

第1章 逻辑回归

统计学习泰斗 Trevor Hastie 曾经说过：“每一个机器学习研究者应该像熟悉自家后院一样熟悉线性模型”。线性模型是许多非线性模型的基础。它良好的数学结构为理解非线性模型提供了深刻的洞见。神经网络的基本组成单元——神经元，正是一个线性模型。本书对于神经网络的介绍从线性模型开始。

1.1 作为一个神经元的逻辑回归

我们以逻辑回归（logistic regression）为例讨论线性模型。假如样本由 n 个数值型特征组成： x_1, x_2, \dots, x_n 。逻辑回归模型的计算式是：

$$f(x_1, x_2, \dots, x_n) = \frac{1}{1 + e^{-(b + w_1 x_1 + w_2 x_2 + \dots + w_n x_n)}} = \frac{1}{1 + e^{-(b + \sum_{i=1}^n w_i x_i)}} \quad (1.1)$$

其中 e 是自然对数的底。式（1.1）将各个特征 x_i 乘上对应的权重系数 w_i 后相加，之后再加上 b 。 b 称作偏置（bias）。这种计算称为对 x_1, x_2, \dots, x_n 的进行仿射变换（affine transform）。下文用 Σ 表示该仿射变换的结果：

$$\Sigma = b + \sum_{i=1}^n w_i x_i \quad (1.2)$$

仿射变换的几何意义将在下一节介绍。接下来对 Σ 施加 logistic 函数：

$$\text{logistic}(\Sigma) = \frac{1}{1 + e^{-\Sigma}} \quad (1.3)$$

logistic 函数的图像是一条S型曲线（sigmoid curve）。当 Σ 趋向于负无穷时函数值趋向于 0；当 Σ 趋向于正无穷时函数值趋向于 1。logistic 函数值在(0,1)区间内。当 $\Sigma = 0$ 时函数值是 0.5。如图 1.1 所示。

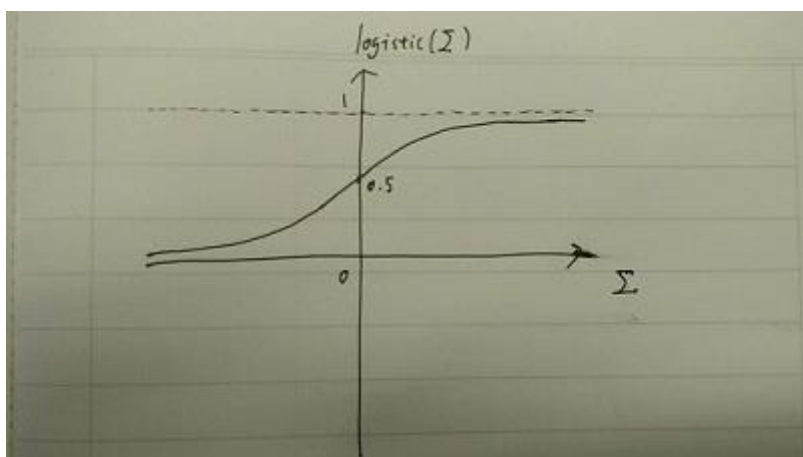


图 1-1 logistic 函数的图形

在二分类问题中，用 n 个数值型特征表示一个样本。例如用体重、身高、年龄表示一个人。这些样本属于两个互斥的类别——A 类或 B 类。如果要求模型告诉我们某一个样本属于哪一类，则可以把逻辑回归的输出当作样本属于 A 类的概率：

$$p_A = \frac{1}{1+e^{-\Sigma}} \quad (1.4)$$

因为样本必属于 A 类或 B 类，且不能既属于 A 类又属于 B 类，所以样本属于 A 类的概率与属于 B 类的概率之和为 1。于是样本属于 B 类的概率 p_B 就是：

$$p_B = 1 - p_A = 1 - \frac{1}{1+e^{-\Sigma}} = \frac{e^{-\Sigma}}{1+e^{-\Sigma}} \quad (1.5)$$

p_A 与 p_B 之比的对数（以 e 为底）是：

$$\log \frac{p_A}{p_B} = \log \frac{1}{1+e^{-\Sigma}} \cdot \frac{1+e^{-\Sigma}}{e^{-\Sigma}} = \log(e^{\Sigma}) = \Sigma \quad (1.6)$$

根据概率判定样本类别的决定权在人。如果规定当 $p_A \geq 1/2$ ，即 $p_A/p_B \geq 1$ 时将样本判定为 A 类，那么根据式（1.6）有：

$$\Sigma = \log \frac{p_A}{p_B} > \log^1 = 0 \quad (1.7)$$

以 0.5 为概率阈值时，类别的判定取决于 Σ 是否大于 0。 Σ 大于 0 具有什么几何意义？逻辑回归为什么属于线性模型？将在下文揭晓。在那之前我们先看看逻辑回归的一种图示。如图 1-2 所示。

