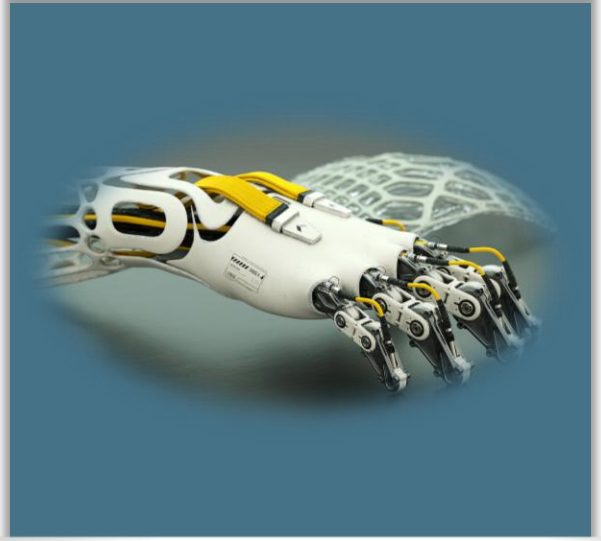
机器学习第三讲-II

授课人：王闻博

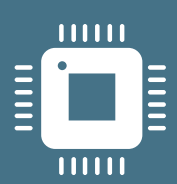
Email: wenbo_wang@kust.edu.cn

昆明理工大学 机电工程学院

2026年03月27日



统计机器学习中的分类问题




1. 判别函数
2. 概率生成模型和判别模型
3. 贝叶斯逻辑回归（Logistic Regression）
4. 分类模型的评价指标

最简单的二分类模型

- **二分类模型**

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$

  - $\mathbf{w}$为权重向量，w_0为偏置量（bias）；
  - 若$y(\mathbf{x}) \geq 0$，则x属于class 1 （t=1）；反之则属于class 2 （t=-1） ⟹ 决策边界：$y(\mathbf{x}) = 0$
  - 对比回归问题的连续输出：更一般地，y可以是类别的后验概率$p(C_i|\mathbf{x})$的估计值：$y \in [0,1]$。

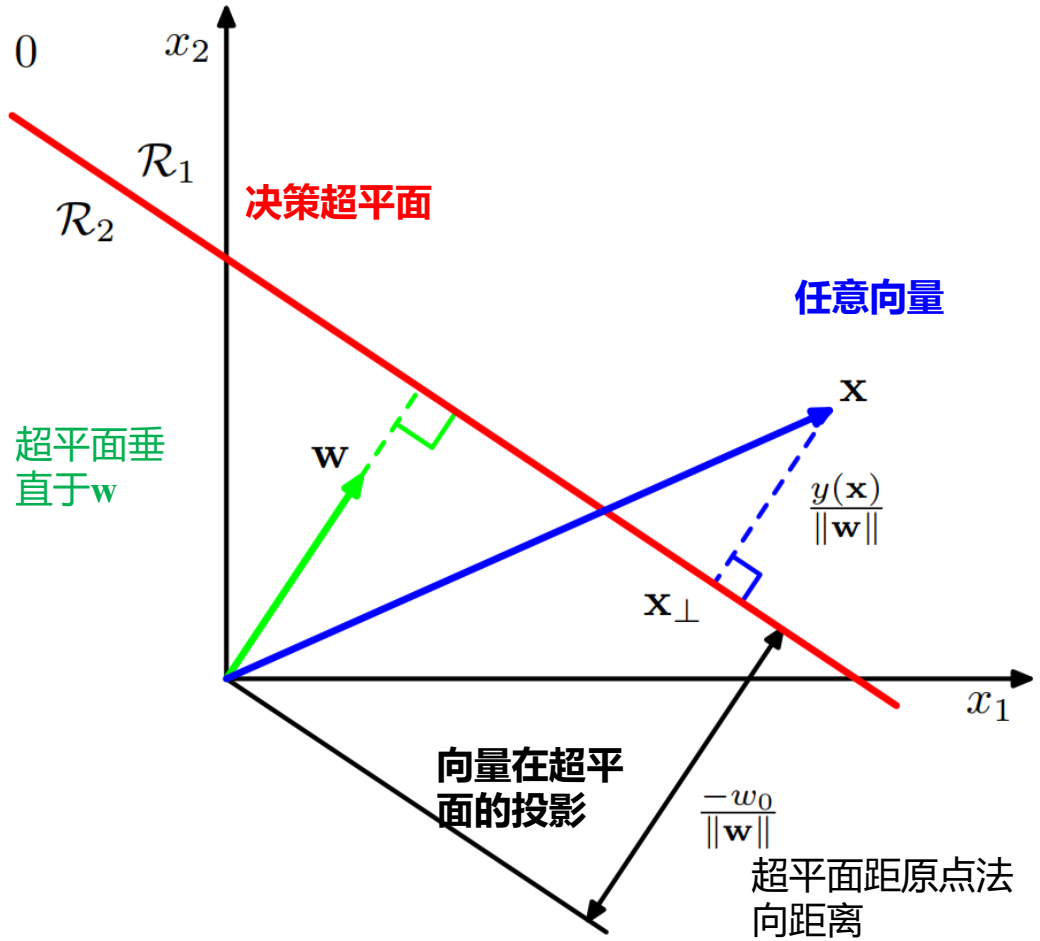
- **广义线性模型**

  - 在线性模型基础上引入非线性函数，或称激活函数（Activation Function）：$y(\mathbf{x}) = \boldsymbol{f}(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0)$；
  - 决策分界面 $y(\mathbf{x}) = 0$；

