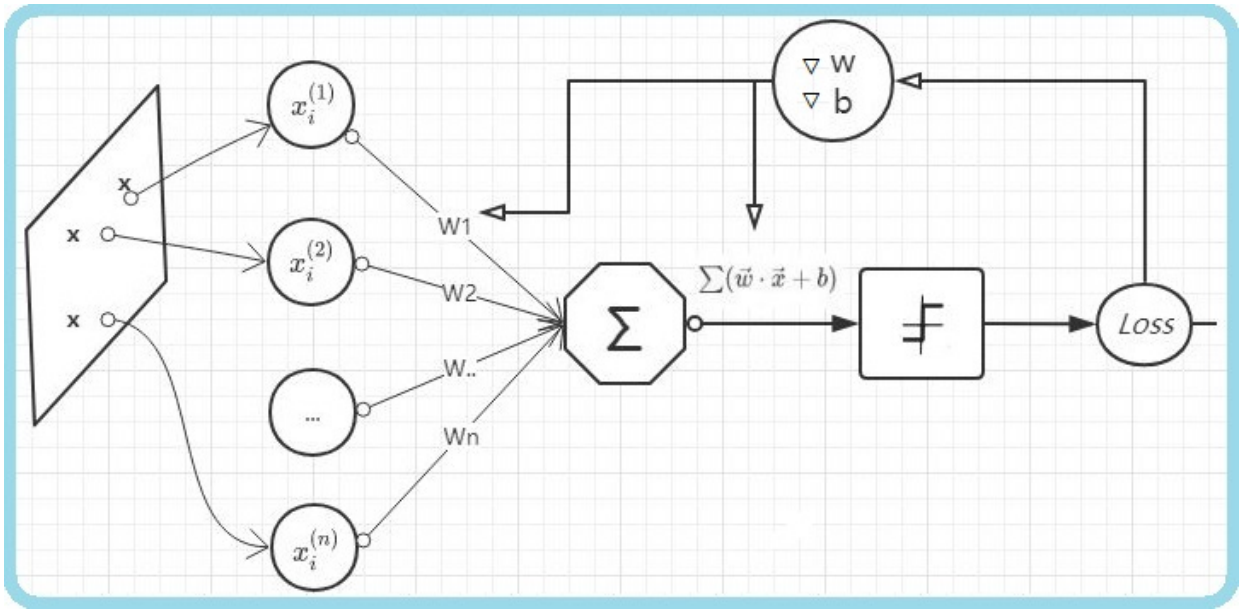



# 从0到1：神经网络实现图像识别（上）

**SIGAI特邀作者：无双谱 JerryX007Srv**

**作者简介**：金融应用架构专家

**研究方向**：机器学习的特定场景应用，复杂项目群管理

---

> 纸上得来终觉浅，绝知此事要躬行。

“神经网络”是“机器学习”的利器之一，常用算法在**TensorFlow**、**MXNet**计算框架上，有很好的支持。

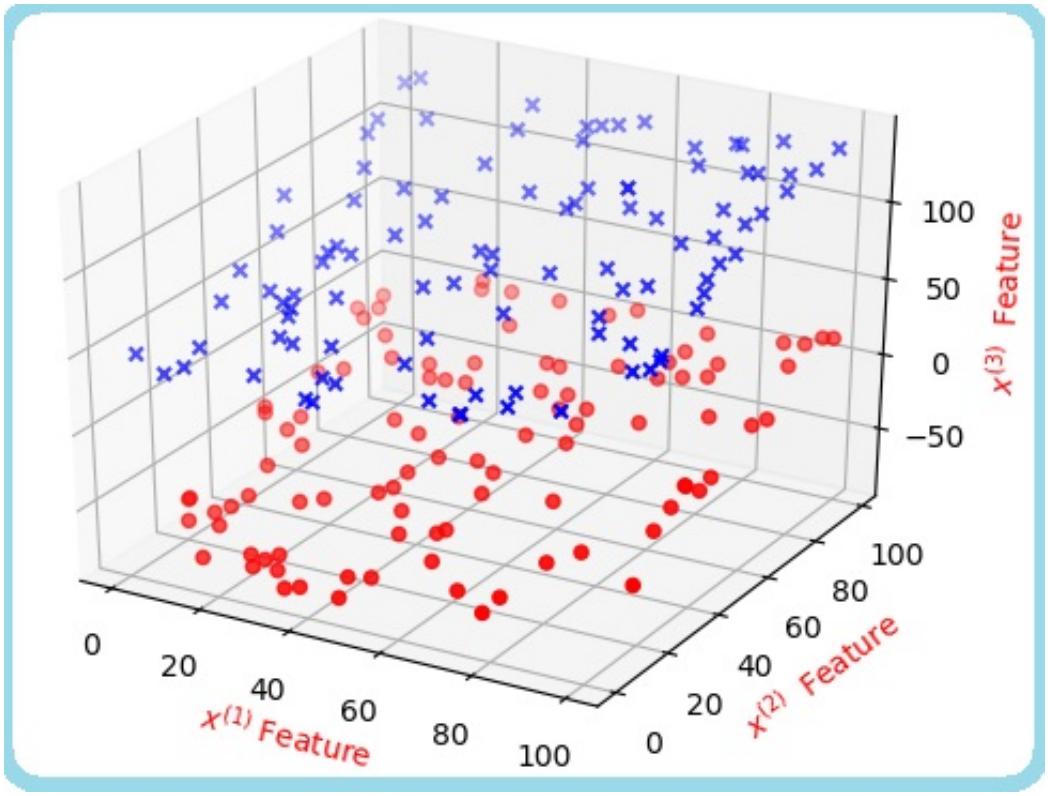
为了更好的理解与使用这件利器，我们可以**不借助计算框架**，从零开始，一步步构建模型，实现学习算法，并在一个图像识别数据集上，训练这个模型，再验证模型预测的准确率。

首先，我们来了解一个简洁的分类模型-感知机（perceptron）模型，感知机是1957年由Rosenblatt提出的线性二类分类模型，也是人工神经网络方法的理论基石。

> By the study of systems such as the perceptron, it is hoped that those fundamental laws of organization which are common to all information handling systems, machines and men included,may eventually understood.
>
> 探究感知机这类体系，我们有望**最终理解**那些**基本法则**，那些将"信息认知"，赋能于机器和人类的基本法则。
>
> - Frank Rosenblatt@Connell Aeronautical Laboratory

## 感知机模型

想像D维空间里分布着$N$个线性可分的实例点$x_i$，当D=3时，这个空间即是一个便于理解的三维空间，其上的任意实例点 $x_i$，都三个特征 $x_i^{(1)}$ 、 $x_i^{(2)}$ 和 $x_i^{(3)}$；实例点看作是3维实数向量 $\vec{x} = (x^{(1)}, x^{(2)}, x^{(3)})^T$，三个特征决定了实例点 $x_i$的二分类类别 $y_i = \{+1, -1\}$。




由输入实例点$x_i$特征，到输出类别$y_i$的映射，可表示为如下**感知机**函数:

$$f(x) = sign(\vec{w} \cdot \vec{x} + b)$$

其中" · "表示两个向量的内积(inner product) 运算, $\vec{w}$称为权值向量(weight vector)， $b$称为偏置(bias)。