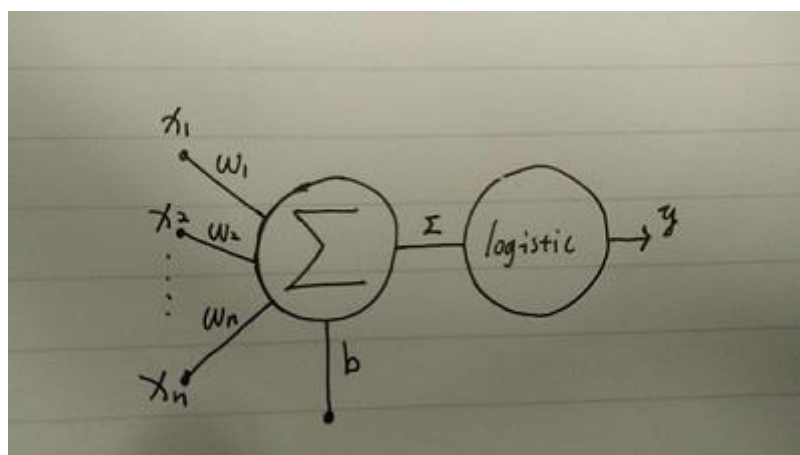


图 1-2 逻辑回归的神经元表示

该图示表达的就是逻辑回归：将输入值加权求和再加偏置之后施加 logistic 函数。在人工神经网络（本书以下省略“人工”二字）语境下，图示的结构就是一个神经元。在这里 logistic 函数被称为激活函数（activation function）。激活函数不限于 logistic 函数。如果激活函数取阶跃函数：

$$f(\Sigma) = \begin{cases} 0, & \Sigma < 0 \\ 1, & \Sigma \geq 0 \end{cases} \quad (1.8)$$

这就是最早的神经元模型——感知机（perceptron）。感知机这个名称一直流传下来，以至于今天多层全连接神经网络（multi-layer full connected network）还被称为 MLP（multi-layer perceptron）。关于激活函数的类型和性质本书后文还有阐述。无论取哪种激活函数，仿射+激活都是神经元的基本结构。

1.2 基础向量几何

神经元是神经网络乃至深度学习的基本砖石。在深入考察神经元的能力和局限之前有必要回顾一下基础的向量几何。

1.2.1 向量

一个包含 n 个数值型特征 x_1, x_2, \dots, x_n 的样本可用 n 维向量表示：

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = (x_1 \quad x_2 \quad \cdots \quad x_n)^T \quad (1.9)$$

本书用粗斜体小写字母表示向量，例如 \mathbf{x}, \mathbf{y} 。斜体小写字母表示标量（实数），例如 a, b 或 x 。用下标表示向量的分量，例如 x_2 表示 \mathbf{x} 的第2分量。 n 是分量的个数。样本有多少特征 \mathbf{x} 就有多少分量。本书向量指列向量（分量竖着排列）。有时为了省纸面空间把列向量写成行向量的转置（transpose），如式（1.9）第二个等号后。转置就是将行向量变成列向量，或将列向量变成行向量：

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}^T = (x_1 \quad x_2 \quad \cdots \quad x_n), \quad (x_1 \quad x_2 \quad \cdots \quad x_n)^T = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad (1.10)$$

为了可视化，本书多以2或3维向量为例。在现实问题中向量的维数会远远超过3。但是理论结果可以扩展到任意维度。

如果将各个分量看作坐标值，那么向量 \mathbf{x} 表示坐标系上的一个点（point）。 \mathbf{x} 也可以看作是从原点指向这个点的一个有长度有方向的“箭头”。这二者都是向量的几何表现形式。如图1-3所示。论述中根据视角不同采用这两种表现形式。

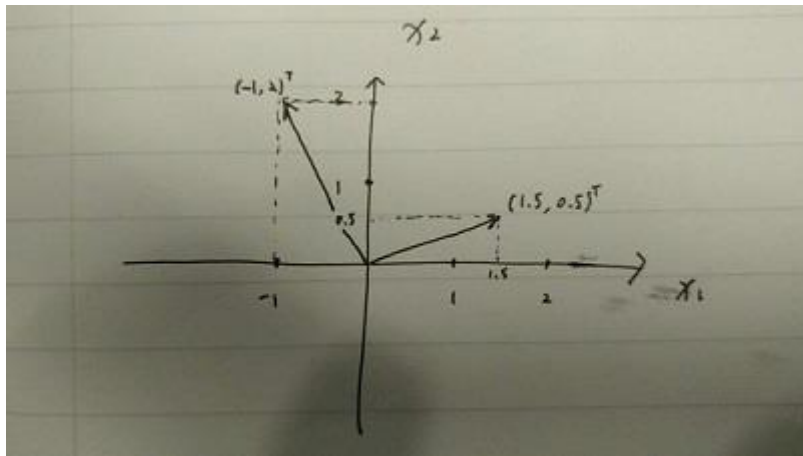


图 1-3 向量的几何表示

1.2.2 向量的和、数乘与零向量

向量可以求和。令 \mathbf{x} 和 \mathbf{y} 是两个维数相同的向量，定义它们的和 $\mathbf{x} + \mathbf{y}$ 是：

$$\mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix} \quad (1.11)$$

将 \mathbf{x} 和 \mathbf{y} 的对应分量相加作为 $\mathbf{x} + \mathbf{y}$ 的分量。如果用“箭头”来表示向量，那么 $\mathbf{x} + \mathbf{y}$ 是以 \mathbf{x} 和 \mathbf{y} 为相邻两边组成的平行四边形的对角线，从原点指向相对的顶点。该顶点也是 $\mathbf{x} + \mathbf{y}$ 的点表示。如图 1-4 所示。

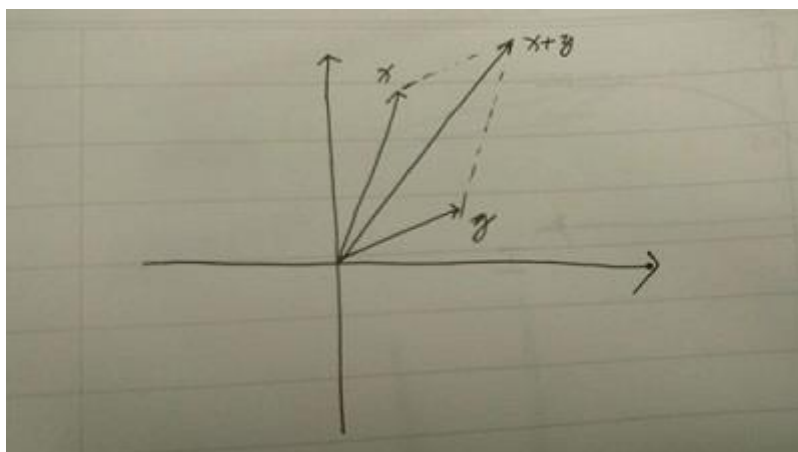


图 1-4 向量和的几何表示

可以用一个标量（实数） k 乘一个向量，即向量的数乘。定义 $k\mathbf{x}$ 为：

$$k\mathbf{x} = \begin{pmatrix} kx_1 \\ kx_2 \\ \vdots \\ kx_n \end{pmatrix} \quad (1.12)$$

