统计学习方法

第二章 感知机（Perceptron）

王长海

2018.10.19

第二章 感知机

2.1 感知机模型

- 定义 2.1：
  - 输入空间：$\chi \subseteq \mathbb{R}^{n}$，输出空间：$\mathcal{Y} = \{+1, \boxed{-1}\}$
  - 模型参数：w，b

$$f(x) = sign(w \cdot x + b)$$ ，其中

$$sign(x) = \begin{cases} +1, & \mathrm{x} \geq 0 \\ -1, & \mathrm{x} < 0 \end{cases}$$

2.1 感知机模型

- 几何解释：

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

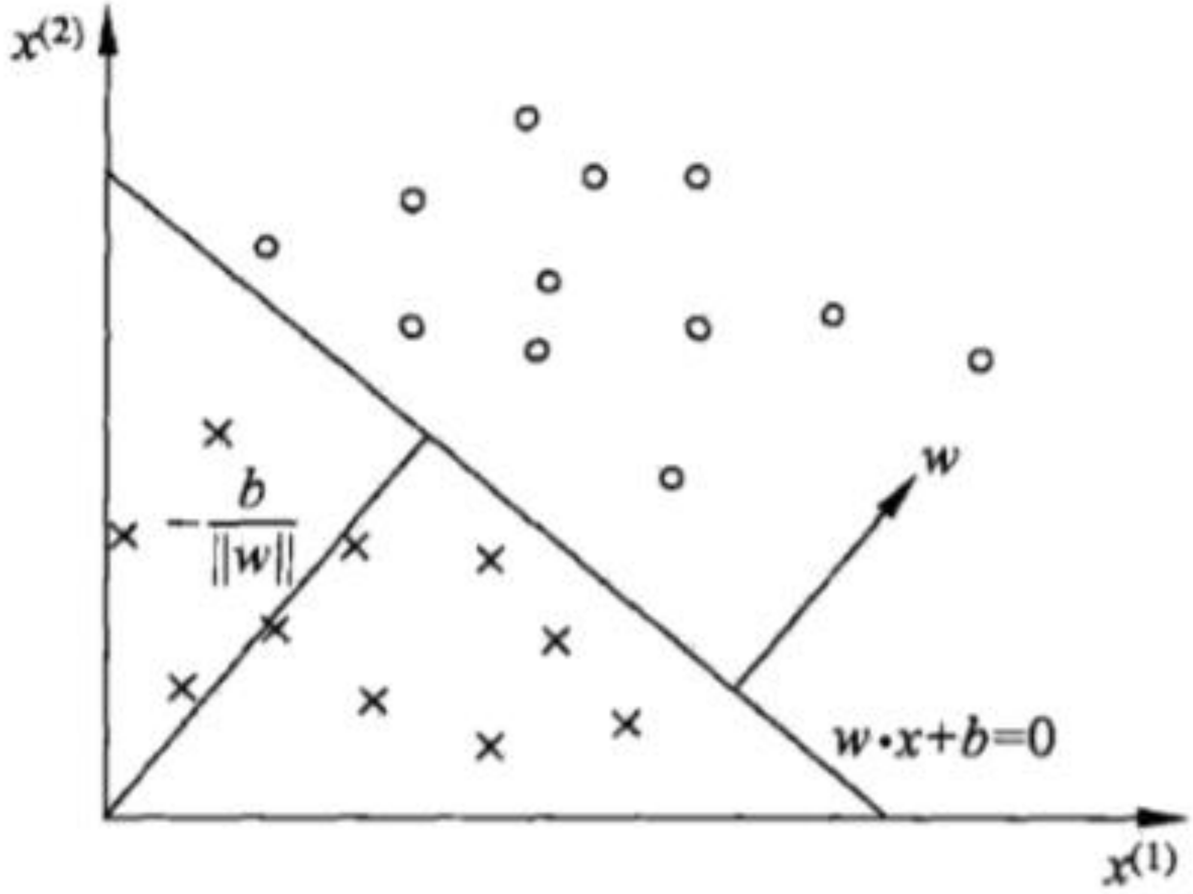


图 2.1 感知机模型

第二章 感知机

2.2 感知机学习策略

- 数据集线性可分性：
  - 数据集：$T=\{(x_1,y_1),(x_2,y_2),\cdots(x_n,y_n)\}$
  - 超平面： $S\rightarrow w\cdot x+b=0$