$sign(\text{自变量})$是符号函数，将**自变量**，进一步映射到$y_i$的输出类别$\{+1, -1\}$上。

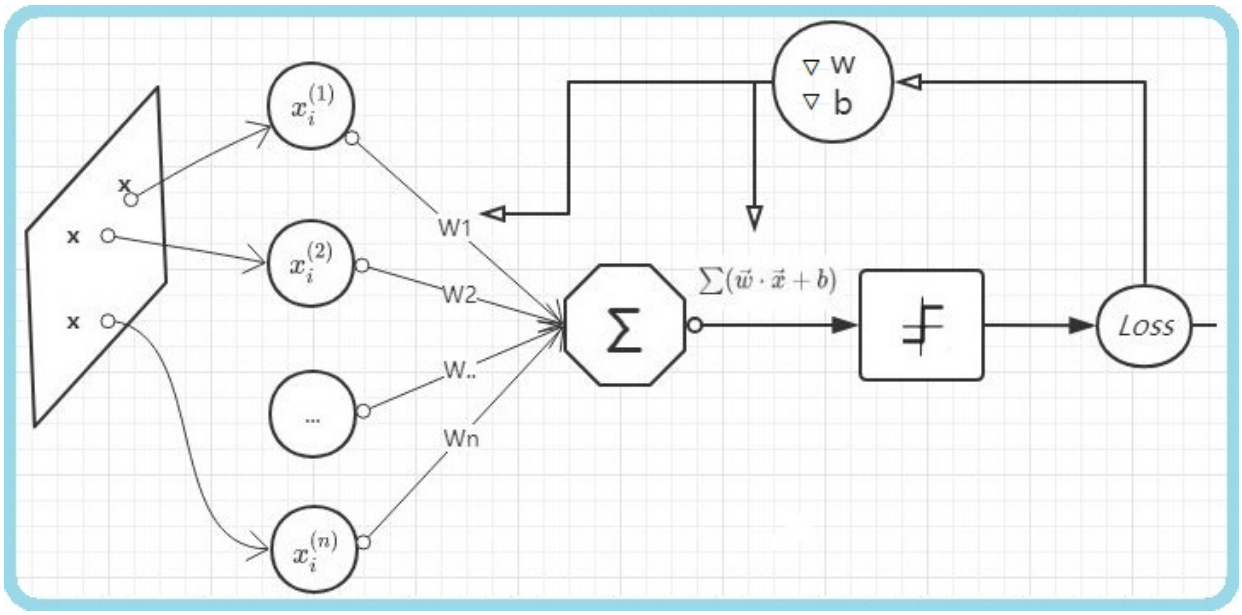
接下来定义D维空间里超平面，用线性方程表示:

$$\vec{w} \cdot \vec{x} + b = 0$$

这个超平面按照输出类别$y_i = \{+1, -1\}$将实例点 $\vec{x}_i$划分在平面两侧，

对正实例点$y_i = +1$ 有 $\vec{w} \cdot \vec{x}_i + b > 0$

对负实例点$y_i = -1$ 有 $\vec{w} \cdot \vec{x}_i + b < 0$




根据平面点法式方程的定义，$\vec{w}$是超平面的法向量，需要回顾的话，可以通过以下步骤反证：

1. 由于平面上任意两个不同实例点，$\vec{x}_i$和$\overrightarrow{x_{i+n}}$ 都满足

   $$\vec{w}\cdot(\vec{x}_i-\vec{x}_{i+n})=\vec{w}\cdot\overrightarrow{(x_{i+n}-x_i)}=0$$

2. 而内积的几何意义，是一个向量在另一个向量方向上的投影长度，与另一个向量长度的乘积。

可知$\vec{w}$是与分离超平面上任意切向量正交的法向量。

通过某种学习策略找到权值向量（法向量） $\vec{w}$ 和偏置（截距） $b$ 这两个参数，确定划分正负实例点的超平面，就可以对输入的D维空间上散布的线性可分实例点$x_i$，做二类分类预测。

## 学习策略

---

为了找到合适的权值向量$\vec{w}$ 和偏置 $b$，首先定义连续可导的损失函数（loss function），再将损失函数极小化，以找到所有可能的分离超平面中，较优的一个；如果损失函数是凸函数，还可以用数值方法，得到全局最优解，学习策略就成为求解最优化问题：

$$\min_{f\in F}\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))$$

$F$是所有能划分输入样本点的感知机模型$f$的集合，$N$是（训练）样本容量，$L$是模型$f$的损失函数。

损失函数仅仅是一次预测的好坏度量，然而根据**大数定理**，当样本容量$N$足够大时，样本点集合上，由损失函数得到的平均损失，趋近于总体分布在分类模型上的期望损失，从而可以用数理统计方法得到理想概率模型的近似解。

损失函数有多种经典选择，对二类分类问题，可以选择造成模型损失的误分类点，到分离超平面的总距离，来度量损失：

对任意一个样本点 $\vec{x}_i$ , 我们可以根据点到平面的距离公式，得出它到超平面 $\vec{w}\cdot\vec{x}+b=0$的距离：

$$\frac{|\vec{w}\cdot\vec{x}_i+b|}{\sqrt{\sum_{j=1}^{D}w_j^2}}$$

其中，$\sqrt{\sum_{j=1}^{D} w_j^2}$ 是法向量 $\vec{w}$ 到D维空间原点的欧式距离，也称为$\vec{w}$的$L_2$ 范数 ，记作 $\|w\|$ ，