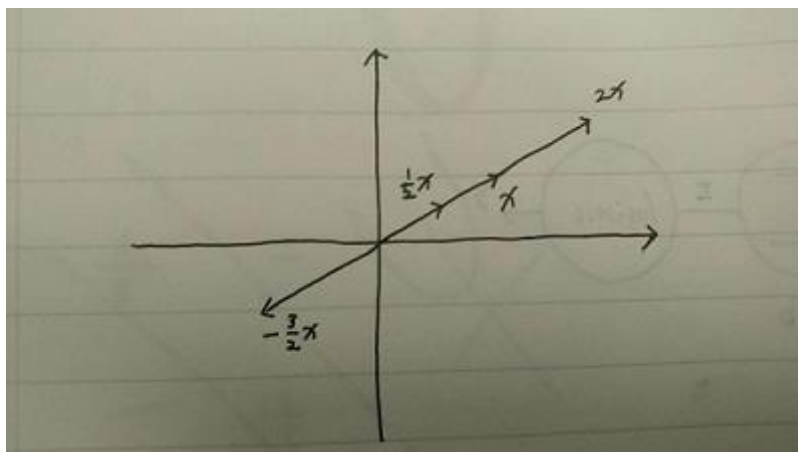


图 1-5 向量的数乘——缩放

数乘的几何意义是对向量进行缩放。如图 1-5 所示。用 2 乘向量 \mathbf{x} 得到 $2\mathbf{x}$ 。 $2\mathbf{x}$ 与 \mathbf{x} 方向相同，长度变成了 \mathbf{x} 的 2 倍。用 0 乘任何向量得到“零向量”：

$$\mathbf{0} = 0\mathbf{x} = \begin{pmatrix} 0x_1 \\ 0x_2 \\ \vdots \\ 0x_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (1.13)$$

$\mathbf{0}$ 是坐标系原点。 $\mathbf{0}$ 与任何向量相加都得到该向量本身。如果用 -1 乘向量 \mathbf{x} ，有：

$$-\mathbf{x} = (-1)\mathbf{x} = \begin{pmatrix} -x_1 \\ -x_2 \\ \vdots \\ -x_n \end{pmatrix} \quad (1.14)$$

显然 $-\mathbf{x}$ 与 \mathbf{x} 相加得 $\mathbf{0}$ 。向量的减法定义为：

$$\mathbf{y} - \mathbf{x} = \mathbf{y} + (-\mathbf{x}) \quad (1.15)$$

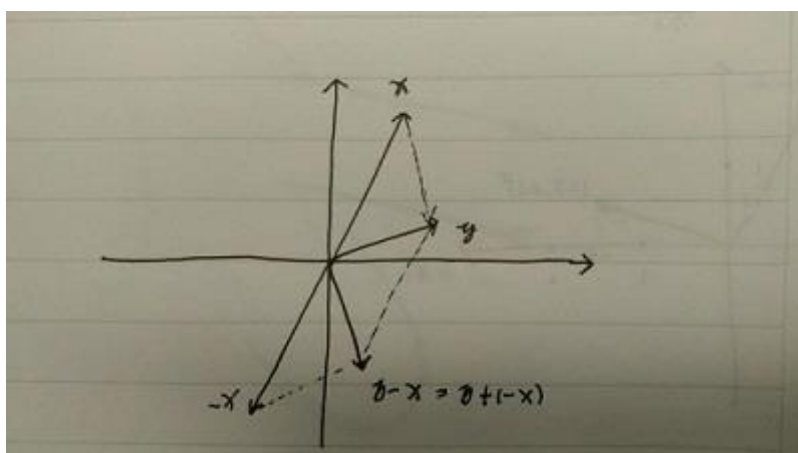


图 1-6 向量的差——三角形的第三边

将箭头 $y - x$ 平移，使其尾部与 x 重合，头部与 y 重合。则 $y - x$ 构成以 x 和 y 为临边的三角形的第三边。如图 1-6 所示。

“箭头”有长度。这个长度是向量表示的点与原点的距离。对于 2 维向量来说长度是：

$$\text{length}(x) = \sqrt{x_1^2 + x_2^2} \quad (1.16)$$

3 维乃至更高维依此类推。

1.2.3 向量的内积、模与投影

向量 x 和 y 的内积（inner product）的定义是：

$$\langle x, y \rangle = \sum_{i=1}^n x_i y_i \quad (1.17)$$

也就是将 x 和 y 的对应分量相乘再相加。根据式（1.17）， x 与自身的内积是：

$$\langle x, x \rangle = \sum_{i=1}^n x_i^2 \quad (1.18)$$

x 与自身的内积一定大于等于 0。只有当 x 的所有分量都是 0，即 $x = \mathbf{0}$ 时它与自身的内积为 0。内积满足交换律：

$$\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle \quad (1.19)$$

内积对向量加法满足结合律：

$$\langle \mathbf{w}, \mathbf{x} + \mathbf{y} \rangle = \langle \mathbf{w}, \mathbf{x} \rangle + \langle \mathbf{w}, \mathbf{y} \rangle, \quad \langle \mathbf{x} + \mathbf{y}, \mathbf{w} \rangle = \langle \mathbf{x}, \mathbf{w} \rangle + \langle \mathbf{y}, \mathbf{w} \rangle \quad (1.20)$$

