感知机（Perceptron）

翟婷婷

扬州大学
信息工程（人工智能）学院
zhtt@yzu.edu.cn

2023年春

课程目标

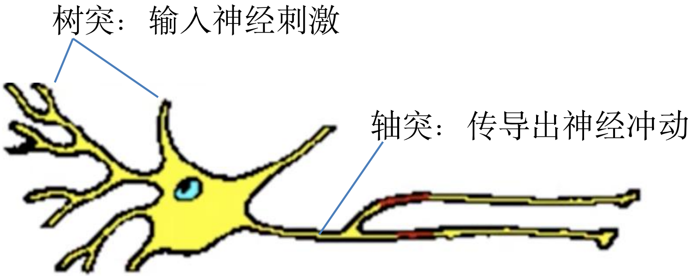
- 理解感知机的工作原理。
- 掌握两种不同的感知机学习算法，并能编程实现。
- 能够分析感知机算法的收敛性。

问题的提出

- 回顾最简单的一种分类器形式——线性分类器，由线性判别函数及相应的分类决策规则构成的。
- 线性判别函数的参数$(\mathbf{w}, b)$是如何得到的呢?
- 本节关注——线性分类器的训练问题：基于一个训练数据集学习得到线性判别函数的参数$(\mathbf{w}, b)$。
- 本节学习如何使用感知机进行线性分类器的训练。

感知机

- 1957年由美国计算机科学家Rosenblatt提出，是神经网络与支持向量机的基础。
- 感知机是模拟**神经元**功能的一种数学模型。仿生学模型



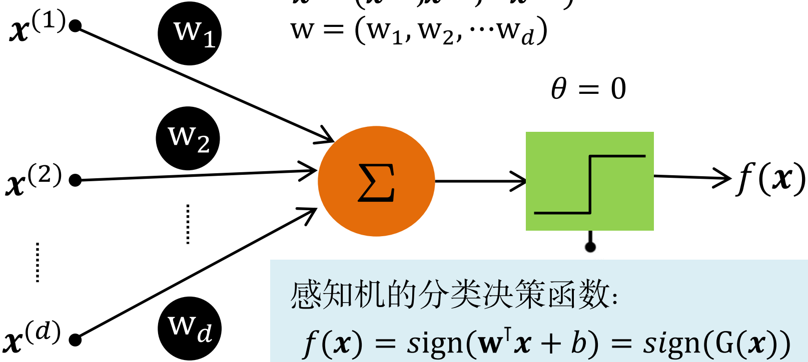
神经元细胞的工作机理：多个树突所搜集到的输入信号，经过神经元的集中处理，如果达到一定的激发阈值，就会激活轴突的输出。

感知机

- 感知机模拟一个神经元细胞的工作机理。

$$\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots \boldsymbol{x}^{(d)})$$

$$\mathrm{w} = (\mathrm{w}_1, \mathrm{w}_2, \cdots \mathrm{w}_d)$$



感知机的分类决策函数:

$$f(\boldsymbol{x}) = \mathrm{sign}(\mathbf{w}^{\mathsf{T}}\boldsymbol{x} + b) = sign(\mathrm{G}(\boldsymbol{x}))$$

$$\mathrm{sign}(a) = \begin{cases} +1, & a > 0 \\ -1, & a \leq 0 \end{cases}$$

准备工作

- 线性判别函数的一般形式为：

$$G(\boldsymbol{x}) = \mathbf{w}^{\mathsf{T}}\boldsymbol{x} + b$$

其未知量为权向量$\mathbf{w} = (\mathrm{w}_1, \mathrm{w}_2, \cdots \mathrm{w}_d)$和偏置$b$，线性分类器训练的过程就是找到$\mathbf{w}$和$b$的合适取值的过程。