    由此有$\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = \text{const.}$，对应一个向量空间里的（D − 1）维的超平面；
  - 由于超平面的形式所限，即便$f(\mathbf{x})$是非线性的，决策分界面（Decision Surface）依然是线性的，称为广义线性模型（Generalized Linear Model）。

二分判别函数（Discriminant Functions）

- **决策分界超平面**
  - 考虑决策平面上的两个点$y(\mathbf{x}_A)=0$ 和$y(\mathbf{x}_B)=0$，有$\mathbf{w}^T(\mathbf{x}_A-\mathbf{x}_B)=\mathbf{0}$；
  - 因此，$\mathbf{w}$正交于所有决策超平面中的向量，即权重向量$\mathbf{w}$决定超平面的朝向；
  - 对任意超平面上的点$\mathbf{x}$，$y(\mathbf{x})=0$，可计算原点到超平面的法向距离：$\frac{\mathbf{w}^T\mathbf{x}}{\|\mathbf{w}\|}=-\frac{w_0}{\|\mathbf{w}\|}$
    即，偏置量w_0决定超平面的的位置。
  - 任意样本点距超平面距离为$\frac{f(\mathbf{x})}{\|\mathbf{w}\|}$，因此距离值越大，样本点离超平面越远，分类置信度越高。

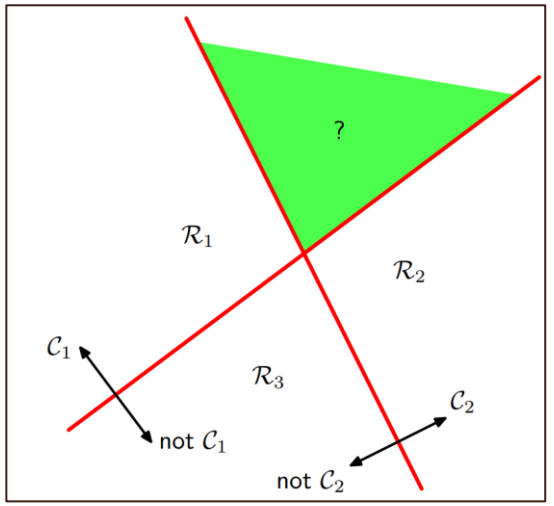


$$r=\frac{y(\mathbf{x})}{\|\mathbf{w}\|}$$ 任意一点到超平面的距离

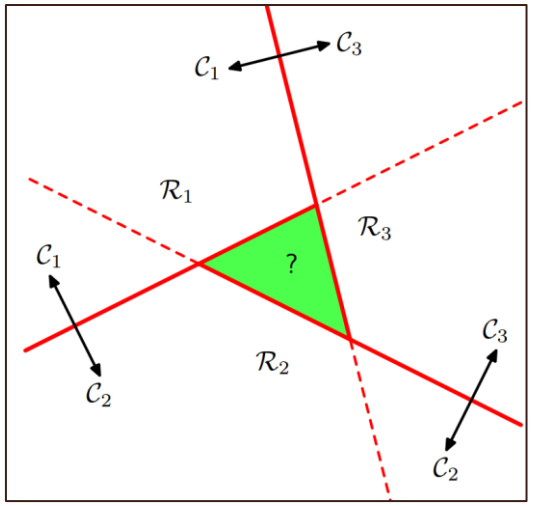
图来源：Bishop《PRML》

多分类模型

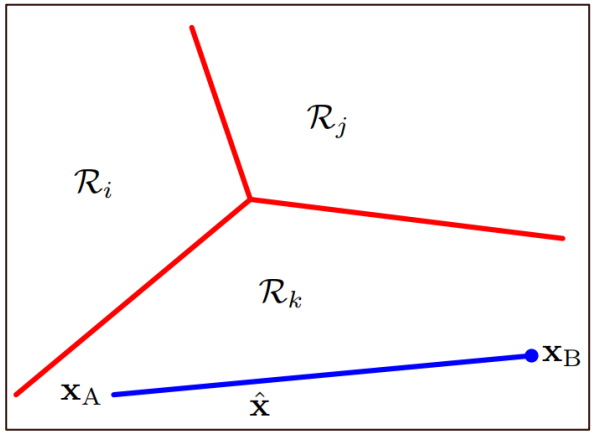
- **在二分类模型基础上扩展到K分类模型**
  - 一对其他 (One-vs-Rest)划分：每个分类器使用多个针对一某个类C_k的二分类器
    - 二分类判别函数数量为$K-1$;
  - 一对一 (One-vs-One)划分：对每两个可能的类（组成一个二分类对）引入一个二分判别函数，
    - 二分类判别函数数量为$K(K-1)/2$;
- **单一的K类判别函数（无模糊域）**
  - 引入K个线性函数：$y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$
  - 决策规则：if $y_k(\mathbf{x}) > y_j(\mathbf{x})$ for all $j \neq k$. ➡ $\mathbf{x} \in C_k$
  - 决策边界：$y_k(\mathbf{x}) = y_j(\mathbf{x})$
  - 超平面方程：$(\mathbf{w}_k - \mathbf{w}_j)^{\mathrm{T}}\mathbf{x} + (w_{k0} - w_{j0}) = 0$