内积对向量数乘也满足结合律：

$$\langle k\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, k\mathbf{x} \rangle = k\langle \mathbf{x}, \mathbf{y} \rangle \quad (1.21)$$

根据内积的定义，以上三条性质很容易证明。将向量 \mathbf{x} 的模 $\|\mathbf{x}\|$ 定义为 \mathbf{x} 与自身的内积的平方根：

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\sum_{i=1}^n x_i^2} \quad (1.22)$$

当 n 为2时 $\|\mathbf{x}\|$ 的定义与式(1.16)中定义的2维向量长度一致。所以可以将任意维向量的“长度”定义为模的平方根。向量 $k\mathbf{x}$ 的长度是：

$$\|k\mathbf{x}\| = \sqrt{\sum_{i=1}^n k^2 x_i^2} = \sqrt{k^2 \sum_{i=1}^n x_i^2} = |k| \|\mathbf{x}\| \quad (1.23)$$

根据式(1.23)，用标量 $1/\|\mathbf{x}\|$ 乘 \mathbf{x} 得到的向量 $\mathbf{x}/\|\mathbf{x}\|$ 的长度为1，方向与 \mathbf{x} 一致。也就是将 \mathbf{x} 缩放到长度1。长度为1的向量称为单位向量（unit vector）。

对于2维或3维的情况，上文提到 $\mathbf{y} - \mathbf{x}$ 经过平移是以 \mathbf{x} 和 \mathbf{y} 为临边的三角形的第三边。 $\mathbf{y} - \mathbf{x}$ 的模就是该第三边的长度。三角形第三边长度可用余弦公式求得：假如三角形三条边长是 a, b, c ， a 和 b 之间的夹角是 θ ，则 c 是：

$$c^2 = a^2 + b^2 - 2ab \cos \theta \quad (1.24)$$

当 $\theta = \pi/2$ 时 $\cos \theta = 0$ ，式(1.24)就是毕达哥拉斯定理。将三角形的边长代换为向量长度，有：

$$\|\mathbf{y} - \mathbf{x}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\|\mathbf{x}\|\|\mathbf{y}\|\cos\theta \quad (1.25)$$

根据模的定义将上式展开并消除等号两边相同项，得到：

$$\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\|\|\mathbf{y}\|\cos\theta \quad (1.26)$$

向量 \mathbf{x} 和 \mathbf{y} 的内积等于它们的长度之积再乘以它们之间夹角的余弦。在 2 或 3 维情况下向量之间的夹角有直观的定义。运用余弦定理得到式 (1.26)。在更高维情况下向量的夹角反过来由式(1.26)定义。如果向量 \mathbf{x} 和 \mathbf{y} 的内积为 0，则它们之间夹角的余弦是 0，夹角 θ 是 $\pi/2$ 。这种情况称 \mathbf{x} 和 \mathbf{y} 正交 (orthogonal)。如图 1-7 所示。若 \mathbf{x} 和 \mathbf{y} 都不是 \mathbf{o} ，则它们的“箭头”互相垂直。 \mathbf{o} 与任何向量正交。

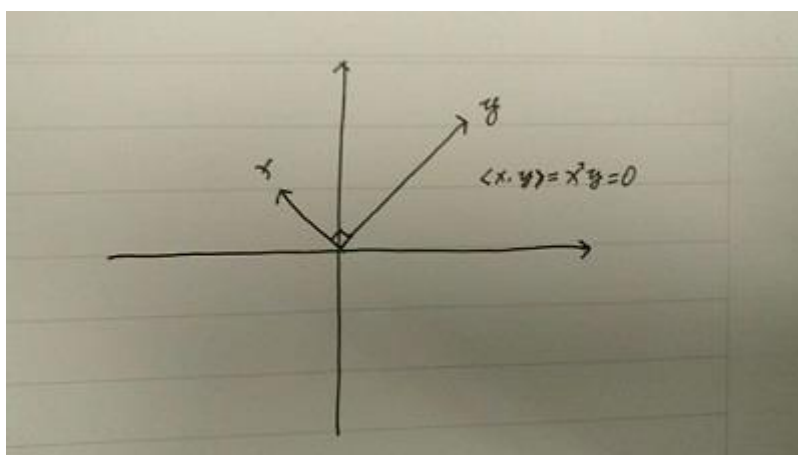


图 1-7 向量正交

如果 \mathbf{x} 和 \mathbf{y} 之间夹角为 θ ，那么 $\|\mathbf{x}\|\cos\theta$ 是 \mathbf{x} 向 \mathbf{y} 方向投影的长度。如图 1-8 所示。它等于 $\langle \mathbf{x}, \mathbf{y} \rangle / \|\mathbf{y}\|$ 。如果 \mathbf{y} 是单位向量，则 \mathbf{x} 向 \mathbf{y} 方向投影的长度就等于 $\langle \mathbf{x}, \mathbf{y} \rangle$ 。