$$\mathrm{y}_i=+1,\ \ w\cdot x_i+b>0$$

$$\mathrm{y}_i=-1,\ \ w\cdot x_i+b<0$$

则数据集T为线性可分数据集

2.2 感知机学习策略

- 学习策略：

超平面： $S \to \boxed{w} \cdot x + \boxed{b} = 0$

a） 最小化误分类点总数

点到直线距离

b） 最小化误分类点到超平面的距离：

$$\frac{1}{\|w\|}|w \cdot x_i + b| \xrightarrow{\text{去绝对值符号}} -\frac{1}{\|w\|}y_i(w \cdot x_i + b)$$

2.2 感知机学习策略

- 学习策略：
  - 输入空间：$\chi \subseteq \mathbb{R}^{n}$，输出空间：$\mathcal{Y} = \{+1, \ -1\}$
  - 训练数据集：$T = \{(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)[x_i \epsilon \chi, y_i \epsilon \mathcal{Y}\}$
  - 误分类集合:$\mathrm{M} = \{(x_1, y_1), (x_2, y_2), \cdots (x_m, y_m)[x_j \epsilon \chi, y_j \epsilon \mathcal{Y}\}$
  - 损失函数：$L(w, b) = -\sum_{x_i \epsilon M} y_i (w \cdot x_i + b)$

最小化

第二章 感知机

2.3 感知机学习算法

- 学习算法：
  - $\min_{w,b} L(w,b) = -\sum_{x_i \epsilon M} y_i (w \cdot x_i + b)$
  - 算法：随机梯度下降算法（SGD）

$$\frac{\partial L}{\partial w} = -\sum_{x_i \epsilon M} y_i x_i \qquad \frac{\partial L}{\partial b} = -\sum_{x_i \epsilon M} y_i$$

误分类驱动：一次随机选取一个误分类点使其梯度下降

2.3 感知机学习算法

- 学习算法（原始形式）：
  - 输入：$T = \{(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)\}$，学习率$\eta (0 < \eta \leq 1)$
  - 输出：w, b；模型$f(x) = sign(w \cdot x + b)$
  - Step 1:选取初值w_0, b_0
  - Step 2:在训练集中选取（x_i, y_i）
  - Step 3:如果$\mathrm{y}_i (w \cdot x_i + b) \leq 0$

$$w \leftarrow w + \eta y_i \mathrm{x}_i$$

$$b \leftarrow b + \eta y_i$$

  - Step 4:转到Step2，直至算法收敛。

解不唯一
增加约束？

线性不可分时，无法收敛

2.3 感知机学习算法

- 收敛性：
  - 设$\widehat{\mathrm{w}} = (w^T, b)^T$，$\hat{x} = (x^T, 1)^T$，有$\widehat{\mathrm{w}} \cdot \hat{x} = \mathrm{w} \cdot x + b$
  - Novikoff定理：
    - 若$T = \{(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)\}$线性可分，则

（1）存在满足$\|\widehat{\mathrm{w}}_{\mathrm{opt}}\| = 1$的超平面$\widehat{\mathrm{w}}_{\mathrm{opt}} \cdot \hat{x} = 0$将数据集完全正确分开，且存在$\gamma > 0$，对所有$i = 1,2, \cdots, N$

$$y_i(\widehat{\mathrm{w}}_{\mathrm{opt}} \cdot \hat{x}_i) = y_i(w_{opt} \cdot x_i + b_{opt}) \geq \gamma$$

（2）令$R = \max_{1 \leq i \leq N} \|\hat{x}_i\|$，感知器算法在训练集上误分类次数k满足

$$\mathrm{k} \leq \left(\frac{R}{\gamma}\right)^2$$

2.3 感知机学习算法

- 定理证明：

（1）存在$\gamma = \min\limits_{i}\{y_i(\hat{w}_{opt}\cdot\hat{x}_i)\}$

使得$y_i(\hat{w}_{opt}\cdot\hat{x}_i) = y_i(w_{opt}\cdot x_i + b_{opt}) \geq \gamma$