一对其他划分情况下，（K-1=2）个分类判别函数导致模糊区域（图中绿色区域）



一对一划分情况下，K(K-1)/2=3个分类判别函数导致模糊区域（图中绿色区域）



凸性：任意位于同一决策区域的两点连线一定仍在这一决策区域（见蓝色直线）

图来源：Bishop《PRML》

最小二乘法分类器

- **训练集数据：**
  - 样本-标签对：$\{\mathbf{x}_n, \mathbf{t}_n\}\ n = 1, \ldots, N$
  - 统一权重和偏差量为向量表示：$y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$

把K个判别函数写成矩阵形式 →

$$\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{x}}$$

$\widetilde{\mathbf{W}}$的第k列为（D+1）维向量：

$\widetilde{\mathbf{w}} = \left(w_{k0}, \mathbf{w}_k^T\right)^T$；

$\tilde{\mathbf{x}} = \left(1, \mathbf{x}^{\mathrm{T}}\right)^T$；

  - 类似多元线多输出性回归的操作，平方和误差（SSE）损失可以写为：

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left\{(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^{\mathrm{T}}(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})\right\}$$

  - 参数集的最小二乘解形式（参见多元线多输出性回归）：

$$\widetilde{\mathbf{W}} = (\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^{\mathrm{T}}\mathbf{T} = \widetilde{\mathbf{X}}^{\dagger}\mathbf{T}$$

对于任意向量 $\mathbf{a} \in \mathbb{R}^{n\times 1}$，有 $\mathbf{a}^{\top}\mathbf{a} = \mathrm{tr}(\mathbf{a}\mathbf{a}^{\top})$；

  - 判别函数的形式：

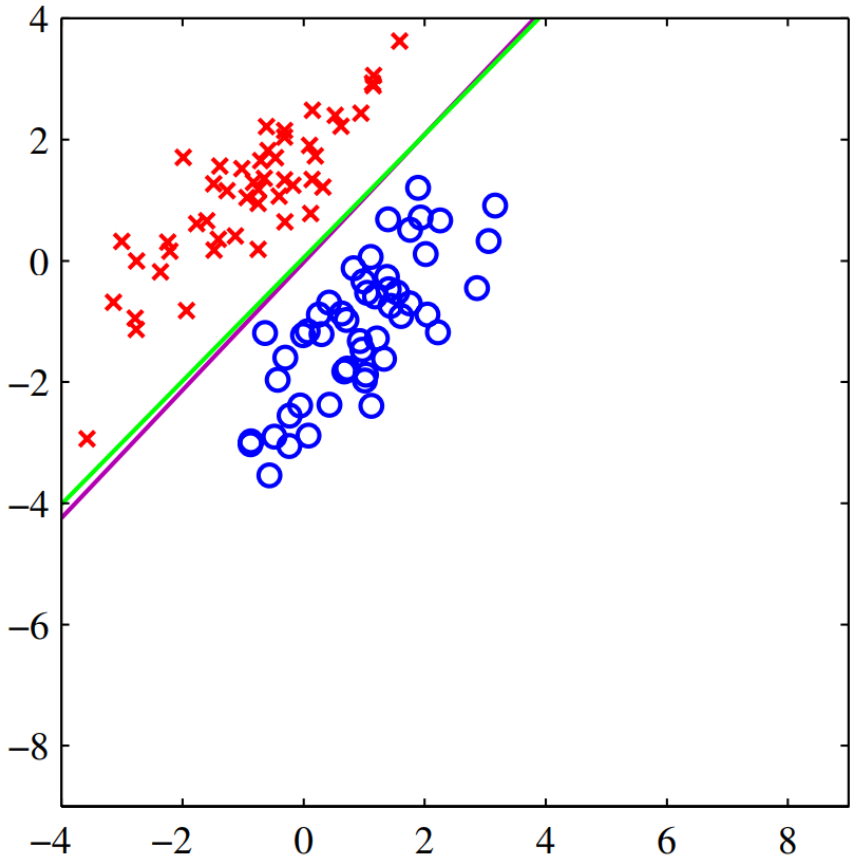
$$\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{x}} = \mathbf{T}^{\mathrm{T}}\left(\widetilde{\mathbf{X}}^{\dagger}\right)^{\mathrm{T}}\widetilde{\mathbf{x}}.$$