- 为了便于表述，将偏置b并入到权向量$\mathbf{w}$中：

$$\mathbf{w}^{\mathsf{T}}\boldsymbol{x} + b = (\mathrm{w}_1, \cdots \mathrm{w}_d, b) \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_d \\ 1 \end{pmatrix} = \mathbf{w}'^{\mathsf{T}}\boldsymbol{x}'$$

其中，$\mathbf{w}' = (\mathrm{w}_1, \cdots \mathrm{w}_d, b)$，$\boldsymbol{x}' = (x_1, \cdots x_d, 1)$。

- 判别函数的形式变为：

$$G(\boldsymbol{x}') = \mathbf{w}'^{\mathsf{T}}\boldsymbol{x}'$$

线性分类器的训练就是寻找模型参数$\mathbf{w}'$的合适取值。

感知机模型的训练问题

- 给定一个线性可分的二分类训练数据集：

$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots\cdots (\boldsymbol{x}_N, y_N)\}$$

其中$\boldsymbol{x}_i \in \mathbb{R}^d$为第$i$个实例，$y_i \in \{-1, +1\}$为$\boldsymbol{x}_i$的类标记，$(\boldsymbol{x}_i, y_i)$称为一个样本。($\boldsymbol{x}_i$是扩充后的特征向量)

- 感知机模型的训练问题描述为：

利用给定的训练数据集，求得权向量$\mathbf{w}$，使得它能对所有训练样本进行正确分类，也即是：

对于所有$y_i = +1$的实例$\boldsymbol{x}_i$，有$\mathbf{w}^\top \boldsymbol{x}_i > 0$

对于所有$y_i = -1$的实例$\boldsymbol{x}_i$，有$\mathbf{w}^\top \boldsymbol{x}_i < 0$

(y_i的符号与$\mathbf{w}^\top \boldsymbol{x}_i$的符号一致，表示$\boldsymbol{x}_i$被正确分类)

等价于： 求得这样的$\mathbf{w}$，对于所有的训练样本，都有

$$y_i \mathbf{w}^\top \boldsymbol{x}_i > 0, \ \forall i = 1, \cdots N.$$

感知机的学习策略

- 一般地，分类器的学习策略为：

① 设计一个关于**w**的准则函数(损失函数)，其值能够代表**w**的优劣程度，准则函数值越小，说明**w**越符合要求，越好;

② 通过寻找准则函数的极小值，找到最优的一个**w**。

- 感知机的准则函数：
  - ✓ 定义：所有误分类的样本到当前决策超平面的总距离。
  - ✓ 只要存在错分类的样本，准则函数值就大于 0 的，只有当所有样本都正确地被分类了，准则函数才能取得极小值 0.

感知机的准则函数的表示

- 记$G(\boldsymbol{x})=\mathbf{w}^\top\boldsymbol{x}=0$表示当前的决策超平面；
- 记M为被$G(\boldsymbol{x})$误分类的样本的集合；
- 对于$\forall(\boldsymbol{x}_i,y_i)\in M$，满足$y_i\mathbf{w}^\top\boldsymbol{x}_i\le 0$
- $\forall(\boldsymbol{x}_i,y_i)\in M$，其到决策超平面的距离为：

$$\frac{|G(\boldsymbol{x}_i)|}{||\mathbf{w}_{1:d-1}||}=\frac{|\mathbf{w}^\top\boldsymbol{x}_i|}{||\mathbf{w}_{1:d-1}||}=\frac{|y_i\mathbf{w}^\top\boldsymbol{x}_i|}{||\mathbf{w}_{1:d-1}||}=\frac{-y_i\mathbf{w}^\top\boldsymbol{x}_i}{||\mathbf{w}_{1:d-1}||}$$

- M中所有样本到决策超平面的总距离为：

$$\frac{1}{||\mathbf{w}_{1:d-1}||}\sum_{\boldsymbol{x}_i\in M}-y_i\mathbf{w}^\top\boldsymbol{x}_i$$