（2）设$\hat{w}_0 = 0$，$\hat{w}_{k-1}$是第k个误分类实例之前的扩充权重向量，则不难推导

$$\hat{w}_k \leftarrow \hat{w}_{k-1} + \eta y_i \hat{x}_i$$

若$\hat{w}_k \cdot \hat{w}_{opt} \geq k\eta\gamma$ 及 $\|\hat{w}_k\| \leq \sqrt{k}\eta R$ 成立，则有

$$k\eta\gamma \leq \hat{w}_k \cdot \hat{w}_{opt} \leq \|\hat{w}_k\|\|\hat{w}_{opt}\| = \|\hat{w}_k\| \leq \sqrt{k}\eta R$$

$$k \leq \left(\frac{R}{\gamma}\right)^2$$

$\boxed{\|\hat{w}_{opt}\| = 1}$

2.3 感知机学习算法

- 定理证明：（续）

$$\boxed{y_i(\hat{w}_{opt}\cdot\hat{x}_i)\geq\gamma}$$

$$\hat{w}_k\cdot\hat{w}_{opt}\geq k\eta\gamma$$

$$\hat{w}_k\cdot\hat{w}_{opt}=\hat{w}_{k-1}\cdot\hat{w}_{opt}+\eta y_i\hat{w}_{opt}\hat{x}_i\geq\hat{w}_{k-1}\cdot\hat{w}_{opt}+\eta\gamma$$

递推不等式$\hat{w}_k\cdot\hat{w}_{opt}\geq\hat{w}_{k-1}\cdot\hat{w}_{opt}+\eta\gamma\geq\hat{w}_{k-2}\cdot\hat{w}_{opt}+2\eta\gamma\geq\cdots\geq k\eta\gamma$

$$\|\hat{w}_k\|\leq\sqrt{k}\eta R$$

两边平方得 $\|\hat{w}_k\|^2\leq k\eta^2R^2$ $\quad\|\hat{w}_k\|^2=\|\hat{w}_{k-1}\|^2+2\eta y_i\hat{w}_{k-1}\hat{x}_i+\eta^2\|x_i\|^2$

$$\leq\|\hat{w}_{k-1}\|^2+\eta^2\|x_i\|^2 \longrightarrow \boxed{R=\max_{1\leq i\leq N}\|\hat{x}_i\|}$$

$$\leq\|\hat{w}_{k-1}\|^2+\eta^2R^2\leq k\eta^2R^2$$

2.3 感知机学习算法

- 学习算法(对偶形式)：

$$w \leftarrow w + \eta y_i \mathrm{x}_i$$

$$b \leftarrow b + \eta y_i$$

$$\alpha_i = \mathrm{n}_i \eta$$

$$\Longrightarrow$$

$$w = \sum_{i=1}^{N} \boxed{\alpha_i} y_i \mathrm{x}_i$$

$$b = \sum_{i=1}^{N} \boxed{\alpha_i} y_i$$

2.3 感知机学习算法

- 学习算法(对偶形式):
  - 输入:$T=\{(x_1,y_1),(x_2,y_2),\cdots(x_n,y_n)\}$, 学习率$\eta(0<\eta\leq 1)$
  - 输出:α,b; 模型$f(x)=sign(\sum_{j=1}^{N}\alpha_j y_j \underline{x_j\cdot x}+b)$
  - Step 1:$\alpha\leftarrow 0,b\leftarrow 0$
  - Step 2:在训练集中选取 (x_i,y_i)
  - Step 3:如果$y_i\left(\sum_{j=1}^{N}\alpha_j y_j x_j\cdot x_i+b\right)\leq 0$

