北京航空航天大學
COLLEGE OF SOFTWARE BEIHANG UNIVERSITY 软件学院

人工智能

第6讲：机器学习-有监督学习II

张晶

2023年春季

- 参考教材： 吴飞，《人工智能导论：模型与算法》，高等教育出版社
- 在线课程： https://www.icourse163.org/course/ZJU-1003377027?from=searchPage

北京航空航天大學 软件学院

提纲

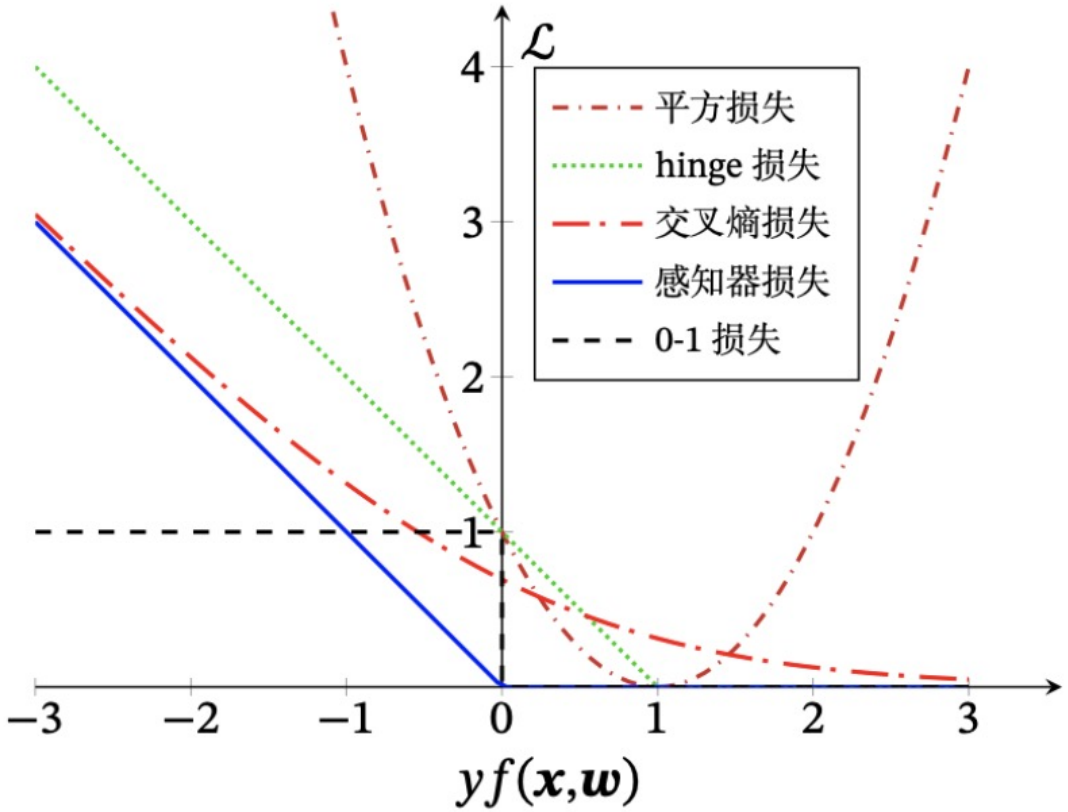
北京航空航天大學 软件学院

提纲

北京航空航天大學
COLLEGE OF SOFTWARE BEIHANG UNIVERSITY 软件学院

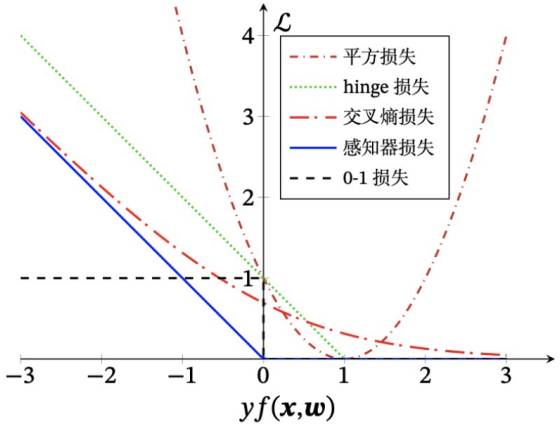
线性分类模型

- 线性回归：$\frac{1}{N}\sum(y_i - f(x_i))^2$
- 对数几率回归：$-logP(y_i|x_i)$
- 线性判别分析
- 感知器
- 支持向量机
- …



二分类问题中不同损失函数的对比
（横轴表示$yf(\boldsymbol{x},\boldsymbol{w})$，纵轴表示损失）

线性分类模型



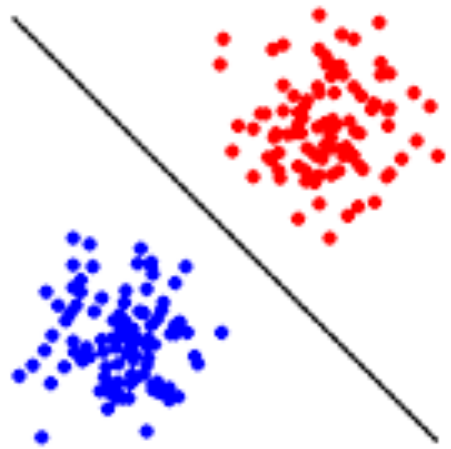
线性模型	激活函数	损失/目标函数	损失/目标函数定义	优化方法
		0-1损失	$\begin{cases}1, f(\boldsymbol{x}_i) \neq y_i \\ 0, f(\boldsymbol{x}_i) = y_i\end{cases}$	
线性回归	-	平方损失	$(y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2$	最小二乘、梯度下降
对数几率回归	$sigmoid(\boldsymbol{w}^T\boldsymbol{x})$	二值交叉熵损失	$-y_i \log\sigma(\boldsymbol{w}^T\boldsymbol{x}_i) + (1-y_i)\log(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}_i))$	梯度下降
Softmax分类	$softmax(\boldsymbol{W}^T\boldsymbol{x})$	交叉熵损失	$-y_i \log softmax(\boldsymbol{w}^T\boldsymbol{x}_i)$	梯度下降
线性判别分析	-	Fisher准则	$\frac{\|m_2 - m_1\|_2^2}{s_1^2 + s_2^2}$	广义特征值分解
感知器	$sign(\boldsymbol{w}^T\boldsymbol{x})$	感知器准则	$\max(0, -y_i\boldsymbol{w}^T\boldsymbol{x}_i)$	随机梯度下降
支持向量机	$sign(\boldsymbol{w}^T\boldsymbol{x})$	Hinge损失	$\max(0,1 - y_i\boldsymbol{w}^T\boldsymbol{x}_i)$	二次规划、SMO等

线性分类模型

- 二分类问题定义：
  - 数据：假设训练集$\mathrm{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$，样本$\boldsymbol{x}_i \in \mathbb{R}^d$的类别标签为$y_i$。其中，$y_i$的取值范围是{0,1}，即一共有2类样本。
  - 模型：线性分类器

$$\hat{y}(\boldsymbol{x}) = f(\boldsymbol{w}^T\boldsymbol{x} + b)$$

  - 目标：正确分类问题



线性分类模型

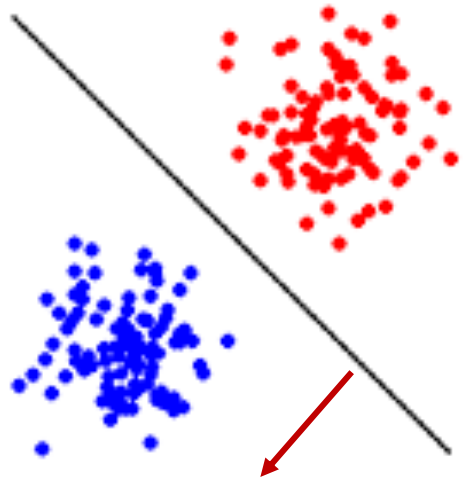
- 线性分类模型的一般形式

$$\hat{y}(\boldsymbol{x}) = f(\boldsymbol{w}^T\boldsymbol{x} + b)，\ 其中 f(z) = \begin{cases} 1, & z > 0 \\ 0, & z < 0 \end{cases}$$

  - $f(\boldsymbol{w}^T\boldsymbol{x} + b) > 0.5 \Rightarrow \mathcal{C}_1$
  - $f(\boldsymbol{w}^T\boldsymbol{x} + b) < 0.5 \Rightarrow \mathcal{C}_2$
  - f为激活函数