- 不考虑$\frac{1}{||\mathbf{w}_{1:d-1}||}$，感知机的准则函数为：

$$J(\mathbf{w})=\sum_{\boldsymbol{x}_i\in M}-y_i\mathbf{w}^\top\boldsymbol{x}_i$$

感知机的学习算法

- 求解最优化问题：

$$\min_{\mathbf{w}} J(\mathbf{w}) = \min_{\mathbf{w}} \sum_{\boldsymbol{x}_i \in \boldsymbol{M}} -y_i \mathbf{w}^\top \boldsymbol{x}_i$$

- 梯度下降法(批量梯度下降法)：
  - ✓ 沿着准则函数的负梯度方向修正权向量**w**；
  - ✓ 准则函数对**w**的梯度为：

$$\nabla J(\mathbf{w}) = \sum_{\boldsymbol{x}_i \in \boldsymbol{M}} -y_i \boldsymbol{x}_i$$

  - ✓ 权向量更新方程为：

$$\mathbf{w} = \mathbf{w} - \eta \nabla J(\mathbf{w}) = \mathbf{w} + \eta \sum_{\boldsymbol{x}_i \in \boldsymbol{M}} y_i \boldsymbol{x}_i$$

其中，η为学习步长或学习速率，需要提前设定。

感知机的学习算法－批量梯度下降法

- 算法流程：(批量梯度下降)

① 初始随机选取一个超平面，$\mathbf{w}$;

② 令集合$\boldsymbol{M} = \emptyset$；依次对训练集中每一个样本点$(\boldsymbol{x}_i, y_i)$进行如下处理：如果$y_i \mathbf{w}^\top \boldsymbol{x}_i \leq 0$，则将$(\boldsymbol{x}_i, y_i)$加入$\boldsymbol{M}$中；

③ 如果集合$\boldsymbol{M}$不为空，进行模型更新：

$$\mathbf{w} = \mathbf{w} + \eta \sum_{\boldsymbol{x}_i \in \boldsymbol{M}} y_i \boldsymbol{x}_i$$

然后返回到步骤②；如果集合$\boldsymbol{M}$为空，则终止程序。

- 算法终止时找到的$\mathbf{w}$满足：

$$\text{对于} i = 1, \cdots N\text{，均有} y_i \mathbf{w}^\top \boldsymbol{x}_i > 0$$

- 所获的感知机模型为：$f(\boldsymbol{x}) = \mathrm{sign}(\mathbf{w}^\top \boldsymbol{x})$

感知机的学习算法

- 批量梯度下降法：
  - ✓ 每次更新权向量**w**时，需要首先遍历一遍训练数据集，计算被当前的权向量**w**误分类的样本的集合$\boldsymbol{M}$，利用$\boldsymbol{M}$中的样本对权向量**w**进行一次更新。
  - ✓ 需要很多次遍历训练数据集，计算量大。
- 随机梯度下降法：
  - ✓ 不是在每次更新时将所有被错分类的样本都找出来用于修正权向量，而是每次处理一 个样本，如果发现分类错误，就用这一个被错分类的样本来修正权向量。
  - ✓ 每次更新仅需要一个错分类的样本，计算量大大减小。

感知机的学习算法 – 随机梯度下降法

- 算法流程：(随机梯度下降)

① 初始随机选取一个超平面，$\mathbf{w} = \mathbf{0}$;

② 依次对训练数据集中每一个样本点$\boldsymbol{x}_i$进行如下处理：

如果$y_i \mathbf{w}^\top \boldsymbol{x}_i \leq 0$，则进行模型更新：

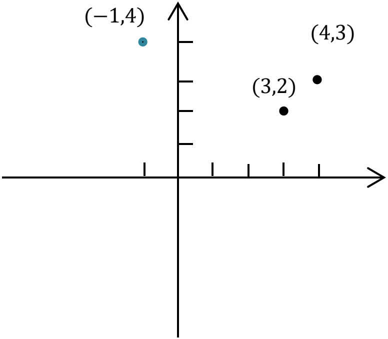
$$\mathbf{w} = \mathbf{w} + \eta\, y_i \boldsymbol{x}_i$$