$$\alpha_i\leftarrow\alpha_i+\eta$$

$$b\leftarrow b+\eta y_i$$

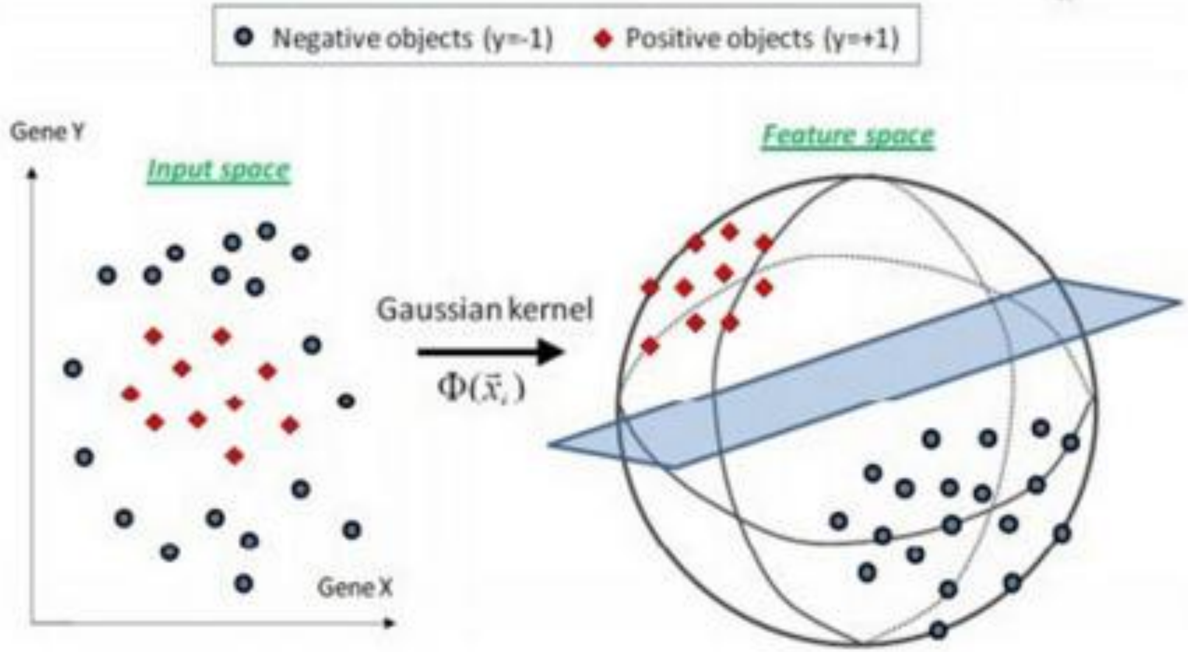
  - Step 4:转到Step2,直至算法收敛。

内积形式可以用
Gram矩阵预处理

2.3 感知机学习算法

可以使用核函数

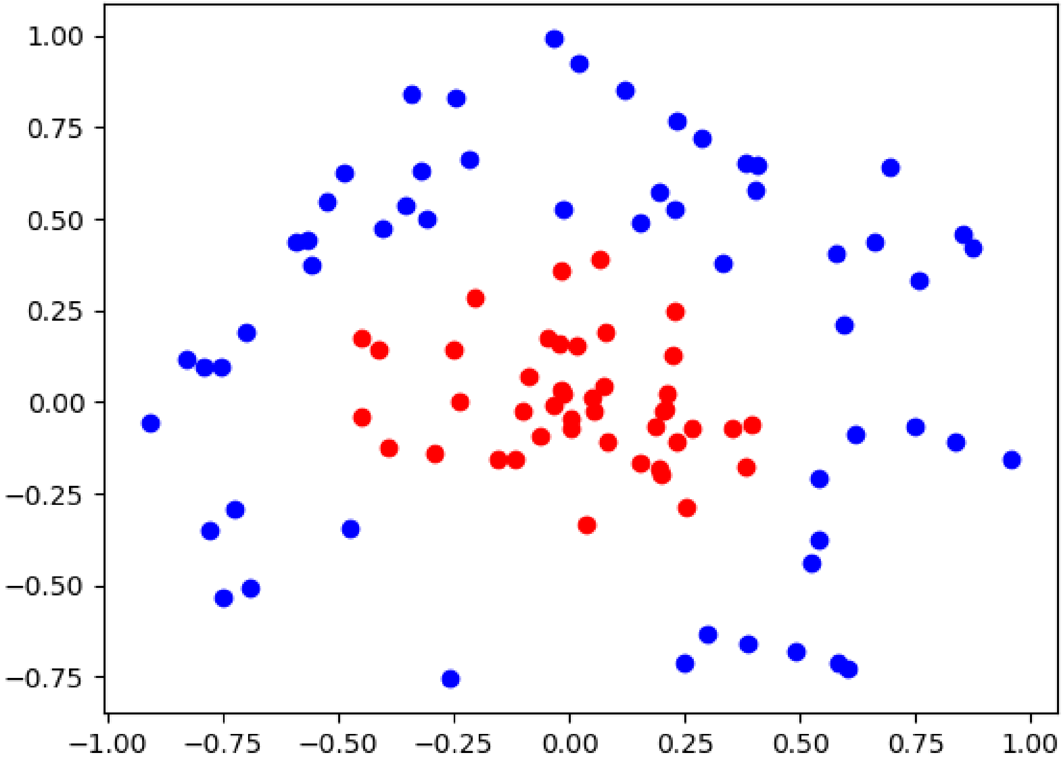
- 核函数：
  - $f(x) = sign(\sum_{j=1}^{N} \alpha_j y_j \underline{x_j \cdot x} + b)$
  - 径向基（高斯）核函数：
  - $K(x, z) = \exp\left(\frac{-\|x-z\|^2}{2\sigma^2}\right)$