可以通过定义和求解线性判别函数实现

$$g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$$



决策边界：$f(\boldsymbol{w}^T\boldsymbol{x} + b) = 0.5$

北京航空航天大學 软件学院

线性分类-最小二乘法

- 回顾：线性回归模型参数求取：$\hat{y}(\boldsymbol{x}_i) = \boldsymbol{w}^T\boldsymbol{x}_i + b\ (1 \le i \le n)$
  - 记在当前参数下第i个训练样本x_i的预测值为$\hat{y}_i = \boldsymbol{w}^T\boldsymbol{x}_i + b$
  - x_i的标注值（实际值）y_i与预测值$\hat{y}_i$之差记为$(y_i - \hat{y}_i)^2$
  - 训练集中n个样本所产生误差总和为：$L(\boldsymbol{w}, b) = \sum_{i=1}^{n}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b\ )^2$
- 将x_i的标注值（实际值）$\boldsymbol{y}_i$编码为K维独热向量:

$$\boldsymbol{y}_i = [1,0,\dots,0]^T \in \mathbb{R}^K$$

  - 即可得到基于最小二乘法的线性分类模型

$$L(\boldsymbol{w}, b) = \sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{w}^T\boldsymbol{x}_i - b\ )^2$$

线性分类-对数几率回归

- 回顾：对数几率回归

$$\hat{y}(\boldsymbol{x}) = f(\boldsymbol{w}^T\boldsymbol{x} + b) = sigmoid(\boldsymbol{w}^T\boldsymbol{x} + b) = \frac{1}{1 + e^{-(\boldsymbol{w}^T\boldsymbol{x}+b)}}$$

$$\text{其中} f(z)\begin{cases} \geq 0.5, & z \geq 0 \\ < 0.5, & z < 0 \end{cases}$$

  - $f(\boldsymbol{w}^T\boldsymbol{x} + b) > 0.5 \Rightarrow \mathcal{C}_1$
  - $f(\boldsymbol{w}^T\boldsymbol{x} + b) < 0.5 \Rightarrow \mathcal{C}_2$
  - f为sigmoid激活函数

线性分类-Fisher线性判别分析

- 线性判别分析(linear discriminant analysis， LDA)是一种基于监督学习的分类/降维方法，也称为Fisher线性判别分析（fisher's discriminant analysis，FDA) [Fisher 1936]。
- 对于一组具有标签信息的高维数据样本，LDA利用其类别信息，将其线性投影到一个低维空间上，在低维空间中同一类别样本尽可能靠近，不同类别样本尽可能彼此远离。

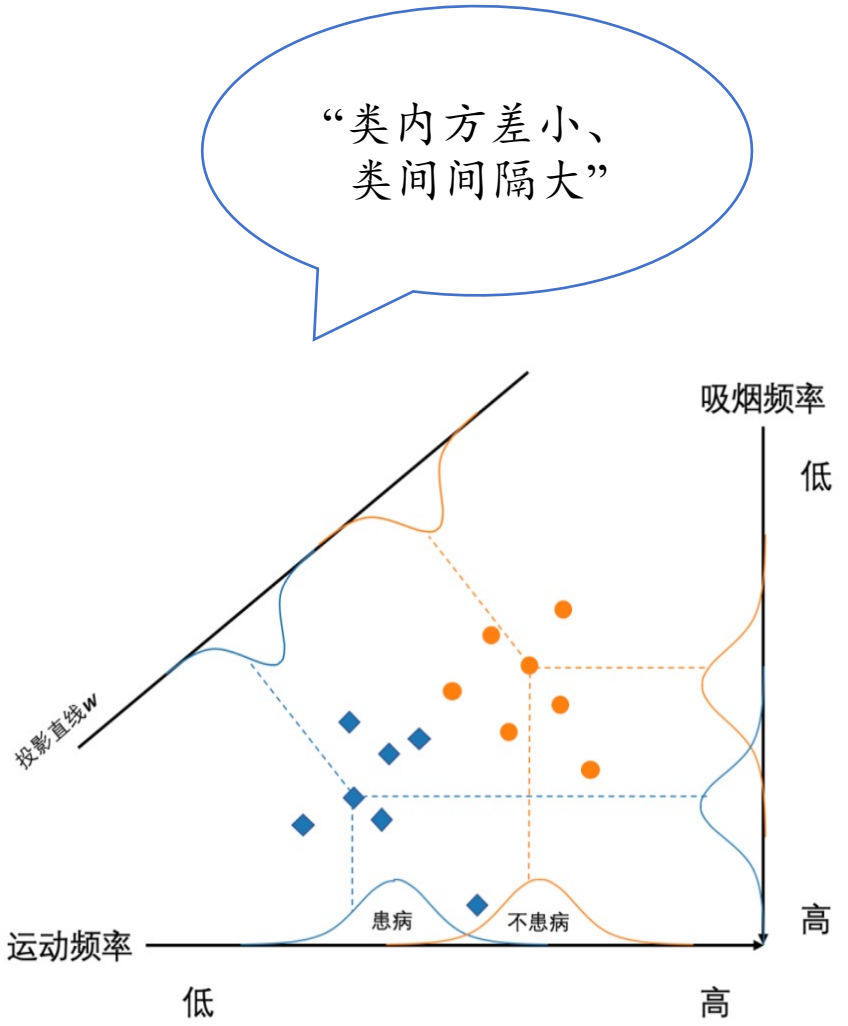


图4. 8 两个类别数据所对应的不同投影方式
君子和而不同、小人同而不和

北京航空航天大学 软件学院

Fisher线性判别分析：符号定义

- 假设样本集为$\mathrm{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$，样本$\boldsymbol{x}_i \in \mathbb{R}^d$的类别标签为$y_i$。其中，$y_i$的取值范围是$\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$，即一共有$K$类样本。
- 定义$\boldsymbol{X}$为所有样本构成集合、N_i为第i个类别所包含样本个数、X_i为第i类样本的集合、$\boldsymbol{m}$为所有样本的均值向量、$\boldsymbol{m}_i$为第i类样本的均值向量。Σ_i为第i类样本的散度矩阵，其定义为:

$$\Sigma_i = \sum_{\boldsymbol{x} \in \boldsymbol{X}_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^T$$

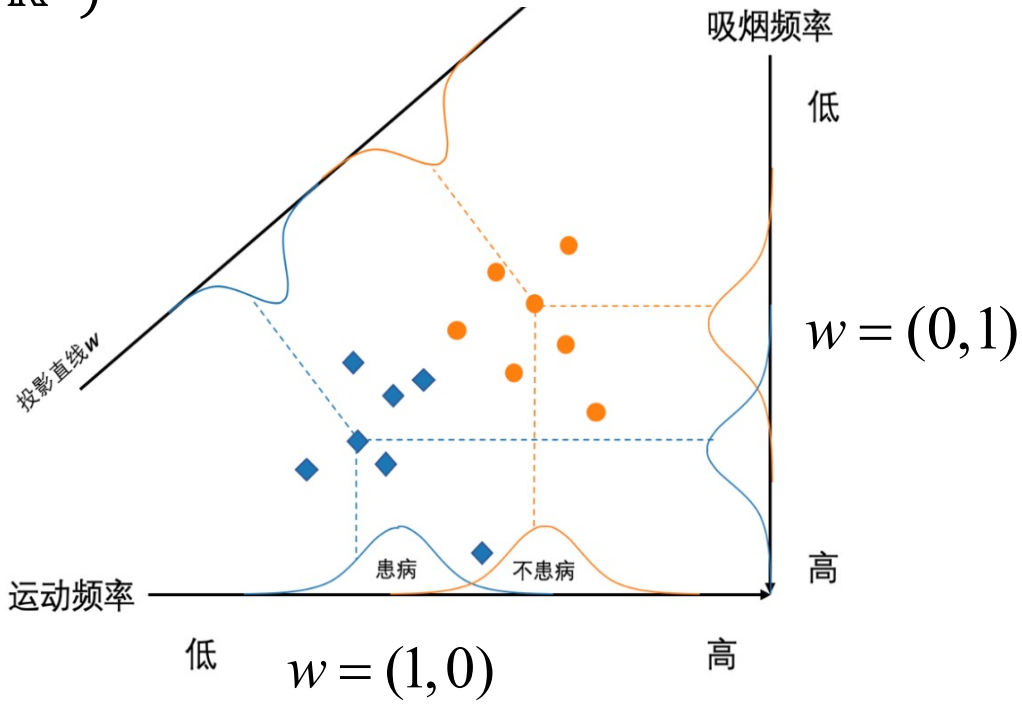
Fisher线性判别分析：二分类问题

- 先来看$K = 2$的情况，即二分类问题。在二分类问题中，训练样本归属于C_1或C_2两个类别，并通过如下的线性函数投影到一维空间上：

$$\hat{y}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} \ (\boldsymbol{w} \in \mathbb{R}^d)$$