如果$y_i \mathbf{w}^\top \boldsymbol{x}_i > 0$，则保持$\mathbf{w}$不变，返回步骤②，

直到训练数据集中所有的样本都被正确分类。

感知机的学习算法-举例

- 例子：给定一个训练数据集，如图所示，其中正例样本为$\boldsymbol{x}_1 = (3,2)$，$\boldsymbol{x}_2 = (4,3)$，负例样本为$\boldsymbol{x}_3 = (-1,4)$，请用**批量梯度下降法**实现的感知机学习算法找到一个分类超平面。



感知机的学习算法-例子

解：①对训练数据集进行拓展得到：

$$\boldsymbol{x}_1 = (3,2,1), y_1 =+ 1;$$
$$\boldsymbol{x}_2 = (4,3,1), y_2 =+ 1;$$
$$\boldsymbol{x}_3 = (-1,4,1), y_3 =- 1;$$

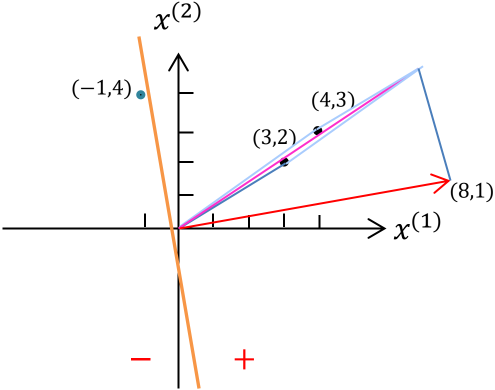
②利用批量梯度下降法求解：

a) 初始化$\mathbf{w} = \mathbf{0}$； 取学习步长$\eta = 1$；
b) 对于$i = 1,2,3$， $y_i \mathbf{w}^\top \boldsymbol{x}_i = 0$， 所以$\boldsymbol{M} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$；
c) 修正权向量：
$$\mathbf{w} = \mathbf{w} + \eta \sum_{\boldsymbol{x}_i \in \boldsymbol{M}} y_i \boldsymbol{x}_i = \boldsymbol{x}_1 + \boldsymbol{x}_2 - \boldsymbol{x}_3 = (8,1,1)$$
d) 对于$i = 1,2,3$， $y_i \mathbf{w}^\top \boldsymbol{x}_i > 0$， 所以$\boldsymbol{M} = \emptyset$， 终止程序。

最后得到的感知机模型为:$f(\boldsymbol{x}) = \text{sign}(8x^{(1)} + x^{(2)} + 1)$，

其中$\boldsymbol{x} = (x^{(1)}, x^{(2)})$为未扩充的实例。

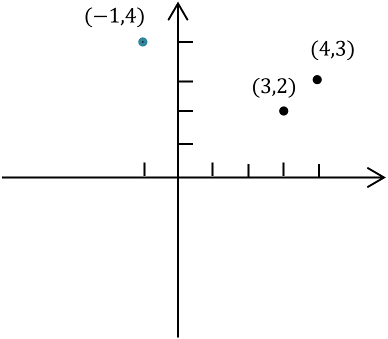
感知机的学习算法-例子



$$8x^{(1)} + x^{(2)} + 1 = 0$$

感知机的学习算法-举例

- 例子：给定一个训练数据集，如图所示，其中正例样本为$\boldsymbol{x}_1=(3,2)$，$\boldsymbol{x}_2=(4,3)$，负例样本为$\boldsymbol{x}_3=(-1,4)$，请用**随机梯度下降法**实现的感知机学习算法找到一个分类超平面。