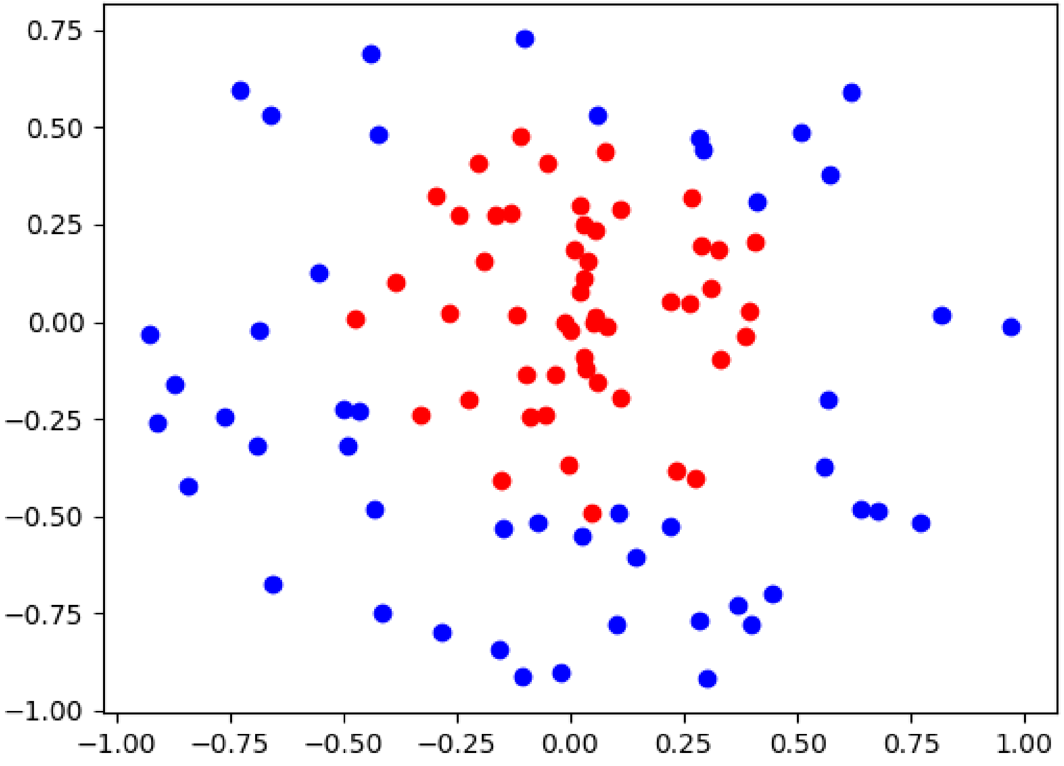
第二章 感知机

2.4 感知机实验

- 数据集：



训练集



测试集

2.4 感知机实验

- 代码实现（核函数形式）：

```
def train(dataArr,labelArr,iter_num=500,learning_rate=0.01):

        w=0.0
        b=0.0
        dataMat=np.array(dataArr)
        labelMat=np.array(labelArr)
        m=np.shape(dataMat)[0]
        K=np.mat(np.zeros((m,m)))
        alpha=[0 for i in range(m)]

        for i in range(m):
                K[:,i]=kernelTrans(np.mat(dataMat),np.mat(dataMat[i,:]),('rbf',1,3))
        for idx in range(iter_num):
                tmp=0
                i = random.randint(0,m-1)
                yi=labelMat[i]