数据点$(1,1),(2,2),(3,3),(4,4)...$

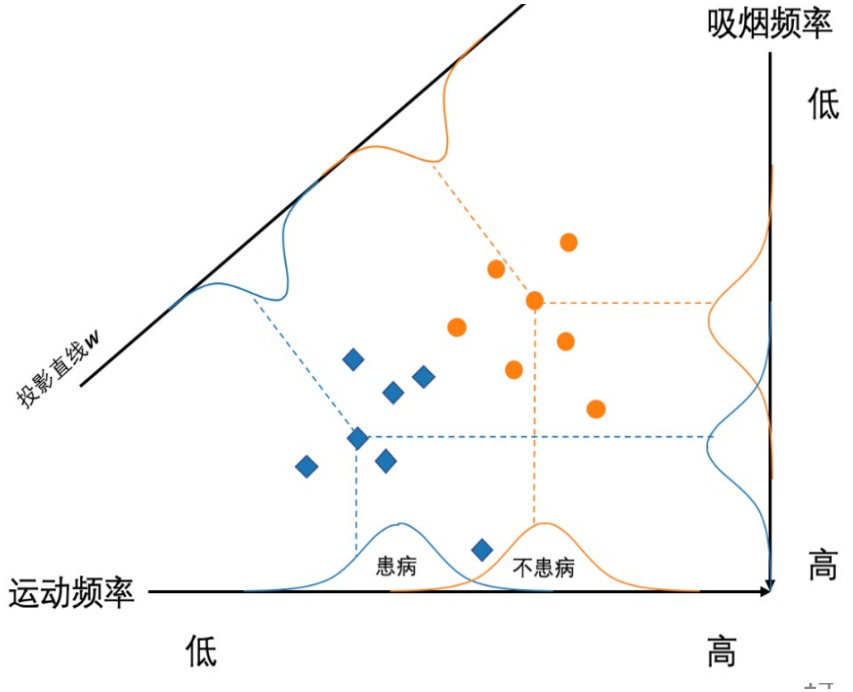
都会投影到同一个点。



Fisher线性判别分析：二分类问题

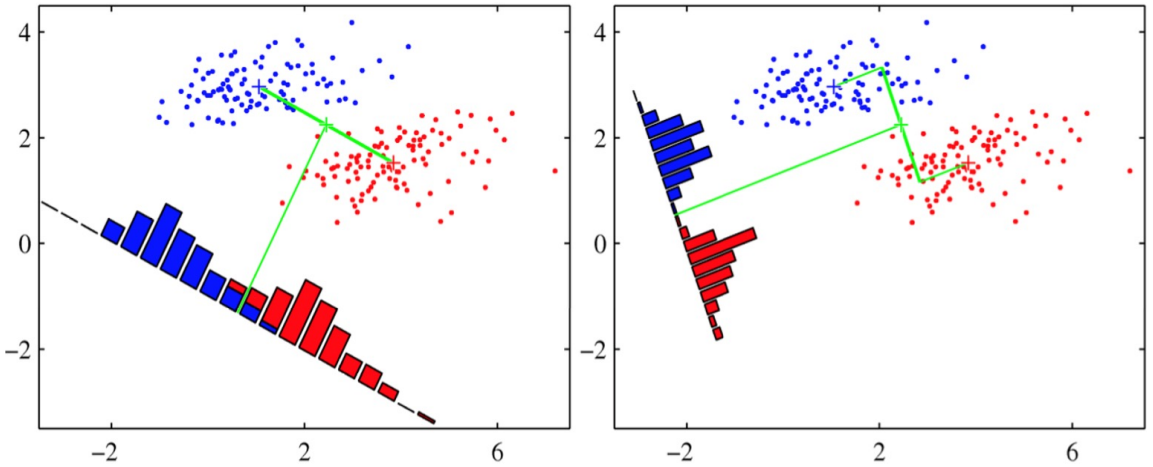
- 希望寻找一个投影方向$\boldsymbol{w}$，使得两个类别的数据在投影以后$\hat{y}(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$容易被分开

- 两个类各自的均值为
  - $\boldsymbol{m}_1 = \frac{1}{N_1}\sum_{\boldsymbol{x}\in C_1}\boldsymbol{x}$,
  - $\boldsymbol{m}_2 = \frac{1}{N_2}\sum_{\boldsymbol{x}\in C_2}\boldsymbol{x}$
  - 投影以后的均值为
    - $m_1 = \boldsymbol{w}^T\boldsymbol{m}_1$,
    - $m_2 = \boldsymbol{w}^T\boldsymbol{m}_2$



Fisher线性判别分析：二分类问题

- 怎样描述“分开”的程度？
- 最大化$\|m_2 - m_1\|_2^2$？问题？
  - 这个值可以无限大。
  - 如图所示，这个值不是越大越好。

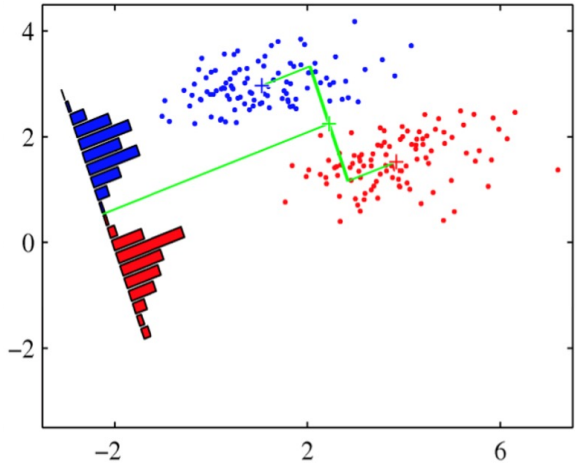


- Fisher准则
  - 在要求$\|m_2 - m_1\|_2^2$尽量大的同时，要求两类在投影以后尽量集中，或者不分散。怎么度量分散程度？

$$J(\boldsymbol{w}) = \frac{\|m_2 - m_1\|_2^2}{s_1^2 + s_2^2}$$

Fisher线性判别分析：二分类问题

- Fisher准则：在要求$\|m_2 - m_1\|_2^2$尽量大的同时，要求两类在投影以后尽量集中，或者不分散。怎么度量分散程度？



$$J(\boldsymbol{w}) = \frac{\|m_2 - m_1\|_2^2}{s_1^2 + s_2^2}$$

投影之后类别$\mathcal{C}_1$的散度矩阵s_1为：

$$s_1 = \sum_{\boldsymbol{x} \in \mathcal{C}_1} (\boldsymbol{w}^T \boldsymbol{x} - \boldsymbol{w}^T \boldsymbol{m}_1)^2 = \boldsymbol{w}^T \sum_{\boldsymbol{x} \in \mathcal{C}_1} [(\boldsymbol{x} - \boldsymbol{m}_1)(\boldsymbol{x} - \boldsymbol{m}_1)^T] \boldsymbol{w}$$

同理可得到投影之后类别$\mathcal{C}_2$的散度矩阵s_2。

Fisher线性判别分析：二分类问题

$$J(\boldsymbol{w}) = \frac{\|m_2 - m_1\|_2^2}{s_1^2 + s_2^2} = \frac{\|\boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)\|_2^2}{\boldsymbol{w}^T\Sigma_1\boldsymbol{w} + \boldsymbol{w}^T\Sigma_2\boldsymbol{w}} = \frac{\boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T\boldsymbol{w}}{\boldsymbol{w}^T(\Sigma_1 + \Sigma_2)\boldsymbol{w}} = \frac{\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}}$$

- 其中，$\boldsymbol{S}_b$称为类间散度矩阵(between-class scatter matrix)，即衡量两个类别均值点之间的“分离”程度，可定义如下:

$$\boldsymbol{S}_b = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T$$

- $\boldsymbol{S}_w$则称为类内散度矩阵(within-class scatter matrix)，即衡量每个类别中数据点的“分离”程度，可定义如下：

$$\boldsymbol{S}_w = \Sigma_1 + \Sigma_2$$

- 由于$J(\boldsymbol{w})$的分子和分母都是关于$\boldsymbol{w}$的二项式，因此最后的解只与$\boldsymbol{w}$的方向有关，与$\boldsymbol{w}$的长度无关，因此可令分母$\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w} = 1$，然后用拉格朗日乘子法来求解这个问题。