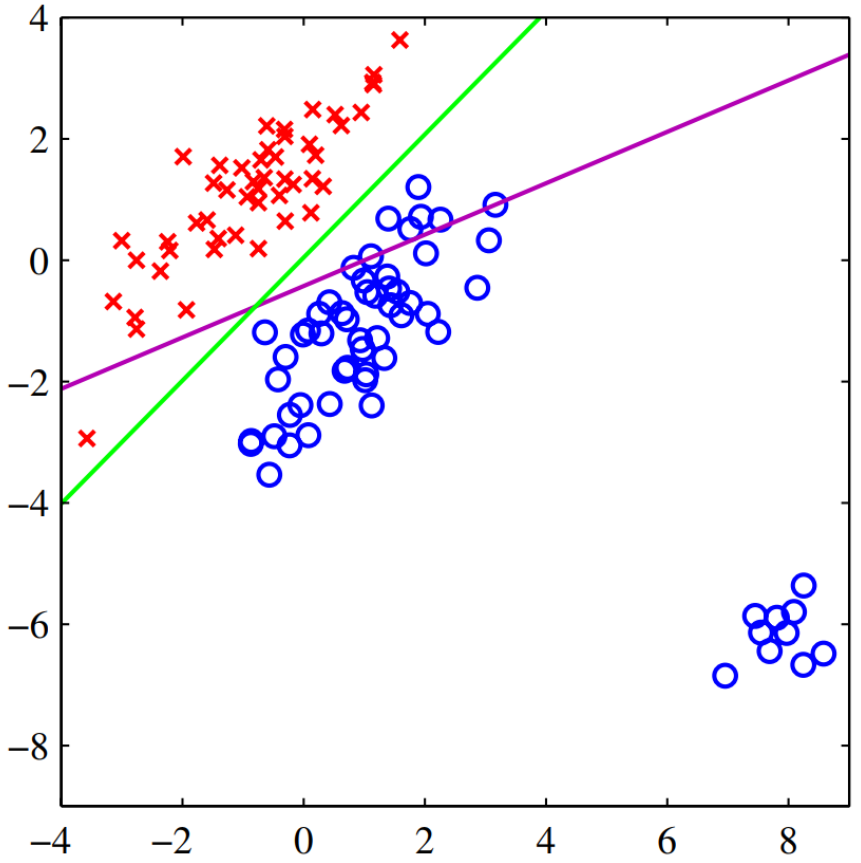
- **在新样本下的分类准则：**

新输入$\mathbf{x} \in C_k$, if $y_k = \widetilde{\mathbf{w}}_k^{\mathrm{T}}\widetilde{\mathbf{x}}$ 取最大值（即某一行输出取最大值）

最小二乘法分类特性

- 两个类别的数据：红色叉号（类别1）和蓝色圆点（类别2）
- 分类结果对比：最小二乘法（紫色曲线）逻辑回归模型（**绿色曲线**，带正则项）的决策边界。
- 最小二乘法对异常值高度敏感（见右图右下角的异常值采样点），而逻辑回归模型不受影响。





图来源：Bishop《PRML》

适用于二分类问题的感知器（Perceptron）方法

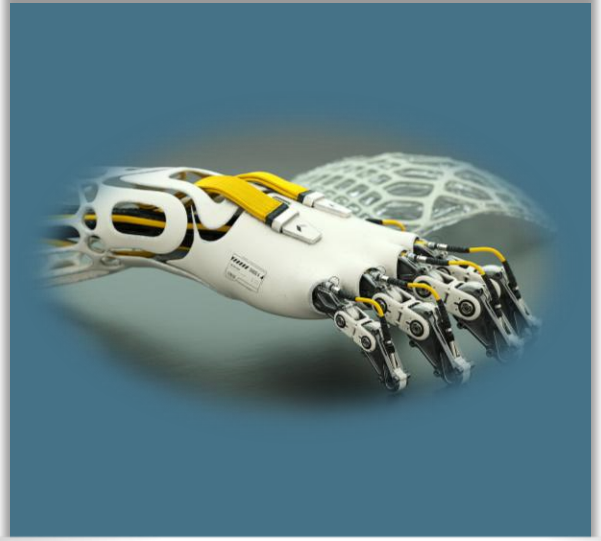
- **定理：若存在一个精确解（即，如果训练数据集是线性可分的），则感知器算法保证在有限步骤内找到一个精确解。**
- **方法描述：迭代算法**
  - 感知器函数：$y(\mathbf{x}) = f\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right)$ 取非线性激活函数 $f(a) = \begin{cases} +1, & a \geqslant 0 \\ -1, & a < 0. \end{cases}$
  - 目标值：$t = +1$ for class $\mathcal{C}_1$ and $t = -1$ for class $\mathcal{C}_2$.
  - 构建损失函数（分类判别准则）：

$$E_{\mathrm{P}}(\mathbf{w}) = -\sum_{n \in \mathcal{M}} \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)t_n$$

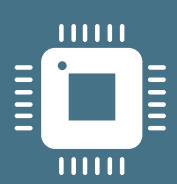
其中 $\mathcal{M}$ 代表所有错误分类的样本集合

  - 基于梯度下降法的参数更新：

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta\nabla E_{\mathrm{P}}(\mathbf{w}) = \mathbf{w}^{(\tau)} + \eta\boldsymbol{\phi}_n t_n$$



统计机器学习中的分类问题




1. 判别函数
2. 概率生成模型和判别模型
3. 贝叶斯逻辑回归（Logistic Regression）
4. 分类模型的评价指标

概率生成模型：高斯朴素贝叶斯分类器

- **原理：建立基于类别的条件密度函数$p(\mathbf{x}|C_k)$，以及关于类别的先验概率密度函数$p(C_k)$，根据Bayes定理，计算后验概率$p(C_k|\mathbf{x})$。**
  - 在**二分类**情况下，假设类条件概率为高斯分布，且所有类别具有同样的协方差矩阵

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}$$

  - 后验概率：

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)+p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} = \frac{1}{1+\exp(-a)} = \sigma(a)$$

  - 上式中，定义：$a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$ 且 $\sigma(a) = \frac{1}{1+\exp(-a)}$ logistic sigmoid函数
  - 后验概率扩展到**K分类**情况下：

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \quad \text{其中} \quad a_k = \ln p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$$

此归一化指数函数即为著名的Softmax函数

概率生成模型（续）

- **考虑二分类情况**

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)+p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2} = \frac{1}{1+\exp(-a)} = \sigma(a)$$

  - 用线性模型替代a:

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}+w_0)$$

我们的目的在于确定w和w_0

  - 则根据高斯分布假设（见上页）：

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}$$

  - 模型参数可求得如下:

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

$$w_0 = -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1+\frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2+\ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

