39}$$

因为$\boldsymbol{A}$是实对称矩阵，有$\boldsymbol{A} = \bar{\boldsymbol{A}} = \boldsymbol{A}^{\mathrm{T}} = \bar{\boldsymbol{A}}^{\mathrm{T}}$，所以有：

$$\lambda\bar{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{v} = \bar{\boldsymbol{v}}^{\mathrm{T}}\lambda\boldsymbol{v} = \bar{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{v} = \bar{\boldsymbol{v}}^{\mathrm{T}}\bar{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{v} = (\bar{\boldsymbol{A}}\bar{\boldsymbol{v}})^{\mathrm{T}}\boldsymbol{v} = \left(\bar{\lambda}\bar{\boldsymbol{v}}\right)^{\mathrm{T}}\boldsymbol{v} = \bar{\lambda}\bar{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{v} \tag{4.40}$$

因为$\boldsymbol{x} \neq \boldsymbol{0}$，根据式（4.37）$\bar{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{v} > 0$。所以$\lambda = \bar{\lambda}$，即$\lambda$是实数。$\boldsymbol{A}$是实矩阵，$\lambda$是实数，所以$\boldsymbol{v}$一定是实向量。这就证明了实对称矩阵的特征值都是实数，特征向量是实向量。后文谈到矩阵都是实矩阵。

所以实对称矩阵$\boldsymbol{A}$有n个实特征值（可重复）。有一个结论我们不加证明：如果λ是$\boldsymbol{A}$的k重特征值（方程$|\boldsymbol{A} - \lambda\boldsymbol{I}| = 0$的$k$重根），则$\lambda$对应的特征空间是$k$维，即对于$\lambda$能找到$k$个线性独立的特征向量。