Fisher线性判别分析：二分类问题

对应拉格朗日函数为：

$$L(\boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w} - \lambda(\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} - 1)$$

- 解法1：对$\boldsymbol{w}$求偏导并使其求导结果为零，可得

$$\boldsymbol{S}_b \boldsymbol{w} = \lambda \boldsymbol{S}_w \boldsymbol{w}，\text{或} \boldsymbol{S}_w{}^{-1} \boldsymbol{S}_b \boldsymbol{w} = \lambda \boldsymbol{w}$$

由此可见，λ和$\boldsymbol{w}$分别是$\boldsymbol{S}_w{}^{-1} \boldsymbol{S}_b$的特征根和特征向量

对矩阵和向量求导规则参见：
https://www.sfu.ca/~haiyunc/notes/matrix_calculus.pdf
https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Fisher线性判别分析：二分类问题

对应拉格朗日函数为：

$$L(\boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w} - \lambda(\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} - 1)$$

- 解法2：根据$\boldsymbol{S}_b = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T$，以及矩阵乘法结合律：$(AB)C = A(BC)$，得到

$$\boldsymbol{S}_b \boldsymbol{w} = (\boldsymbol{m}_2 - \boldsymbol{m}_1)\underbrace{((\boldsymbol{m}_2 - \boldsymbol{m}_1)^T \boldsymbol{w})}_{\text{标量}}$$

因此，$\boldsymbol{S}_b \boldsymbol{w}$的方向恒为$\boldsymbol{m}_2 - \boldsymbol{m}_1$，不妨令

$$\lambda_w = (\boldsymbol{m}_2 - \boldsymbol{m}_1)^T \boldsymbol{w}$$

那么

$$\boldsymbol{S}_w{}^{-1} \boldsymbol{S}_b \boldsymbol{w} = \boldsymbol{S}_w{}^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1) \times \lambda_w = \lambda \boldsymbol{w}$$

Fisher线性判别分析：二分类问题

对应拉格朗日函数为：

$$L(\boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w} - \lambda(\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} - 1)$$

- 解法2：（接上页）

那么

$$\boldsymbol{S}_w{}^{-1} \boldsymbol{S}_b \boldsymbol{w} = \boldsymbol{S}_w{}^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1) \times \lambda_w = \lambda \boldsymbol{w}$$

由于对$\boldsymbol{w}$的放大和缩小操作不影响结果，因此可约去上式中的未知数λ和λ_w，得到：

$$\boldsymbol{w} = \boldsymbol{S}_w{}^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1)$$

也被称为Fisher线性判别（Fisher linear discrimination，简称FLD）。

Fisher线性判别分析：多分类问题

- 假设n个原始d维数据包含类别种类为K、每个原始数据被投影映射到低维空间中的维度为r。
- 令投影矩阵$\boldsymbol{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_r)$，可知$\boldsymbol{W}$是一个$d\times r$矩阵。于是，$\boldsymbol{W}^T\boldsymbol{m}_i$为第$i$类样本数据中心在低维空间的投影结果，$\boldsymbol{W}^T\Sigma_i\boldsymbol{W}$为第$i$类样本数据散度在低维空间的投影结果。
- 全局散度矩阵$\boldsymbol{S}_t=\boldsymbol{S}_w+\boldsymbol{S}_b=\sum_{i=1}^{K}(\boldsymbol{x}_i-\boldsymbol{m})(\boldsymbol{x}_i-\boldsymbol{m})^T$，其中$\boldsymbol{m}$为所有样本的均值向量。
- 类内散度矩阵$\boldsymbol{S}_w$重新定义如下：

$$\boldsymbol{S}_w=\sum_{i=1}^{K}\Sigma_i，\text{其中}\Sigma_i=\sum_{\boldsymbol{x}\in\mathcal{C}_i}(\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)^T$$

在上式中，$\boldsymbol{m}_i$是第i个类别中所包含样本数据的均值。

- 类间散度矩阵$\mathbf{S}_b$重新定义如下：

$$\boldsymbol{S}_b=\boldsymbol{S}_t-\boldsymbol{S}_w=\sum_{i=1}^{K}\frac{N_i}{N}(\boldsymbol{m}_i-\boldsymbol{m})(\boldsymbol{m}_i-\boldsymbol{m})^T$$

Fisher线性判别分析：多分类问题

- 将多类LDA映射投影方向的优化目标$J(\boldsymbol{W})$改为：

$$J(\boldsymbol{W}) = \max_{\boldsymbol{W}} \frac{tr(\boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W})}{tr(\boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W})}$$

其中，$\boldsymbol{W} \in \boldsymbol{R}^{n \times r}$，$tr(\cdot)$表示矩阵的迹。

- 可通过如下广义特征值问题进行求解：

$$\boldsymbol{S}_b \boldsymbol{W} = \lambda \boldsymbol{S}_w \boldsymbol{W}$$

$\boldsymbol{W}$的闭式解则是$\boldsymbol{S}_w{}^{-1}\boldsymbol{S}_b$的$d' \leq K-1$个最大非零广义特征值所对应的特征向量组成的矩阵。

Fisher线性判别分析：多分类问题

- K类判别函数可以定义为：

$$\hat{y}_k(\boldsymbol{x}) = \boldsymbol{w}_k^T\boldsymbol{x} + b_k, \qquad k = 1, \dots, K$$

- 对于$\boldsymbol{x}$，如果

$$\hat{y}_k(\boldsymbol{x}) > \hat{y}_j(\boldsymbol{x}), \qquad j \neq k,$$

  那么将$\boldsymbol{x}$分类到$\mathcal{C}_K$。

- 于是，类别$\mathcal{C}_k$和$\mathcal{C}_j$之间分类决策面为$\hat{y}_k(\boldsymbol{x}) = \hat{y}_j(\boldsymbol{x})$，对应于一个$K-1$的超平面，形式为

$$(\boldsymbol{w}_k - \boldsymbol{w}_j)^T\boldsymbol{x} + \left(b_k - b_j\right) = 0$$

北京航空航天大学 软件学院

Fisher线性判别分析：降维/分类器学习步骤

- 对Fisher线性判别分析的降维/分类器学习步骤描述如下：

1. 计算数据样本集中每个类别样本的均值
2. 计算类内散度矩阵S_w和类间散度矩阵S_b
3. 根据${S_w}^{-1}S_bW = \lambda W$来求解${S_w}^{-1}S_b$所对应前$r$个最大特征值所对应特征向量$(w_1, w_2, \dots, w_r)$，构成矩阵$W$
4. 通过矩阵W将每个样本映射到低维空间，实现特征降维或分类器学习。

北京航空航天大學 软件学院

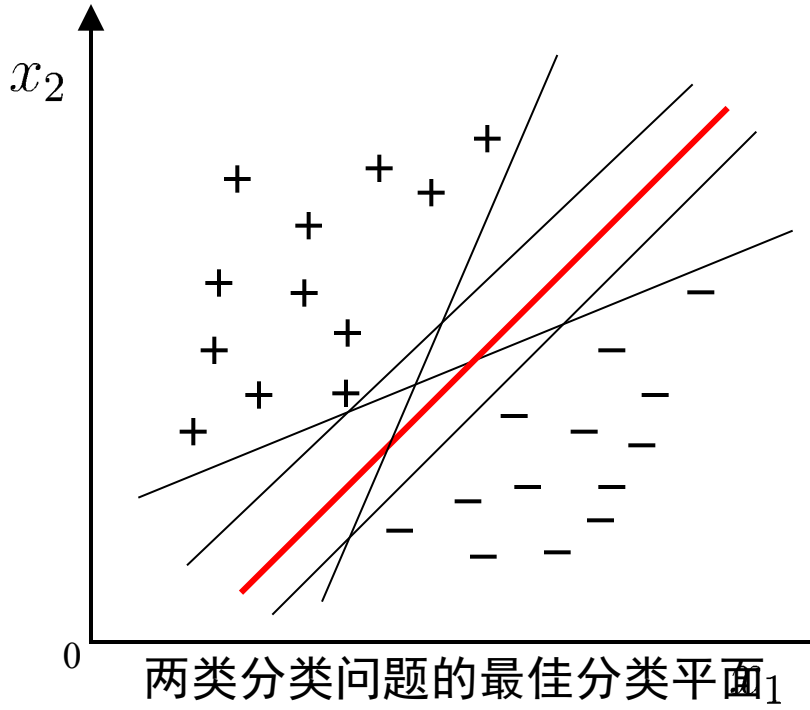
提纲

线性分类-从感知器模型到支持向量机

- 感知器模型

$$\hat{y}(\boldsymbol{x}) = f(\boldsymbol{w}^T\boldsymbol{x}) = sign(\boldsymbol{w}^T\boldsymbol{x}),\ \text{其中} f(z) = \begin{cases} +1, & z \geq 0 \\ -1, & z < 0 \end{cases}$$