同时，根据超平面的定义，一个误分类的实例点 $x_{err}$ ，有:

$$-y_i(w \cdot x_{err} + b) > 0$$

得到所有误分类实例点到超平面的总距离为：

$$-\frac{\sum y_{err}(\vec{w} \cdot \vec{x}_{err} + b)}{\sqrt{\sum_{j=1}^{D} w_j^2}} = -\frac{\sum y_{err}(\vec{w} \cdot \vec{x}_{err} + b)}{\|w\|}$$

在优化损失函数使损失极小时，函数取值的数值**缩放正倍数**不影响优化方法，所以损失函数可以进一步写为：

$$L(\vec{w}, b) = -\sum y_{err}(\vec{w} \cdot \vec{x}_{err} + b)$$

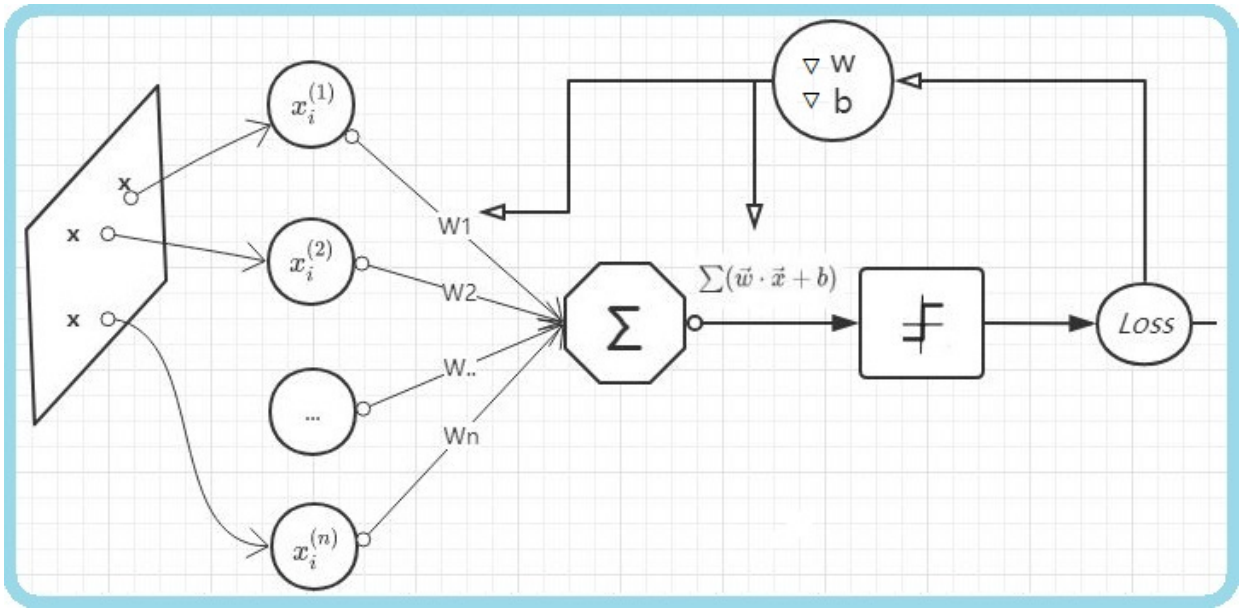
直观理解，损失越小，误分类点距离超平面越近，直到所有样本点都被正确分类，损失为0，可知 $L(\vec{w}, b)$ 是 $\vec{w}$,b 的连续可导函数。

求解最优化问题：

$$\min_{f \in F} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$$

变成了损失函数极小时，参数 $\vec{w}$,b 的求解：

$$\min_{\vec{w}, b} L(\vec{w}, b) = -\sum y_{err}(\vec{w} \cdot \vec{x}_{err} + b)$$