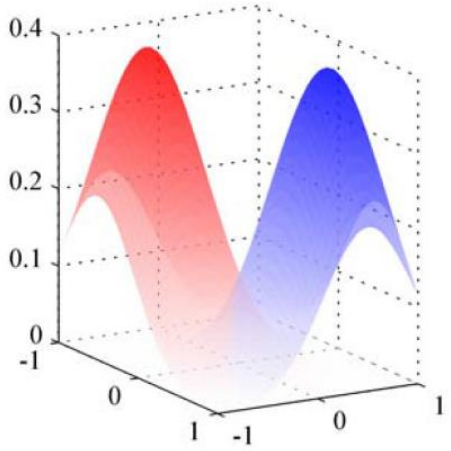
- **扩展到K分类情况（不经证明地）**

$$a_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x}+w_{k0} \quad \text{其中}$$

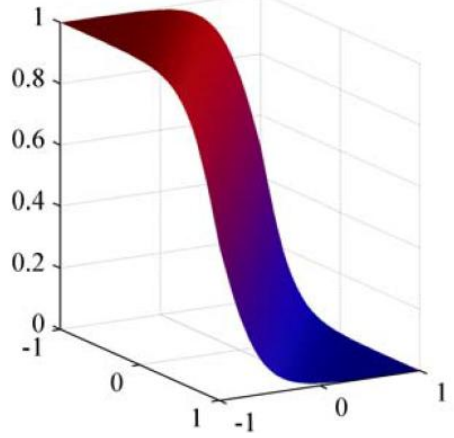
$$\mathbf{w}_k = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k$$

$$w_{k0} = -\frac{1}{2}\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k+\ln p(\mathcal{C}_k)$$

二分类概率生成模型图示

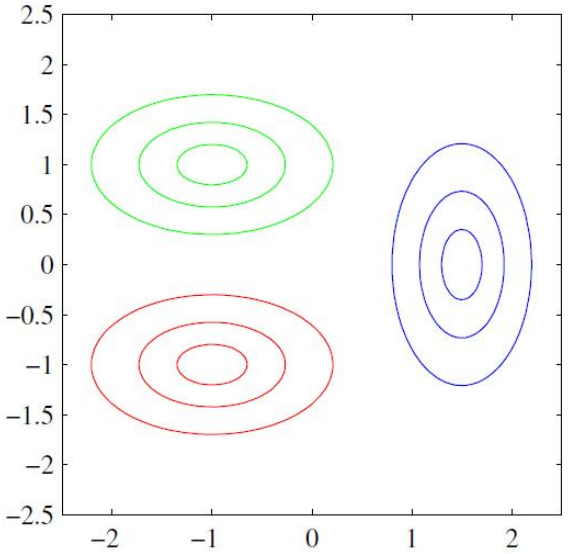


基于类别的条件密度函数

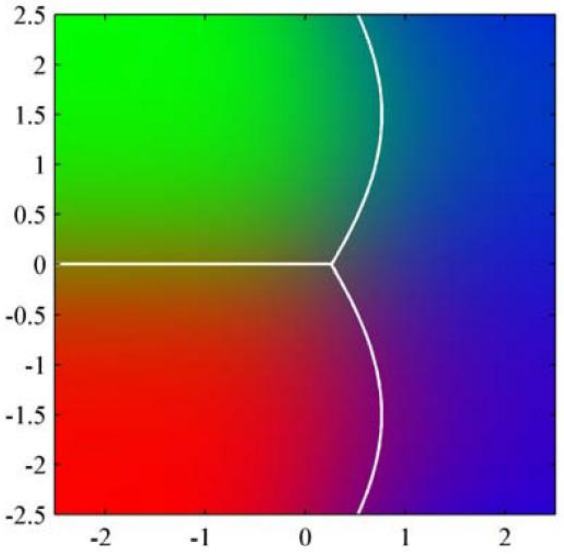


后验概率：logistic sigmoid函数

二分类的情况



基于类别的条件密度函数



后验概率和决策界

三分类的情况，条件概率共享协方差矩阵

阶段性总结：从贝叶斯公式到概率生成模型

后验概率　似然度　先验概率

$$P(Y|X) = \frac{P(X,Y)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$$

边际似然度/模型证据（Evidence）

- **我们希望通过观测数据 $\mathbf{D}=\{\mathbf{X}_1,\ldots,\mathbf{X}_N\}$，得到分类模型，即，给定新特征X属于类别Y的概率。**
- **贝叶斯法是典型的生成学习方法。生成方法由训练数据学习联合概率分布 $P(\mathrm{X},\mathrm{C})$，然后求得后验概率分布 $P(\mathrm{C}|\mathrm{X})$.**
- **分类决策：** $\arg_c\max P(\mathrm{Y}=\mathrm{c}|\mathrm{X}) = \frac{P(X|Y)P(Y)}{P(X)}$
- **一般步骤**
  - **计算先验概率**：估计每个类别的 先验概率$P(Y=c)$。
  - **估计似然度**：对每个特征 X_i，计算 $P(X_i|Y=c)$。
  - **计算联合概率**：对测试样本 X，计算每个类别的 $P(X|Y=c)P(Y=c)$。
  - **选择最大后验概率**：预测类别为 $\arg_c\max P(Y=c|X)$。

- 实际计算中，由于 $P(X)$ 对所有类别相同，只需比较分子部分：$P(Y=c|X) \propto P(X|Y=c)P(Y=c)$.
- 等价表示：
  - $\arg_c\max P(\mathrm{X}|\boldsymbol{Y})\mathrm{P}(\mathrm{Y})$
  - $\arg_c\max \log P(\mathrm{X}|\boldsymbol{Y})\log\mathrm{P}(\mathrm{Y})$