感知机的学习算法-例子

解：①对训练数据集进行规范化得到：

$$\boldsymbol{x}_1 = (3,2,1), y_1 =+1;$$
$$\boldsymbol{x}_2 = (4,3,1), y_2 =+1;$$
$$\boldsymbol{x}_3 = (-1,4,1), y_3 =-1;$$

②利用随机梯度下降法求解：

a) 初始化$\mathbf{w} = \mathbf{0}$； 取学习步长$\eta = 1$；

b) 对于$\boldsymbol{x}_1$，因为$y_1\mathbf{w}^{\mathsf{T}}\boldsymbol{x}_1 = 0$，所以修正权向量：

$\mathbf{w} = \mathbf{w} + \eta y_1 \boldsymbol{x}_1 = \boldsymbol{x}_1 = (3,2,1)$

对于$\boldsymbol{x}_2$，因为$y_2\mathbf{w}^{\mathsf{T}}\boldsymbol{x}_2 > 0$，无需修正权向量；

对于$\boldsymbol{x}_3$，因为$y_3\mathbf{w}^{\mathsf{T}}\boldsymbol{x}_3 < 0$，需要修正权向量：

$\mathbf{w} = \mathbf{w} + \eta y_3 \boldsymbol{x}_3 = \boldsymbol{x}_1 - \boldsymbol{x}_3 = (4,-2,0)$

对于$\boldsymbol{x}_1$，$y_1\mathbf{w}^{\mathsf{T}}\boldsymbol{x}_1 > 0$，无需修正；对于$\boldsymbol{x}_2$，无需修正；

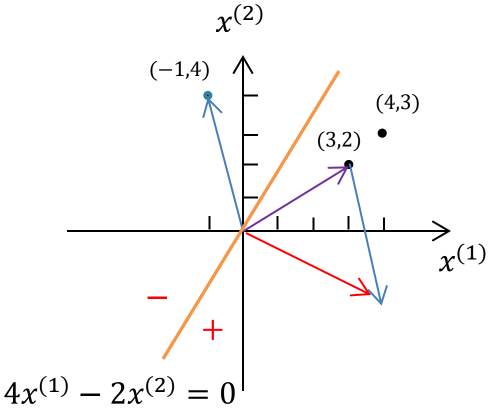
对于$\boldsymbol{x}_3$，无需修正；算法终止。

感知机的学习算法-例子

- 使用随机梯度法得到的感知机模型为:

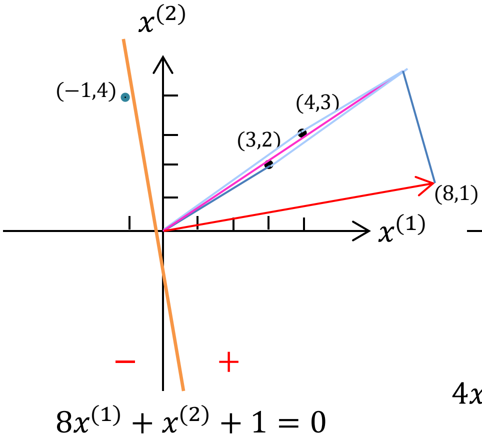
$$f(\boldsymbol{x}) = \mathrm{sign}(4x^{(1)} - 2x^{(2)})$$

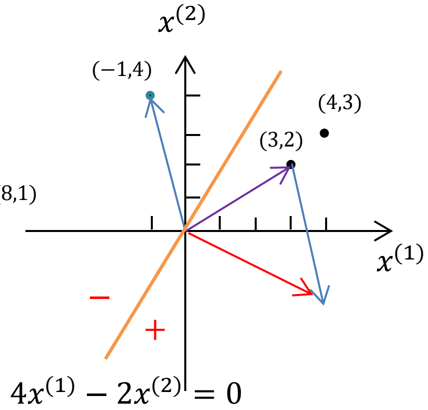
其中$\boldsymbol{x} = (x^{(1)}, x^{(2)})$为未扩充的实例。