## 学习算法

一种求解算法是 **随机梯度下降**（**s**tochastic **g**radient **d**escent），先为 $\vec{w}$,b 设置初始值如 0 ，然后用梯度下降法，让参数不断更新梯度 $\nabla\vec{w}$ 和$\nabla b$，来极小化损失函数。

$$\nabla_w Loss = \frac{\partial[-\sum y_{err}(\vec{w} \cdot \vec{x}_{err} + b)]}{\partial \vec{w}} = -\sum y_{err} x_{err}$$

同样：

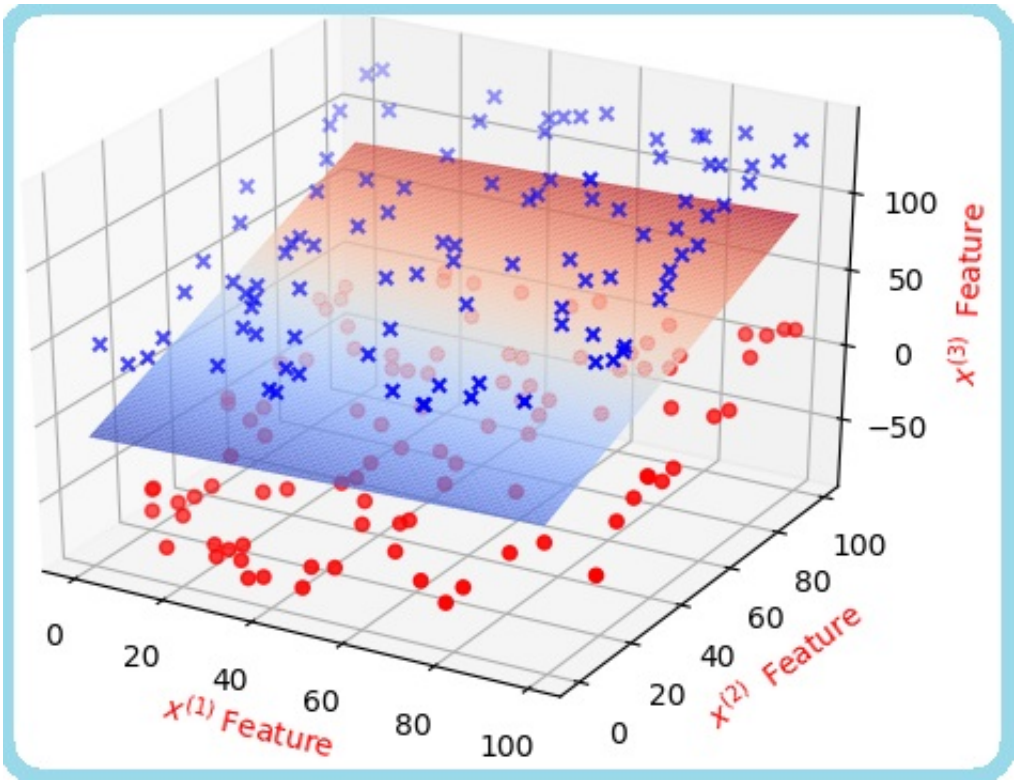
$$\nabla_b Loss = \frac{\partial[-\sum y_{err}(\vec{w}\cdot\vec{x}_{err}+b)]}{\partial b} = -\sum y_{err}$$

为便于灵活调整每次梯度下降的尺度，引入参数“步长” 或“学习率” $\eta$，根据随机选出的误分类点，来更新参数：

$$w = w + \eta\, x_{err}\; y_{err}$$

$$b = b + \eta\, y_{err}$$

可以证明，这个算法在线性可分数据集上是收敛的；通过不断随机选取误分类点，更新$w$和$b$，通过有限次迭代，能找到一个可以把线性可分正负实例点划分在两侧的分离超平面。




---

至此，我们看到感知机解决了D维空间内，N个线性可分实例点的二分类问题；那么这个方法是否能处理**多类分类问题**, 比如之前介绍的，MNIST数据集上的10类分类问题呢?

答案是肯定的。

下一次，我们把感知机模型改进推广到分类类别 K>2 的情况，并根据改进后策略和学习算法，在MNIST手写数字识别数据集上，训练模型参数，初步得到一个识别率尚可（>90%）的结果。

(上篇完)

**参考**

[1] 李航 .统计学习方法. 北京：清华大学出版社，2012

[2] Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review.1958,65(6), 386-408