  - $f(\boldsymbol{w}^T\boldsymbol{x}) > 0 \Rightarrow C_1$
  - $f(\boldsymbol{w}^T\boldsymbol{x}) < 0 \Rightarrow C_2$
  - 问题1：阶跃函数（分段常数函数）难以优化
  - 感知器准则：$J_P(\boldsymbol{w}) = \sum_{i=1}^{n} \max(0, -y_i\boldsymbol{w}^T\boldsymbol{x_i})$
  - 问题2：无限多完全正确分类解，哪个最佳？
  - 支持向量机：最大化分类间隔



两类分类问题的最佳分类平面

支持向量机：线性可分支持向量机

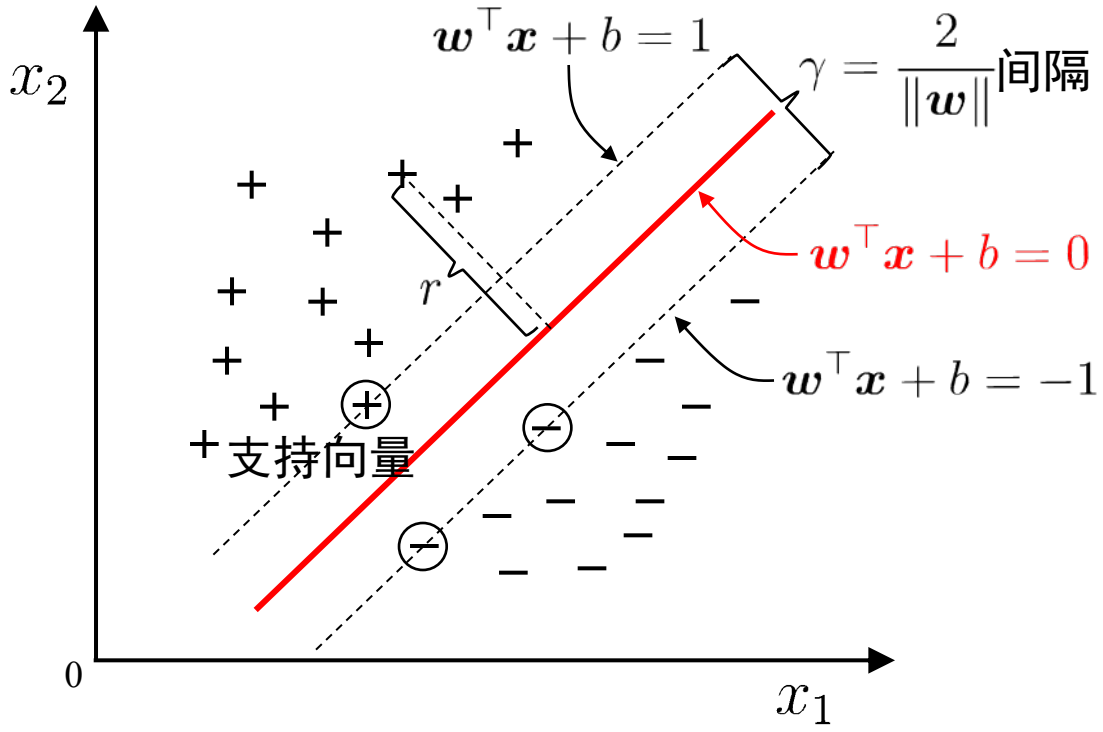


图4.12 线性分类中分类平面及其支持向量

- 寻找一个最优的超平面，其方程为

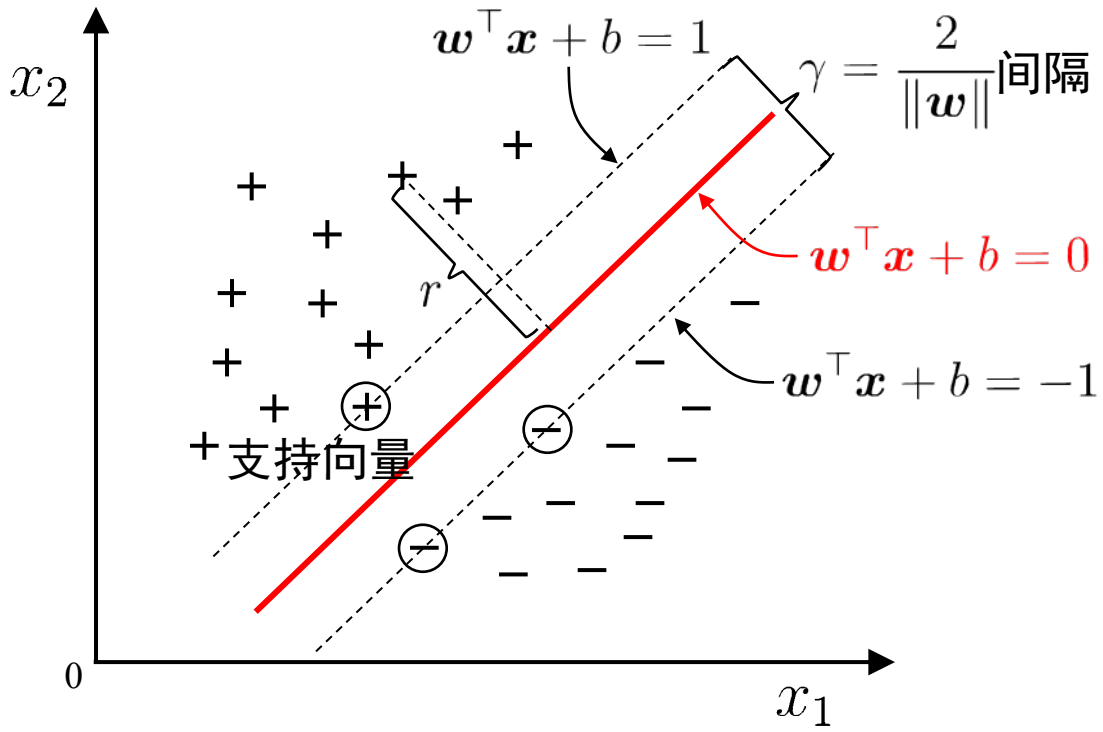
$$\boldsymbol{w}^T \boldsymbol{x} + b = 0$$

- 这里$\boldsymbol{w} = (w_1, w_2, \ldots, w_d)$为超平面的法向量，与超平面的方向有关；
- b为偏置项，是一个标量，其决定了超平面与原点之间的距离。
- 使得分类间隔最大！

支持向量机：线性可分支持向量机

样本空间中任意样本$\boldsymbol{x}$到该平面距离可表示为:

$$r = d(\boldsymbol{w}, b, \boldsymbol{x}) = \frac{|\boldsymbol{w}^T\boldsymbol{x}+b|}{\|\boldsymbol{w}\|_2} \quad (\|\boldsymbol{w}\|_2 = \sqrt{\boldsymbol{w}^T\boldsymbol{w}})$$