构建参数化生成模型：二分类的极大似然解

- **基于类别的条件密度函数$p(\mathbf{x}|C_k)$假设，我们可以在先验概率假设$p(C_k)$ 的基础上通过最大似然方法确定参数值**
  - 假设类条件概率密度为高斯分布： $p(x|C_1) = \mathcal{N}(x|\mu_1, \Sigma), \quad p(x|C_2) = \mathcal{N}(x|\mu_2, \Sigma)$
  - 假设两个类别的先验概率分布分别为： $P(C_1)=\pi$ 和 $P(C_2)=1-\pi$；
  - 对于每个样本 $\mathbf{x}_n$，联合概率为：

$$p(\mathbf{x}_n, \mathcal{C}_1) = p(\mathcal{C}_1)p(\mathbf{x}_n|\mathcal{C}_1) = \pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$
$$p(\mathbf{x}_n, \mathcal{C}_2) = p(\mathcal{C}_2)p(\mathbf{x}_n|\mathcal{C}_2) = (1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}).$$

  - 似然函数：是所有样本的联合概率的乘积

$$p(\mathbf{t}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} [\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n} [(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1-t_n}$$

    - 其中 t_n 是训练集中的样本标签，指示样本 $\mathbf{x}_n$ 属于 C_1 或 C_2。

极大似然解（续）

- **考虑似然函数关于π值的最优解：**
  - 似然函数：$$\prod_{n=1}^{N}[\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1,\boldsymbol{\Sigma})]^{t_n}\,[(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2,\boldsymbol{\Sigma})]^{1-t_n}$$
  - 在不考虑类别条件密度函数的情况下，对数似然函数有如下形式：
$$\sum_{n=1}^{N}\{t_n\ln\pi+(1-t_n)\ln(1-\pi)\}$$
  - 关于π 取偏导，有：
$$\pi=\frac{1}{N}\sum_{n=1}^{N}t_n=\frac{N_1}{N}=\frac{N_1}{N_1+N_2}$$

极大似然解（续）

- **考虑似然函数关于μ_1和μ_2值的最优解：**
  - 似然函数：$$\prod_{n=1}^{N}[\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1,\boldsymbol{\Sigma})]^{t_n}[(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2,\boldsymbol{\Sigma})]^{1-t_n}$$
  - 在不考虑先验分布假设和参数Σ的情况下，对数似然函数有如下形式：
$$\sum_{n=1}^{N}t_n\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1,\boldsymbol{\Sigma})=-\frac{1}{2}\sum_{n=1}^{N}t_n(\mathbf{x}_n-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_1)+\text{const}$$
  - 关于μ_1 取偏导并另其为0，有：
$$\boldsymbol{\mu}_1=\frac{1}{N_1}\sum_{n=1}^{N}t_n\mathbf{x}_n$$

类似地，对μ_2：$$\boldsymbol{\mu}_2=\frac{1}{N_2}\sum_{n=1}^{N}(1-t_n)\mathbf{x}_n$$

极大似然解（续）

- **考虑似然函数关于Σ值的最优解：**
  - 似然函数：$$\prod_{n=1}^{N}\left[\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1,\boldsymbol{\Sigma})\right]^{t_n}\left[(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2,\boldsymbol{\Sigma})\right]^{1-t_n}$$
  - 在不考虑先验分布假设的情况下，对数似然函数有如下形式（类似上页的结论）：

$$-\frac{1}{2}\sum_{n=1}^{N}t_n\ln|\boldsymbol{\Sigma}|-\frac{1}{2}\sum_{n=1}^{N}t_n(\mathbf{x}_n-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_1)$$

$$-\frac{1}{2}\sum_{n=1}^{N}(1-t_n)\ln|\boldsymbol{\Sigma}|-\frac{1}{2}\sum_{n=1}^{N}(1-t_n)(\mathbf{x}_n-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_2)=-\frac{N}{2}\ln|\boldsymbol{\Sigma}|-\frac{N}{2}\mathrm{Tr}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{S}\right\}$$

  - 定义：$\mathbf{S}=\frac{N_1}{N}\mathbf{S}_1+\frac{N_2}{N}\mathbf{S}_2$
  - **则根据高斯分布的极大似然解，有**

$$\boldsymbol{\Sigma}=\mathbf{S}$$

其中

$$\mathbf{S}_1=\frac{1}{N_1}\sum_{n\in\mathcal{C}_1}(\mathbf{x}_n-\boldsymbol{\mu}_1)(\mathbf{x}_n-\boldsymbol{\mu}_1)^{\mathrm{T}}$$

$$\mathbf{S}_2=\frac{1}{N_2}\sum_{n\in\mathcal{C}_2}(\mathbf{x}_n-\boldsymbol{\mu}_2)(\mathbf{x}_n-\boldsymbol{\mu}_2)^{\mathrm{T}}$$

二分类极大似然解——总结

- **二分类模型**
  - 后验概率（分类模型）：

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

  - 类别条件概率密度函数：

$$p(x|C_1) = \mathcal{N}(x|\mu_1, \Sigma), \quad p(x|C_2) = \mathcal{N}(x|\mu_2, \Sigma)$$

  - 类分布先验概率：

$P(C_1)=\pi$ 和 $P(C_2)=1-\pi$

  - **参数：**

$$\pi = \frac{1}{N}\sum_{n=1}^{N} t_n = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$$

$$\boldsymbol{\mu}_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n \mathbf{x}_n \qquad \boldsymbol{\mu}_2 = \frac{1}{N_2}\sum_{n=1}^{N} (1 - t_n)\mathbf{x}_n$$

$$\boldsymbol{\Sigma} = \mathbf{S} = \frac{N_1}{N}\mathbf{S}_1 + \frac{N_2}{N}\mathbf{S}_2$$