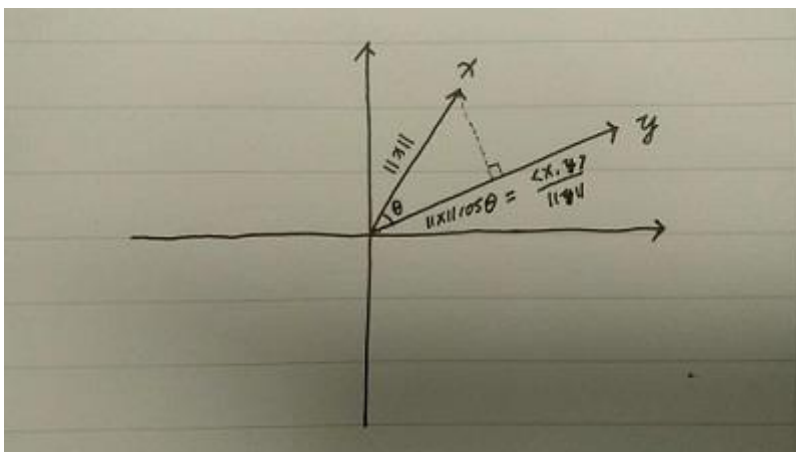


图 1-8 向量投影

如果把列向量看作 $n \times 1$ 矩阵，向量的内积可以用矩阵乘积表示：

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^n x_i y_i = (x_1 \quad x_2 \quad \cdots \quad x_n) \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{x}^T \mathbf{y} \quad (1.27)$$

1.2.4 线性空间、基和线性变换

对任何一个 n 维向量 \mathbf{x} 可以进行如下分解：

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \cdots + x_n \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} = \sum_{i=1}^n x_i \mathbf{e}_i \quad (1.28)$$

每一个 \mathbf{e}_i 是一个 n 维向量，其第 i 分量是 1，其余分量都是 0。称 \mathbf{x} 可以被 $\mathbf{e}_{i=1 \dots n}$ 线性表出。任意 n 维向量都可以被 $\mathbf{e}_{i=1 \dots n}$ 线性表出。而 $\mathbf{e}_{i=1 \dots n}$ 中的任何一个向量 \mathbf{e}_j 都无法由其他向量 $\mathbf{e}_{i \neq j}$ 线性表出。这是因为除了 \mathbf{e}_j ，其他向量的第 j 分量都是 0。这种情况称 $\mathbf{e}_{i=1 \dots n}$ 线性独立 (linear independent)。

全体 n 维向量构成 n 维线性空间 \mathbb{R}^n 。非严格地说，线性空间是向量的集合，其中全部向量都可以由一组线性独立的向量线性表出。这组向量是该线性空间的一组基 (basis)。基向量的个数是该线性空间的维数。 $\mathbf{e}_{i=1 \dots n}$ 是 \mathbb{R}^n 的一组基。或者说它们张成了 \mathbb{R}^n 。

$\mathbf{e}_{i=1 \dots n}$ 称为 \mathbb{R}^n 的标准基。除标准基外，任意 n 个线性独立的向量 $\mathbf{b}_{i=1 \dots n}$ 都能构成 \mathbb{R}^n 的基。因

为 $\mathbf{b}_{i=1\dots n}$ 是线性独立的，所以其中任一个 \mathbf{b}_j 都无法由 $\mathbf{b}_{i\neq j}$ 线性表出。 \mathbf{b}_j 可以分解成两个向量之和 $\mathbf{b}_j = \mathbf{b}_j^1 + \mathbf{b}_j^2$ 。其中 \mathbf{b}_j^1 可由 $\mathbf{b}_{i\neq j}$ 线性表出，而 \mathbf{b}_j^2 正交（垂直于）每一个 $\mathbf{b}_{i\neq j}$ 。 \mathbf{b}_j 提供了其他基向量无法提供的某个方向的信息，为构成 \mathbb{R}^n 做出了它独有的贡献。

线性独立有一个等价定义：不存在一组全为0的系数 w_1, w_2, \dots, w_n ，使 $\sum_{i=1}^n w_i \mathbf{b}_i = \mathbf{0}$ 。如果存在这样一组系数，假如其中 $w_j \neq 0$ ，那么 \mathbf{b}_j 就可以被其他向量线性表出： $\mathbf{b}_j = \sum_{i\neq j} w_i \mathbf{b}_i / w_j$ 。反之，如果 \mathbf{b}_j 可以被其他向量线性表出： $\mathbf{b}_j = \sum_{i\neq j} w_i \mathbf{b}_i / w_j$ ，那么 $\sum_{i\neq j} w_i \mathbf{b}_i / w_j - \mathbf{b}_j = \mathbf{0}$ ，即存在一组不全为0的系数（ \mathbf{b}_j 的系数是-1，不为0）使 $\sum_{i=1}^n w_i \mathbf{b}_i = \mathbf{0}$ 。

现在介绍线性变换（linear transform）。如果变换 f 是线性的，则对于任意两个向量 \mathbf{x} 和 \mathbf{y} 以及任意两个实数 a 和 b 有：

$$f(a\mathbf{x} + b\mathbf{y}) = af(\mathbf{x}) + bf(\mathbf{y}) \quad (1.29)$$

一个重要结论是：如果 f 是线性变换，则必然存在一个向量 \mathbf{w} ，对任意向量 \mathbf{x} 有：

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} \quad (1.30)$$

即线性变换执行的计算一定是用某个向量 \mathbf{w} 去与输入向量 \mathbf{x} 内积。这样构造 \mathbf{w} ：

$$\mathbf{w} = \begin{pmatrix} f(\mathbf{e}_1) \\ f(\mathbf{e}_2) \\ \vdots \\ f(\mathbf{e}_n) \end{pmatrix} \quad (1.31)$$

\mathbf{w} 的第 i 分量是对标准基第 i 个向量 \mathbf{e}_i 施加 f 后得到的标量值。由于 f 是线性变换，有：

$$f(\mathbf{x}) = f(\sum_{i=1}^n x_i \mathbf{e}_i) = \sum_{i=1}^n x_i f(\mathbf{e}_i) = (f(\mathbf{e}_1) \ f(\mathbf{e}_2) \ \dots \ f(\mathbf{e}_n)) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \mathbf{w}^T \mathbf{x} \quad (1.32)$$