- 由于法向量$\boldsymbol{w}$中的值可按比例任意缩放而不改变法向量方向，使得分类平面不唯一。为此，对$\boldsymbol{w}$和b添加如下约束：

$$min_i|\boldsymbol{w}^T\boldsymbol{x}_i + b| = 1$$

- 即离超平面最近的正负样本代入超平面方程后其绝对值为1。于是对超平面的约束变为：

$$y_i(\boldsymbol{w}^T\boldsymbol{x_i} + b) \geq 1$$

- 其中，满足等号成立的样本被称为支持向量（support vector）

支持向量机：线性可分支持向量机

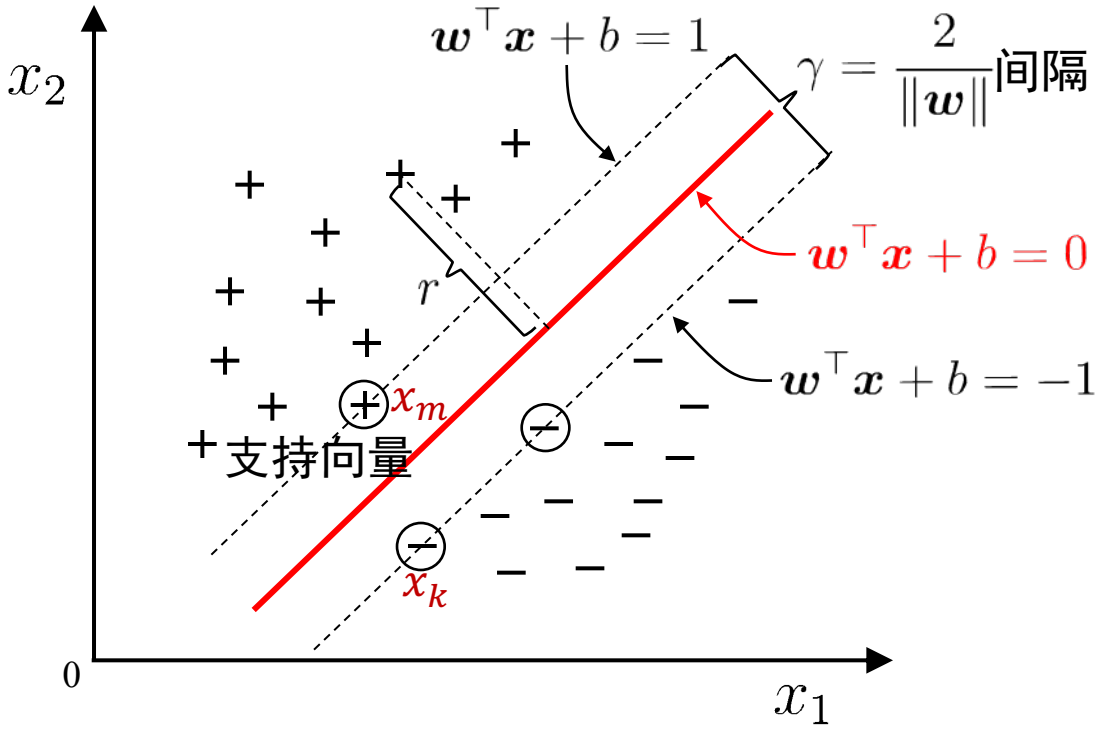


图4.12 线性分类中分类平面及其支持向量

- 两类样本中离分类超平面最近的数据之间的距离可如下计算：

$$d(\boldsymbol{w},b)=min_{(\boldsymbol{x}_k,y_k=1)}d(\boldsymbol{w},b,\boldsymbol{x}_k)+min_{(\boldsymbol{x}_m,y_m=-1)}d(\boldsymbol{w},b,\boldsymbol{x}_m)$$

$$=min_{(\boldsymbol{x}_k,y_k=1)}\frac{|\boldsymbol{w}^T\boldsymbol{x}_k+b|}{\|\boldsymbol{w}\|_2}+min_{(\boldsymbol{x}_m,y_m=-1)}\frac{|\boldsymbol{w}^T\boldsymbol{x}_m+b|}{\|\boldsymbol{w}\|_2}$$

$$=\frac{1}{\|\boldsymbol{w}\|_2}\left(min_{(\boldsymbol{x}_k,y_k=1)}|\boldsymbol{w}^T\boldsymbol{x}_k+b|+min_{(\boldsymbol{x}_m,y_m=-1)}|\boldsymbol{w}^T\boldsymbol{x}_m+b|\right)$$

$$=\frac{2}{\|\boldsymbol{w}\|_2}$$

即：$d(\boldsymbol{w},b)=\frac{2}{\|\boldsymbol{w}\|_2}$

- 支持向量机的基本形式就是最大化分类间隔，即在满足约束的条件下找到参数$\boldsymbol{w}$和b使得γ最大，即：

$$\min_{\boldsymbol{w},b}\frac{\|\boldsymbol{w}\|^2}{2}=\min_{\boldsymbol{w},b}\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

凸二次规划

$$\text{s.t.}\ y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)\geq 1, i=1,2,\ldots,n$$

支持向量机：线性可分支持向量机

- 还可以通过拉格朗日对偶性（Lagrange Duality） 变换到对偶变量 (dual variable) 的优化问题

$$\min_{\boldsymbol{w},b} \frac{||\boldsymbol{w}||^2}{2} = \min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\text{s.t. } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, i = 1,2,\dots,n$$

引入拉格朗日乘子α

$$\min_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1)$$

支持向量机：线性可分支持向量机

$$\min_{\boldsymbol{w},b} L(\boldsymbol{w},b,\boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i (y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1)$$

- 给定一组$\boldsymbol{w}, b$，当$\alpha_i \geq 0$，
  - 如果不满足约束条件$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, i = 1,2,\dots,n$，那么$\max_{\boldsymbol{\alpha}} L(\boldsymbol{w},b,\boldsymbol{\alpha}) = +\infty$
  - 如果满足约束条件$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, i = 1,2,\dots,n$，那么$\max_{\boldsymbol{\alpha}} L(\boldsymbol{w},b,\boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2$
- 因此，等价的优化问题为：$\min_{\boldsymbol{w},b} \max_{\boldsymbol{\alpha}} L(\boldsymbol{w},b,\boldsymbol{\alpha})$
- 根据拉格朗日对偶性，原始问题的对偶问题是：

$$\max_{\boldsymbol{\alpha}} \min_{\boldsymbol{w},b} L(\boldsymbol{w},b,\boldsymbol{\alpha})$$

支持向量机：线性可分支持向量机

$$\max_{\boldsymbol{\alpha}} \min_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i (y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1)$$

先求 $\min_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \boldsymbol{\alpha})$，根据

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

再求原问题的对偶问题 $\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} \min_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \boldsymbol{\alpha})$：

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

$$s.t. \sum_{i=1}^{n} \alpha_i y_i = 0,$$

$$\alpha_i \geq 0, i = 1,2, \ldots n$$

支持向量机：线性可分支持向量机

解出α后，求出$\boldsymbol{w}$与b即可得到模型

$$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$$
$$= \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + b$$

如何求出 b呢？注意对任意支持向量$(\boldsymbol{x}_s, y_s)$都有 $y_s f(\boldsymbol{x_s}) = 1$, 即$y_s\left(\sum_{i\in S} \alpha_i y_i \boldsymbol{x}_i^T \boldsymbol{x_s} + b\right) = 1$，那么可以使用所有支持向量求解的平均值得到

$$b = \frac{1}{|S|}\left(\frac{1}{y_s} - \sum_{i\in S} \alpha_i y_i \boldsymbol{x}_i^T \boldsymbol{x_s}\right)$$

如何测试？

$$sign(\boldsymbol{w}^T\boldsymbol{x} + b) = sign\left(\sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + b\right)$$
$$= sign\left(\sum_{i\in S} \alpha_i y_i \boldsymbol{x}_i^T \boldsymbol{x_s} + b\right)$$

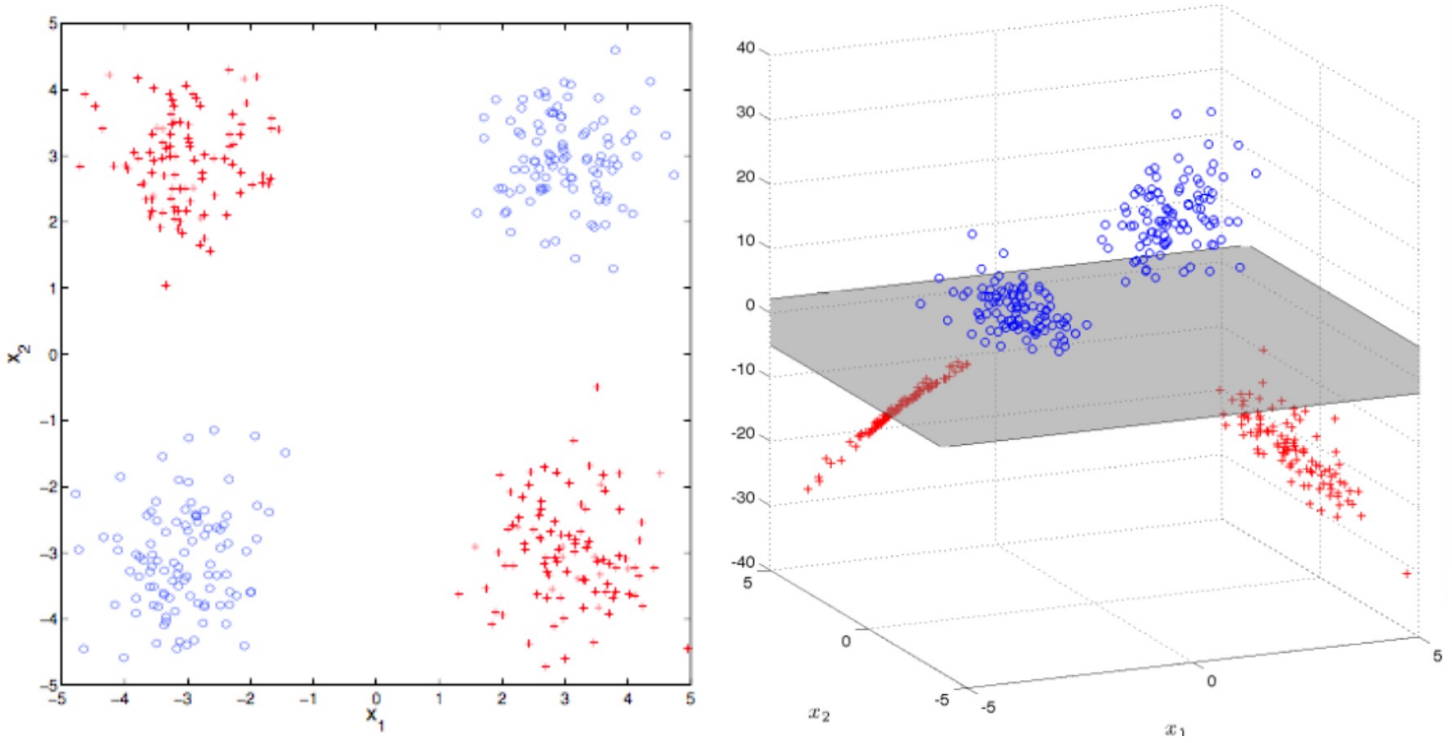
$S = \{i|\alpha_i > 0, i = 1,2,..n\}$表示所有支持向量下标集