其中

$$\mathbf{S}_1 = \frac{1}{N_1}\sum_{n\in\mathcal{C}_1} (\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^{\mathrm{T}}$$

$$\mathbf{S}_2 = \frac{1}{N_2}\sum_{n\in\mathcal{C}_2} (\mathbf{x}_n - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_2)^{\mathrm{T}}$$

从二分类到K分类任务：极大似然解

- **考虑K分类情况（不加证明地），有：**
  - 第k类样本子集有N_k个样本，则先验类分布假设的极大似然解：$\pi_k = \dfrac{N_k}{N}$
  - 若基于类别条件样本分布假设符合如下高斯分布：$p(\boldsymbol{\phi}|\mathcal{C}_k) = \mathcal{N}(\boldsymbol{\phi}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$，则分别有

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}\boldsymbol{\phi}_n \qquad \text{以及} \qquad \boldsymbol{\Sigma} = \sum_{k=1}^{K}\frac{N_k}{N}\mathbf{S}_k$$

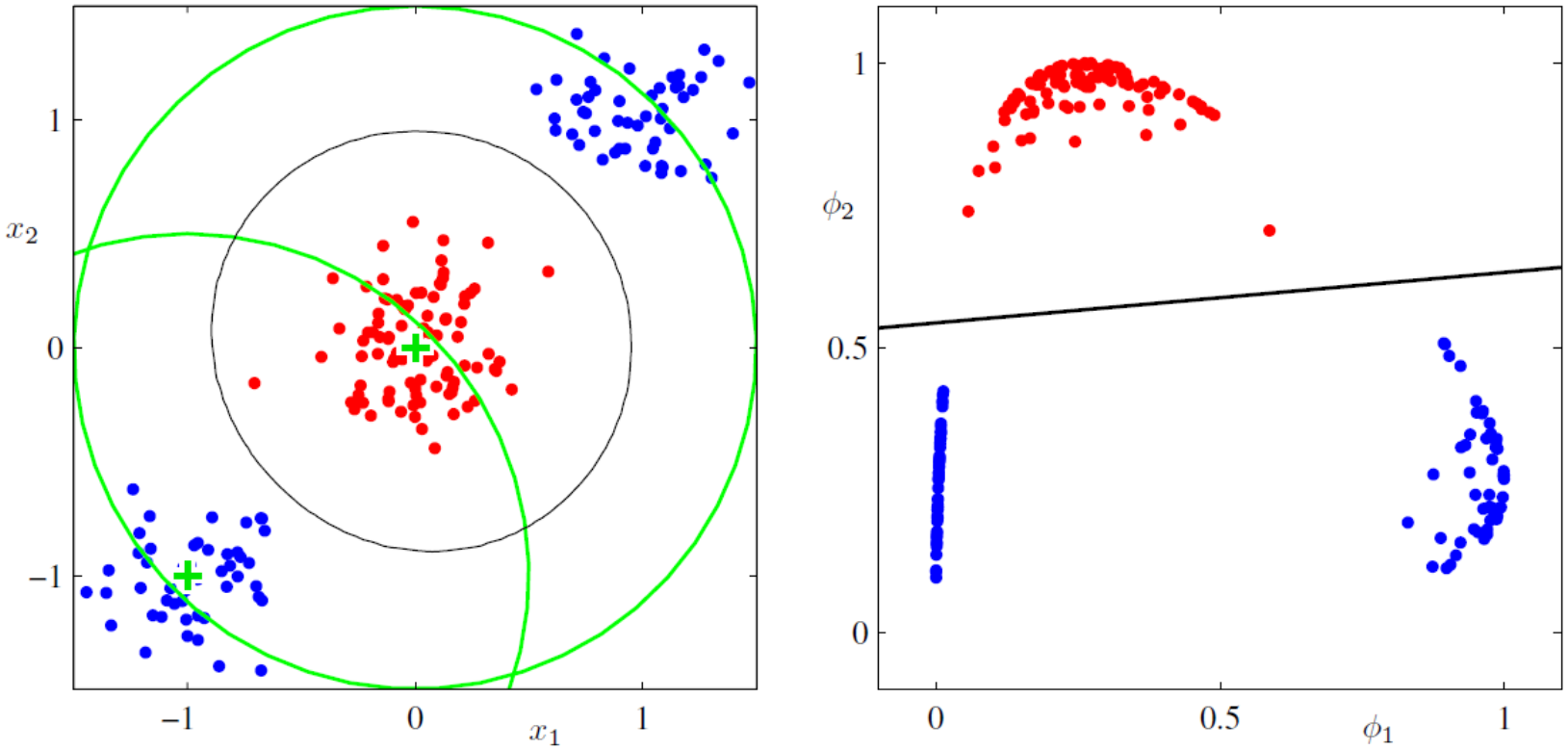
ϕ为非线性核函数

其中，

$$\mathbf{S}_k = \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

思考：根据二分类极大似然解的模型构建思路，写出K分类情况下的极大似然解的分类模型。

概率判别模型：非线性核函数的作用



- 左图：原始输入空间 (x_1,x_2) 以及来自两个类别的数据点（红色和蓝色）。在这个空间中定义了两个高斯基函数 $\phi_1(x)$ 和 $\phi_2(x)$，它们的中心位置用绿色十字标记，轮廓用绿色圆圈表示。