感知机的学习算法-比较

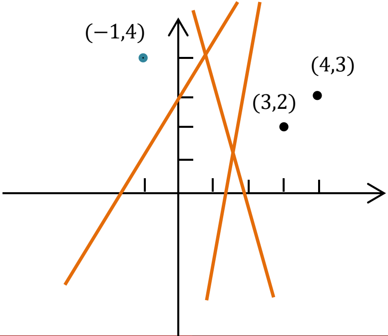
- 观察发现：两种优化方法均能得到一个将所有训练样本正确分类的超平面，只是所获得的超平面不同。





感知机的学习算法

- 事实上，当取不同的初始值、或者对训练样本采用不同的处理顺序、或者采用了不同的学习步长，算法最终求得的**w**可能就不相同。所以，感知机求得的**w**不是唯一的。
- 能将一个线性可分的训练数据集完全分开的超平面有很多，一个自然的问题是：众多的分割超平面中，哪个是最优的？



感知机难以回答这个问题，对于感知机而言，这些超平面都是最优的，因为它们都能使准则函数取得最小值0。

感知机算法的收敛性

- 现在证明：对于任意一个线性可分的二分类训练数据集，感知机算法经过有限次迭代更新后，可以得到一个将该数据集中的两类样本完全正确划分的超平面。
- 定理：设训练数据集$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots\cdots (\boldsymbol{x}_N, y_N)\}$是线性可分的，其中$\boldsymbol{x}_i \in \mathbb{R}^d$，$y_i \in \{-1, +1\}$，$i = 1, \cdots N$，则

① 存在满足$||\mathbf{w}^*|| = 1$的超平面$\mathbf{w}^{*\top}\boldsymbol{x} = 0$能将训练数据集完全正确分开，且存在一个$\gamma > 0$，对于所有$i = 1, \cdots N$，$y_i \mathbf{w}^{*\top} \boldsymbol{x}_i \geq \gamma$。

② 令$\max\limits_i ||\boldsymbol{x}_i|| = R$，则基于随机梯度下降更新的感知机算法在训练数据集上的误分类次数k满足不等式：

$$k \leq \frac{R^2}{\gamma^2}$$

定理证明

- 证明①：因为训练数据集$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots\cdots (\boldsymbol{x}_N, y_N)\}$是线性可分的，所以一定存在一个超平面$\mathbf{w}^{*\top}\boldsymbol{x} = 0$能将训练数据集完全正确分开，且$||\mathbf{w}^*|| = 1$，因此对于所有$i = 1, \cdots N$，有$y_i\mathbf{w}^{*\top}\boldsymbol{x}_i > 0$，取
$$\gamma = \min_i y_i\mathbf{w}^{*\top}\boldsymbol{x}_i$$
则对于所有$i = 1, \cdots N$，$y_i\mathbf{w}^{*\top}\boldsymbol{x}_i \geq \gamma > 0$。

- 证明②：设第k次修正后的权向量为$\mathbf{w}_k$，则第$k-1$次修正后的权向量为$\mathbf{w}_{k-1}$；设第k次修正使用的样本为$(\boldsymbol{x}_{(k)}, y_{(k)})$，因此可以得到
$$y_{(k)}\mathbf{w}_{k-1}{}^{\top}\boldsymbol{x}_{(k)} \leq 0 \qquad (1)$$
$$\mathbf{w}_k = \mathbf{w}_{k-1} + \eta y_{(k)}\boldsymbol{x}_{(k)} \qquad (2)$$

定理证明

$$\mathbf{w}_k = \mathbf{w}_{k-1} + \eta y_{(k)} \boldsymbol{x}_{(\boldsymbol{k})} \quad (2)$$