                for j in range(m):
                        tmp+=alpha[j]*labelMat[j]*K[i,j]

                tmp+=b
                if(yi*tmp<=0):
                        alpha[i]=alpha[i]+learning_rate
                        b=b+learning_rate*yi

        return alpha,b,K
```

预处理

直接调用K[i,j]

2.4 感知机实验

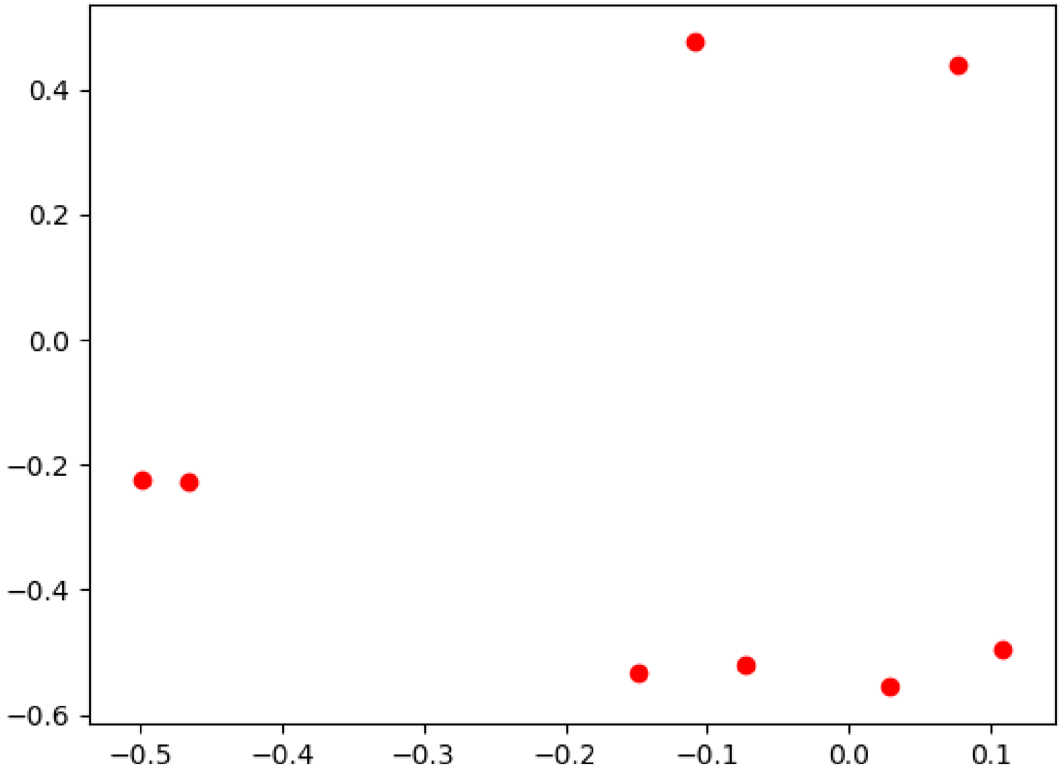
- 测试结果：

同数据感知机测试误差：

```
the test error rate is: 0.080000
```

同数据支持向量机测试误差：

```
the test error rate is: 0.040000
```



错分数据

第二章 感知机

2.1 感知机模型

2.2 感知机学习策略

2.3 感知机学习算法

2.4 感知机实验

2.5 总结

2.5 总结

- 感知器的模型及定义
- 感知器的学习策略
- 梯度下降算法
- 感知器算法的原始形式及对偶形式