这就证明了式（1.30）的结论。线性变换的图像一定过原点，因为：

$$f(\mathbf{0}) = f(0\mathbf{x}) = 0f(\mathbf{x}) = \mathbf{0} \quad (1.33)$$

1.2.5 直线、(超)平面与仿射变换

对于任意 $\mathbf{w} \neq \mathbf{0}$ ，令 \mathbf{x} 为任意与 \mathbf{w} 内积是常数 c 的向量：

$$\mathbf{w}^T \mathbf{x} = \langle \mathbf{w}, \mathbf{x} \rangle = \|\mathbf{w}\| \|\mathbf{x}\| \cos \theta = c \quad (1.34)$$

也就是说 \mathbf{x} 满足 $\|\mathbf{x}\| \cos \theta = c / \|\mathbf{w}\|$ 。即 \mathbf{x} 向 \mathbf{w} 方向投影的长度是常数 $c / \|\mathbf{w}\|$ 。在 2 维情况下将 \mathbf{x} 看作平面上一个点，则所有满足式 (1.34) 的点都位于垂直于 \mathbf{w} 的一条直线上。如图 1-9 所示。

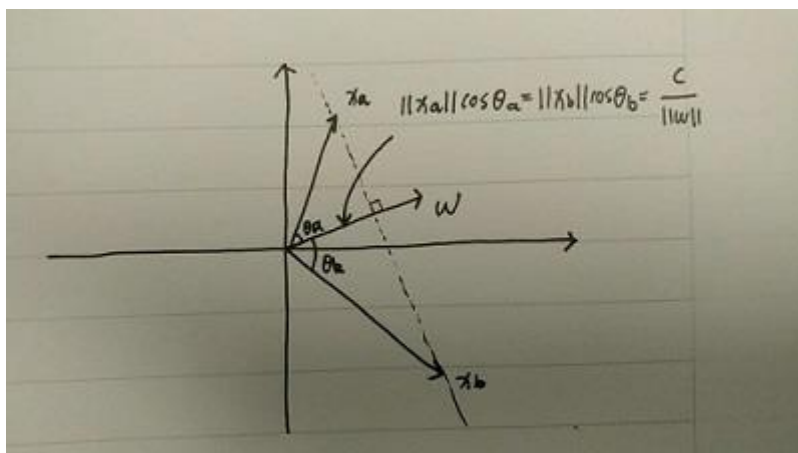


图 1-9 2 维空间中与特定非零向量内积相同的所有点位于垂直于该向量的直线上

在 3 维空间中所有与 $\mathbf{w} \neq \mathbf{0}$ 的内积为常数 c 的点构成垂直于 \mathbf{w} 的平面。如图 1-10 所示。在更高维空间中这样的点构成垂直于 \mathbf{w} 的超平面 (hyperplane)。

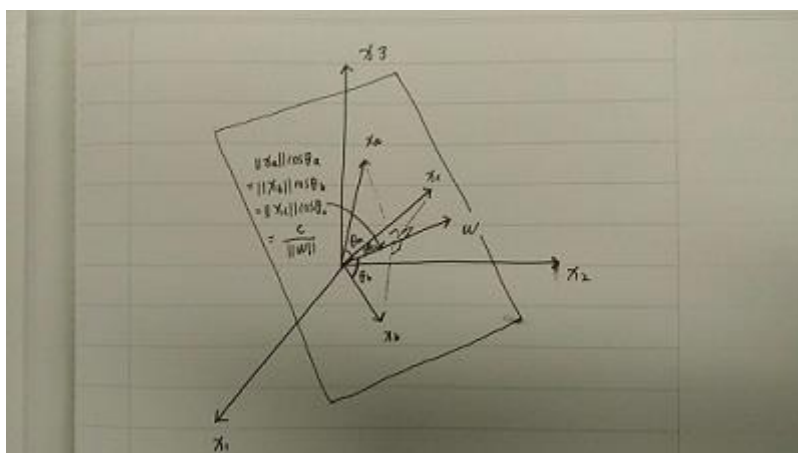


图 1-10 3 维空间中与特定非零向量内积相同的所有点位于垂直于该向量的平面上

\mathbf{w} 称为直线/（超）平面的法向量（norm）。对于任意非 0 实数 k ， $k\mathbf{w}$ 也是同一个直线/（超）平面的法向量。因为该直线/（超）平面上任意向量与 $k\mathbf{w}$ 的内积也是常数：

$$(k\mathbf{w})^T \mathbf{x} = \langle k\mathbf{w}, \mathbf{x} \rangle = |k| \|\mathbf{w}\| \|\mathbf{x}\| \cos \theta = |k|c \quad (1.35)$$

1.1 节曾提到仿射变换，将其重新写在这里：

$$y = b + \sum_{i=1}^n w_i x_i \quad (1.36)$$

如果自变量是 2 维，将 y 移到等号另一侧得到：

$$w_1 x_1 + w_2 x_2 - y = (w_1 \ w_2 \ -1) \begin{pmatrix} x_1 \\ x_2 \\ y \end{pmatrix} = -b \quad (1.37)$$

根据式（1.37），在 $x_1 x_2 y$ 三维空间中所有满足式（1.36）的点与 $\mathbf{w} = (w_1 \ w_2 \ -1)$ 的内积为常数 $-b$ 。也就是说仿射变换的图像是 3 维空间中一张平面。该平面的法向量是 \mathbf{w} 。当 $x_1 = x_2 = 0$ 时 y 的值是 b ，称为该平面的截距。由式（1.36）可以看出仿射变换是线性变换加上一个常量 b 。线性变换的图像是过原点，即截距 $b = 0$ 的直线\（超）平面。