支持向量机：线性可分支持向量机

- 支持向量机的求解转换为对偶问题的好处是
  - 仅需求解一个变量α
  - 对偶问题约束方程简单
  - 优化问题可以转化为高效算法，如SOM（Sequential Minimal Optimization）
  - 模型转化为输入样本之间的内积形式，便于核函数的引入。

支持向量机：线性不可分-核函数

- 将线性不可分样本从原始空间映射到一个更加高维的特征空间中去，使得样本在这个特征空间中高概率线性可分。
- 如果原始空间是有限维，那么一定存在一个高维特征空间使样本可分[Shawe-Taylor, J. 2004]。



支持向量机：线性不可分-核函数

- 支持向量机模型

$$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$$
$$= \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + b$$

- 将 $\boldsymbol{x}$映射到高位空间，得到新的特征向量$\phi(\boldsymbol{x})$，模型转换为

$$f(\boldsymbol{x}) = \boldsymbol{w}^T\phi(\boldsymbol{x}) + b$$
$$= \sum_{i=1}^{n} \alpha_i y_i \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}) + b$$
$$= \sum_{i=1}^{n} \alpha_i y_i k(\boldsymbol{x}, \boldsymbol{x}_i) + b$$

这里 $k(\boldsymbol{x}, \boldsymbol{x}_i)$就是核函数（kernel function）

支持向量机：线性不可分-核函数

- 有效核函数需要满足：
  - 对称性： $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = k(\boldsymbol{x}_j, \boldsymbol{x}_i)$
  - 正定性：对于任意$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$，格拉姆矩阵$K$满足半正定性：

$$K = \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & k(\boldsymbol{x}_1, \boldsymbol{x}_2) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_N) \\ \cdots & \cdots & & \cdots \\ k(\boldsymbol{x}_N, \boldsymbol{x}_1) & k(\boldsymbol{x}_N, \boldsymbol{x}_2) & & k(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{pmatrix}$$

  - 半正定性指：对于任意$\boldsymbol{x}$，$\boldsymbol{x}^T K \boldsymbol{x} \geq 0$

支持向量机：线性不可分-核函数

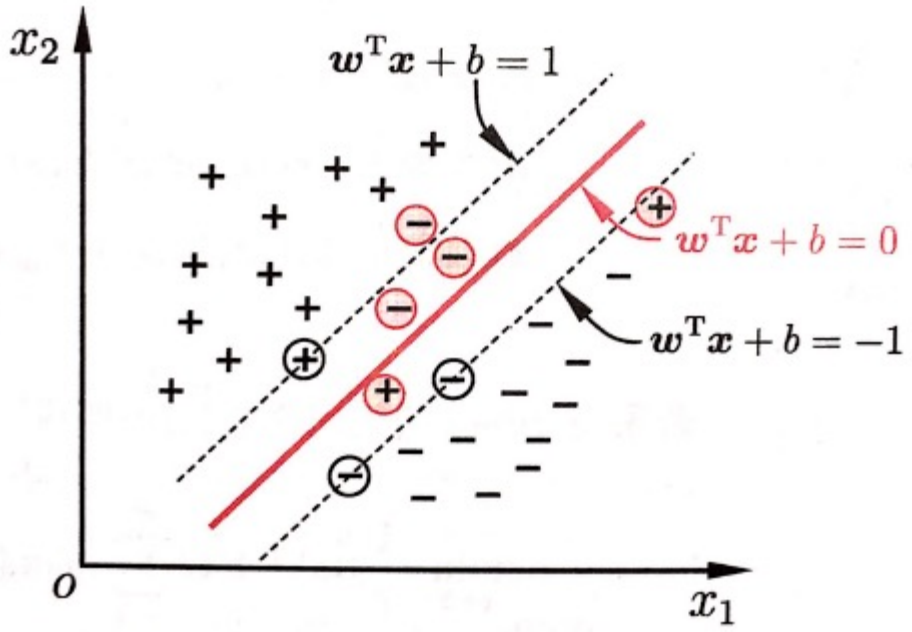
- 常见的核函数包括：

常见核函数

线性	$\kappa(x_i, x_j) = x_i x_j$
多项式	$\kappa(x_i, x_j) = (\gamma x_i x_j + c)^n$
Radial basis function	$\kappa(x_i, x_j) = e^{-\frac{\|\|x_i - x_j\|\|^2}{2\sigma^2}}$
Sigmoid	$\kappa(x_i, x_j) = tanh(x_i, x_j - \gamma)$

支持向量机：线性不可分-松弛变量，软间隔与hinge损失函数

先前介绍中假设所有训练样本数据是线性可分，即存在一个线性超平面能将不同类别样本完全隔开，这种情况称为“硬间隔”（hard margin），与硬间隔相对的是“软间隔”（soft margin）。软间隔指允许部分错分给定的训练样本。



$$\min_{\boldsymbol{w},b}\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+c\times\sum_{i=1}^{n}[\![y_i\neq sign(\boldsymbol{w}^T\boldsymbol{x}_i+b)]\!]$$

$$s.t.\ \ y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)\geq 1\ for\ correct\ \boldsymbol{x}_i$$

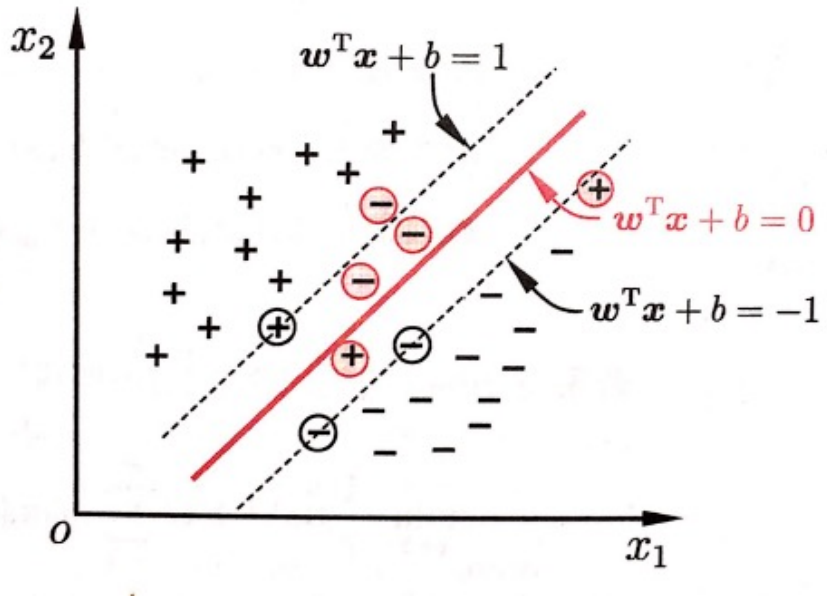
$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)\geq -\infty\ for\ incorrect\ \boldsymbol{x}_i$$