由公式(2)可得:

$$\mathbf{w}^{*\top}\mathbf{w}_k = \mathbf{w}^{*\top}\mathbf{w}_{k-1} + \eta y_{(k)} \mathbf{w}^{*\top} \boldsymbol{x}_{(\boldsymbol{k})} \geq \mathbf{w}^{*\top}\mathbf{w}_{k-1} + \eta\gamma$$

$$\geq \mathbf{w}^{*\top}\mathbf{w}_{k-2} + 2\eta\gamma \geq \cdots\cdots \geq \mathbf{w}^{*\top}\mathbf{w}_0 + k\eta\gamma$$

因为$\mathbf{w}_0 = \mathbf{0}$，所以得到$\mathbf{w}^{*\top}\mathbf{w}_k \geq k\eta\gamma$

根据柯西-施瓦茨不等式(Cauchy-Schwarz):

$$\mathbf{w}^{*\top}\mathbf{w}_k \leq ||\mathbf{w}^*|| \cdot ||\mathbf{w}_k|| = ||\mathbf{w}_k||$$

所以得到：$k\eta\gamma \leq ||\mathbf{w}_k||$

接下来需要求$||\mathbf{w}_k||$的上界。

定理证明

$$\mathbf{w}_k = \mathbf{w}_{k-1} + \eta y_{(k)} \boldsymbol{x}_{(\boldsymbol{k})} \quad (2)$$

由公式(2)可得:

$$\begin{aligned}
||\mathbf{w}_k||^2 &= ||\mathbf{w}_{k-1} + \eta y_{(k)} \boldsymbol{x}_{(\boldsymbol{k})}||^2 \\
&= ||\mathbf{w}_{k-1}||^2 + \eta^2 ||\boldsymbol{x}_{(\boldsymbol{k})}||^2 + 2\eta y_{(k)} \mathbf{w}_{k-1}{}^{\top} \boldsymbol{x}_{(\boldsymbol{k})} \\
&\le ||\mathbf{w}_{k-1}||^2 + \eta^2 ||\boldsymbol{x}_{(\boldsymbol{k})}||^2 \le ||\mathbf{w}_{k-1}||^2 + \eta^2 R^2
\end{aligned}$$

递归下去:

$$||\mathbf{w}_k||^2 \le ||\mathbf{w}_{k-1}||^2 + \eta^2 R^2 \le ||\mathbf{w}_{k-2}||^2 + 2\eta^2 R^2 \le \cdots\cdots$$

$$\le ||\mathbf{w}_0||^2 + k\eta^2 R^2 \le k\eta^2 R^2 \quad (\mathbf{w}_0 = 0)$$

所以得到: $||\mathbf{w}_k|| \le \sqrt{k}\eta R$

联合 $k\eta\gamma \le ||\mathbf{w}_k||$, 从而得到:

$$k \le \frac{R^2}{\gamma^2}$$

小结

- 感知机模型是一个适用于二分类的线性分类模型：

$$f(\boldsymbol{x}) = \mathrm{sign}(\mathbf{w}^{\mathsf{T}}\boldsymbol{x} + b)$$

- 感知机算法是基于梯度下降法对准则函数寻优的一个最优化算法，其两种权向量更新公式为：

  批量梯度下降法： $\mathbf{w} = \mathbf{w} + \eta \sum_{\boldsymbol{x}_i \in \boldsymbol{M}} y_i \boldsymbol{x}_i$

  随机梯度下降法： $\mathbf{w} = \mathbf{w} + \eta y_i \boldsymbol{x}_i$

重点

- 当训练数据集线性可分时，感知机算法存在无穷多的解，这些解的不同是因为算法采用了**不同的初始值**或对训练样本采用了**不同的处理顺序或**采用了**不同的学习步长**。

小结

- 当训练数据集线性可分时，感知机算法是收敛的，且在训练数据集上的误分类次数k满足：

$$k \leq \frac{R^2}{\gamma^2}$$

难点

- 感知机只能用于求解线性可分的问题，对线性不可分问题，感知机算法不能收敛。