若 w_1 和 w_2 的绝对值较大，则 \mathbf{w} 更贴近 $x_1 x_2$ 平面。以 \mathbf{w} 为法向量的平面更竖立。反之若 w_1 和 w_2 的绝对值较小，则 \mathbf{w} 更贴近 y 轴，以 \mathbf{w} 为法向量的平面较平躺。平面的倾斜程度与 w_1 和 w_2 的绝对值大小有关，也就是与 $(w_1 \ w_2)$ 的模有关。

$(w_1 \ w_2 \ 0)$ 是 \mathbf{w} 在 $x_1 x_2$ 平面上的投影。它的方向决定了平面的朝向。总之， $\mathbf{w} = (w_1 \ w_2 \ -1)$ 是 2 维仿射变换的图像（3 维空间中一张平面）的法向量。 $(w_1 \ w_2)$ 的模决定了平面的倾斜程度， $(w_1 \ w_2)$ 的方向决定了平面的朝向。图 1-11 集中展示这几个概念。

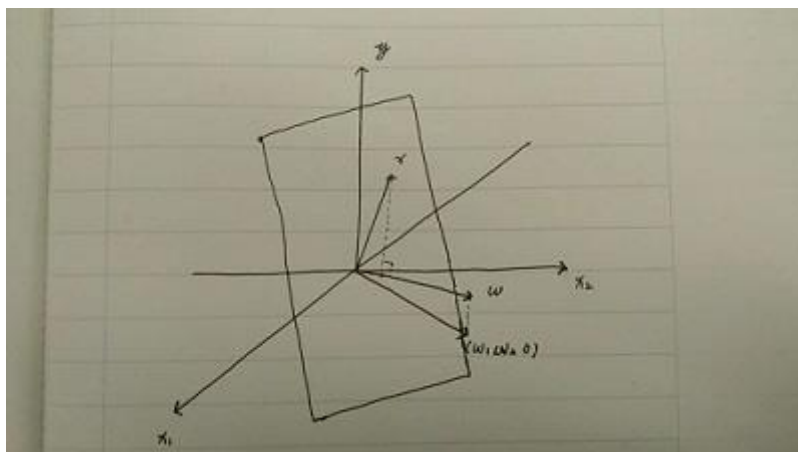


图 1-11 法向量决定平面的朝向和倾角

仿射变换是一类最简单的变换。仿射变换的图像在自变量为 1 维的情况下是直线；在自变量为 2 维的情况下是平面；在更高维情况下是超平面。这类图像在各处的特性是相同的。想象一个平坦的斜坡。无论处于斜坡上的哪一点，朝向相同方向的倾斜程度都是相同的。

仿射函数是更复杂的函数在某一点局部的最粗糙的近似，称一阶近似。训练逻辑回归或神经网络的梯度下降法就是一种基于损失函数局部一阶近似特性的优化算法。这些内容本书后文将详加介绍。

1.3 从几何角度理解逻辑回归的能力和局限

回忆一下逻辑回归的表达式：

$$f(\mathbf{x}) = \frac{1}{1+e^{-(b+\sum_{i=1}^n w_i x_i)}} = \frac{1}{1+e^{-(b+\mathbf{w}^T \mathbf{x})}} \quad (1.38)$$

用我们现在的语言表述，式 (1.38) 是对输入向量 \mathbf{x} 做仿射变换后再施以 logistic 函数。还是以 2 维为例，某条以 $\mathbf{w} = (w_1 \ w_2)$ 为法向量的直线上的点与 \mathbf{w} 的内积为常量。所以对它们施加仿射变换 $b + \mathbf{w}^T \mathbf{x}$ 后的值都相同，再施加 logistic 函数后的结果也都相同。逻辑回归先对输入向量 \mathbf{x} 用 \mathbf{w} 做仿射变换，这样就丢弃了 \mathbf{x} 在垂直于 \mathbf{w} 方向上的信息。于是逻辑回归的输出只在 \mathbf{w} 的方向上有变化；在垂直于 \mathbf{w} 的方向上无变化。如果用阈值 t 去卡这个输出：

$$\text{output}(\mathbf{x}) = \begin{cases} 0, & f(\mathbf{x}) < 0 \\ 1, & f(\mathbf{x}) \geq 0 \end{cases} \quad (1.39)$$

则只能得到垂直于 \mathbf{w} 的直线分界线。这也就是逻辑回归属于线性模型的原因。仿射变换是“罪

罪魁祸首”。它先对信息进行了过滤。即使之后再施以S形的 logistic 函数也于事无补。如图 1-12 所示。

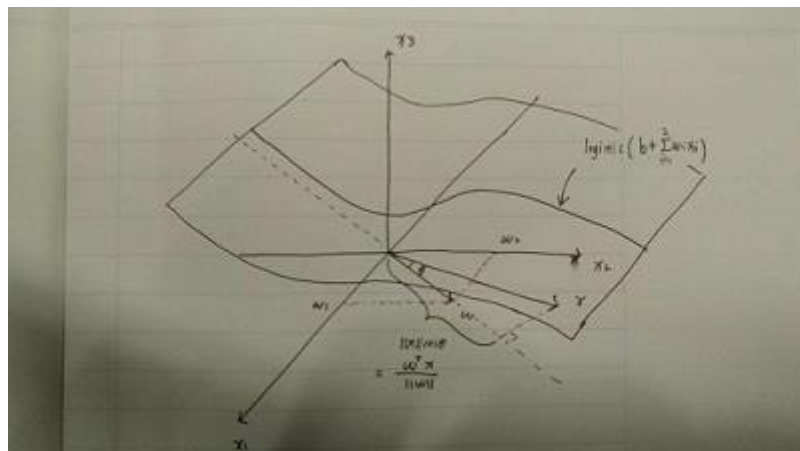


图 1-12 单个神经元只能形成直线/超平面分界线/面

神经元可以选用其他激活函数。但只要激活函数仍是单调的，则只能形成一条直线分界线(在更高维情况下就是超平面分界面)。单调激活函数神经元无法处理异或问题 (XOR)。因为异或问题是线性不可分的——无法用直线将正负样本点分隔开。如图 1-13 所示。