难以直接求解

支持向量机：线性不可分-松弛变量，软间隔与hinge损失函数

hinge损失函数：$\min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + c\sum_{i=1}^{n} \max(0,1-y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b))$

正确分类数据的hinge损失中 $\max\big(0,1-y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)\big)=0$



- 记$\xi_i = \max\big(0,1-y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)\big)$（$\xi_i$被称为第$i$个变量的“松弛变量”，slack variables），显然$\xi_i \geq 0$。每一个样本对应一个松弛变量，用来表示该样本被分类错误所产生的损失。于是，可将上式重写为：

$$\min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + c \times \sum_{i=1}^{n} \xi_i$$

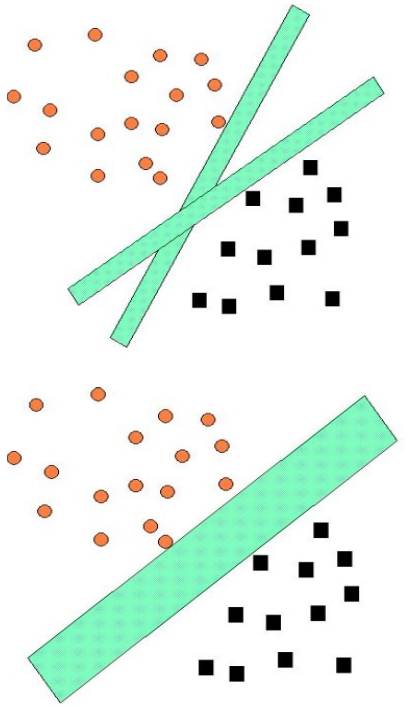
$$\text{s.t.}\quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b) \geq 1-\xi_i$$
$$\xi_i \geq 0, i = 1,2,\ldots,n$$

拉格朗日乘子法

线性分类模型：正则项

- hinge损失函数：

$$\min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + c\sum_{i=1}^{n} \max(0,1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b))$$



- 如果替换成其他损失函数，得到更一般的分类模型形式，

$$min_f\, \Omega(f) + c\sum_{i=1}^{n} l(f(\boldsymbol{x}_i), y_i)$$

  - 其中，$\Omega(f)$称为结构风险（structural risk），也被称为正则项，用于描述模型f的复杂度等性质。常用的正则化方法包括L_p范数，其中L_2范数$\|\boldsymbol{w}\|_2$倾向于$\boldsymbol{w}$的分量取值较小，L_1范数$\|\boldsymbol{w}\|_1$倾向于$\boldsymbol{w}$的分量尽量稀疏。
  - $\sum_{i=1}^{n} l(f(\boldsymbol{x}_i), y_i)$称为经验风险（empirical risk），用于描述模型f与训练数据的契合度。

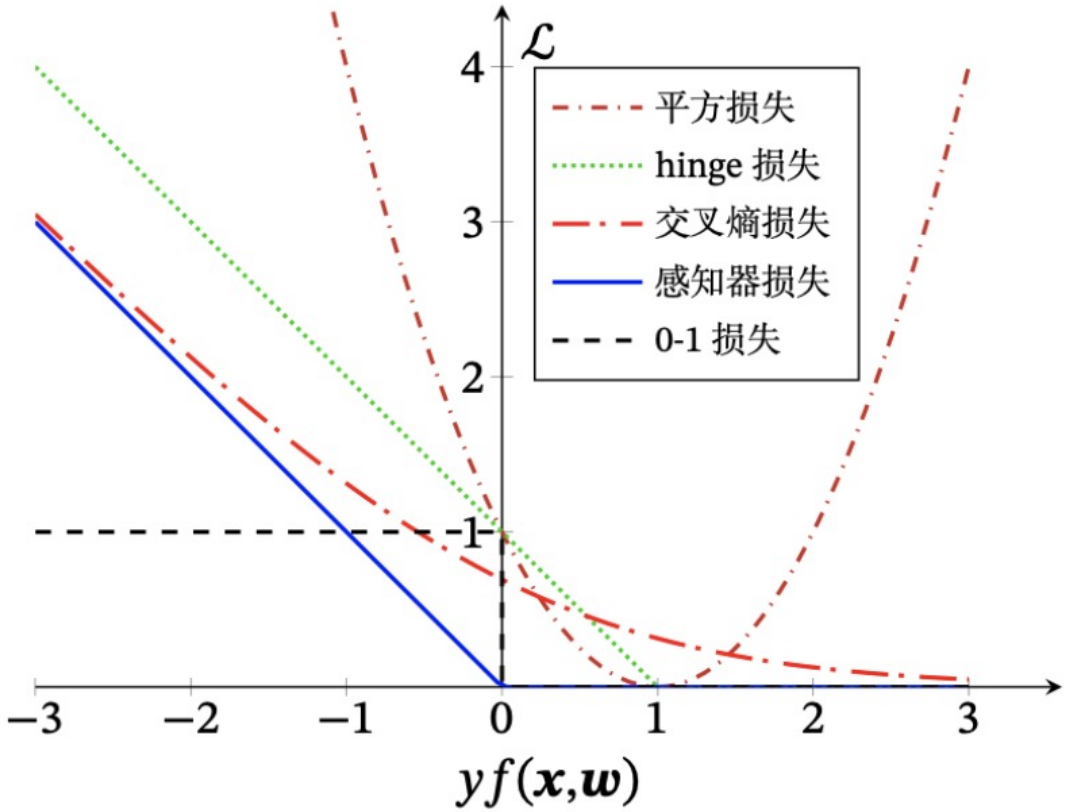
支持向量机

- 更多阅读和视频资源
  - 书籍：机器学习，周志华
  - 视频：

    https://www.bilibili.com/video/BV1jt4y1E7BQ/?vd_source=5afe770cafa42a833cbfdbba5d750438
  - 博客： https://blog.csdn.net/v_JULY_v/article/details/7624837

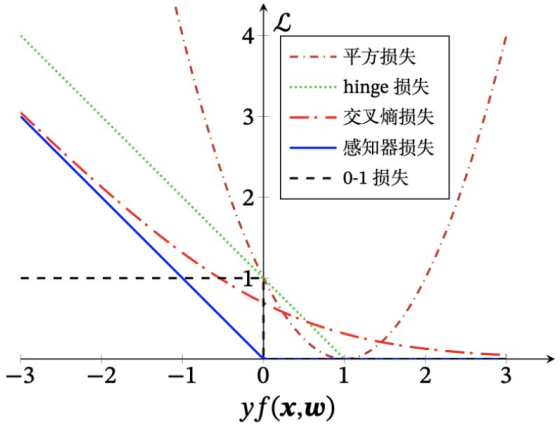
总结：线性分类模型

- 线性回归：$\frac{1}{N}\sum(y_i - f(x_i))^2$
- 对数几率回归：$-logP(y_i|x_i)$
- 线性判别分析
- 感知器
- 支持向量机
- …



二分类问题中不同损失函数的对比
（横轴表示$yf(\boldsymbol{x},\boldsymbol{w})$，纵轴表示损失）

总结：线性分类模型



线性模型	激活函数	损失/目标函数	损失/目标函数定义	优化方法
		0-1损失	$\begin{cases}1, f(\boldsymbol{x}_i) \neq y_i \\ 0, f(\boldsymbol{x}_i) = y_i\end{cases}$	
线性回归	-	平方损失	$(y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2$	最小二乘、梯度下降
对数几率回归	$sigmoid(\boldsymbol{w}^T\boldsymbol{x})$	二值交叉熵损失	$-y_i \log \sigma(\boldsymbol{w}^T\boldsymbol{x}_i) + (1-y_i)\log(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}_i))$	梯度下降
Softmax分类	$softmax(\boldsymbol{W}^T\boldsymbol{x})$	交叉熵损失	$-y_i \log softmax(\boldsymbol{w}^T\boldsymbol{x}_i)$	梯度下降
线性判别分析	-	Fisher准则	$\frac{\|m_2 - m_1\|_2^2}{s_1^2 + s_2^2}$	广义特征值分解
感知器	$sign(\boldsymbol{w}^T\boldsymbol{x})$	感知器准则	$\max(0, -y_i\boldsymbol{w}^T\boldsymbol{x}_i)$	随机梯度下降
支持向量机	$sign(\boldsymbol{w}^T\boldsymbol{x})$	Hinge损失	$\max(0, 1 - y_i\boldsymbol{w}^T\boldsymbol{x}_i)$	二次规划、SMO等