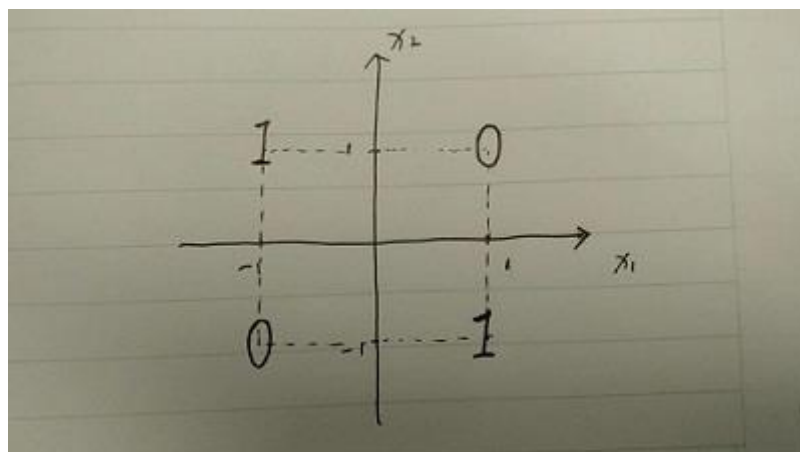


图 1-13 异或问题

如果使用非单调函数作为激活函数，比如平方函数：

$$f(x) = (b + \mathbf{w}^T \mathbf{x})^2 \quad (1.40)$$

通过调整 \mathbf{w} 的方向可以解决异或问题。如图 1-14 所示。

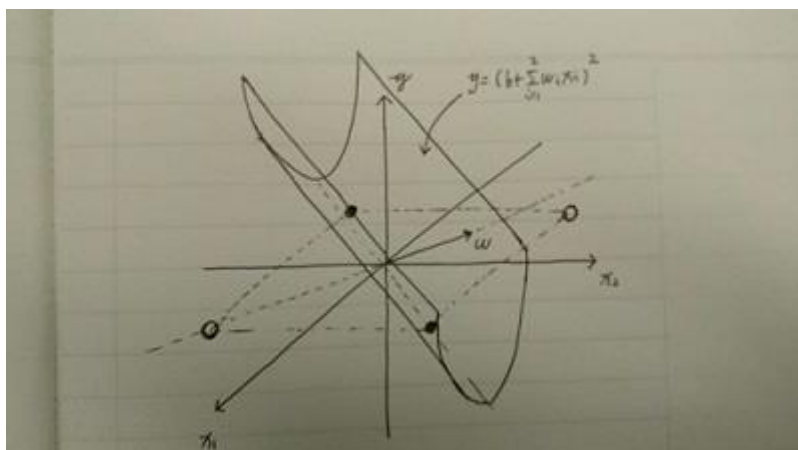


图 1-14 使用平方激活函数的单个神经元可以解决异或问题

但是式 (1.40) 的能力也仅限于用两条平行直线去分割自变量空间。就算使用正弦 \sin 函数作为激活函数也只能形成无数条平行分割线。它们对于同心圆状分布的数据是无能为力的。如图 1-15 所示。

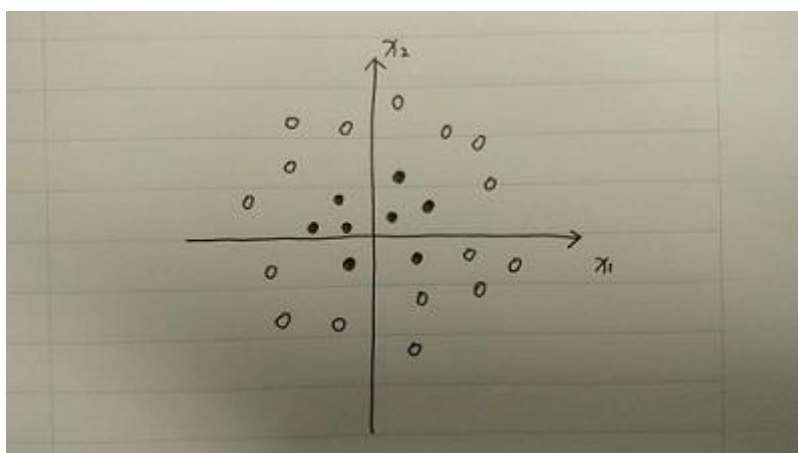


图 1-15 同心圆状数据

同心圆数据的正类点在中央，负类点围绕在外围。在 x_1x_2 平面内不存在一条直线将正负两类点分开。该数据含有各个方向的分类信息，而单个神经元只能辨识一个方向的信息，所以无法用一个神经元分开同心圆数据。要想获得超越线性的能力，必须将多个神经元连接成网络。允许多个神经元合作形成非线性分界线。这就是神经网络的思想。

1.4 小结

本章介绍了作为线性模型的逻辑回归模型。它同时也是一个神经元。在回顾了必要的向量几

何背景知识后，我们从几何角度阐述了逻辑回归的能力及其限制。我们知道逻辑回归如何划分自变量空间取决于权值向量 \mathbf{w} 和偏置 b 。但本章没有阐述对于某一个特定分类问题如何确定 \mathbf{w} 和 b 的取值。这就是逻辑回归模型的训练问题。