- 右图：对应的变换后的特征空间 (ϕ_1,ϕ_2) 以及通过**逻辑回归模型**得到的线性决策边界。对应于原始输入空间中的一个非线性决策边界，如左图中的黑色曲线所示。

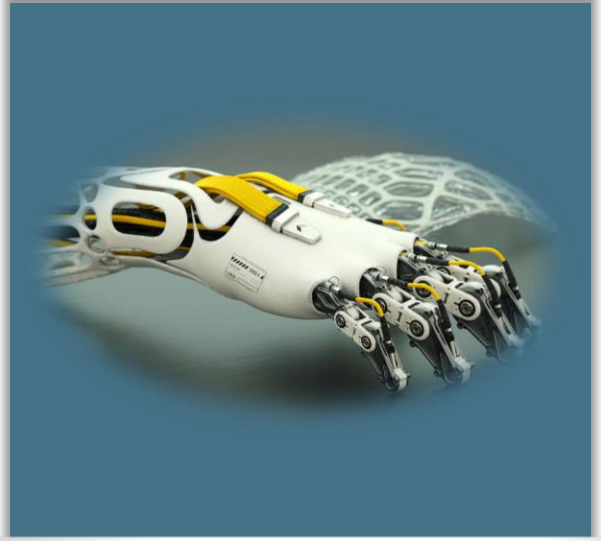
阶段性总结：两种线性分类模型构建方法的对比

- **判别方法**（Discriminative approach）和 **生成方法**（Generative approach）

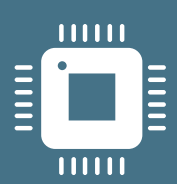
所学到的模型分别称为

- **判别模型**（Discriminative Model）和 **生成模型**（Generative Model）

	判别模型（Discriminative Model）	生成模型（Generative Model）
方法	由数据直接学习决策函数$C=f(X)$或者条件概率分布$P(C\|X)$作为预测的模型，即判别模型。	由训练数据学习联合概率分布 $P(X,C)$，然后求得后验概率分布$P(C\|X)$。
基本思想	在有限样本条件下建立判别函数，不考虑样本的产生模型，直接研究预测模型。 即：直接估计$P(C\|X)$	利用训练数据和对$P(X\|C)$和$P(C)$的估计，得到联合概率分布： $P(X,C)=P(C)P(X\|C)$，再根据它计算后验概率$P(C\|X)$进行分类。 即：估计$P(X\|C)$ 然后推导$P(C\|X)$
归属模型	线性回归、逻辑回归、感知机、决策树、支持向量机……	朴素贝叶斯、HMM、深度信念网络



统计机器学习中的分类问题




1. 判别函数
2. 概率生成模型和判别模型
3. 贝叶斯逻辑回归（Logistic Regression）
4. 分类模型的评价指标

逻辑回归（Logistic Regression）

- **（二分类情况下）将后验概率假设为逻辑Sigmoid函数的形式：**
  - 后验概率：$p(\mathcal{C}_1|\boldsymbol{\phi}) = y(\boldsymbol{\phi}) = \sigma\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\right) \qquad p(\mathcal{C}_2|\boldsymbol{\phi}) = 1 - p(\mathcal{C}_1|\boldsymbol{\phi})$

    其中 $\sigma(\cdot)$ 是**logistic sigmoid** 函数；$\boldsymbol{\phi}$ 是特征向量
  - 对于一个M-维特征空间$\boldsymbol{\phi}$，模型含有 M 个可调参数
    - 对比而言，最大似然估计模型含有2M个关于均值的参数，M(M+1)/2个和个类别共用协方差矩阵相关的参数，以及和先验概率相关的1个参数；
  - 对于数据集 $\{\boldsymbol{\phi}_n, t_n\}$ 而言，我们有 $t_n \in \{0,1\}$，以及 $\boldsymbol{\phi}_n = \boldsymbol{\phi}(\mathbf{x}_n)$，则似然函数为：

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}\{1-y_n\}^{1-t_n} \quad \text{其中} \quad \mathbf{t} = (t_1,\ldots,t_N)^{\mathrm{T}} \quad y_n = p(\mathcal{C}_1|\boldsymbol{\phi}_n)$$

  - 将对数似然函数取负，得到如下误差函数：

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$$

此即著名的“交叉熵函数”

逻辑回归（续）

- **逻辑回归的损失函数：**

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$$

- **对损失函数求梯度，有**

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n$$

  - 损失函数为凸函数，因此可以引入“Newton-Raphson”二阶迭代算法。

- **二阶最优化迭代法求最小二乘解**
  - 更新公式：$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - \mathbf{H}^{-1}\nabla E(\mathbf{w})$ 其中
    $\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n = \boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t})$
    Hessian矩阵 $\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}} = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}$
  - 上面右边式中，$\mathbf{R}$为$N \times N$对角线矩阵：

$$R_{nn} = y_n(1-y_n)$$

逻辑回归（续）

- **Newton-Raphson 迭代公式：**

$$\begin{aligned}\mathbf{w}^{(\text{new})} &= \mathbf{w}^{(\text{old})} - (\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t}) \\ &= (\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi})^{-1}\left\{\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi}\mathbf{w}^{(\text{old})} - \mathbf{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t})\right\} \\ &= (\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{z}\end{aligned}$$

  - 其中，$\mathbf{z} = \mathbf{\Phi}\mathbf{w}^{(\text{old})} - \mathbf{R}^{-1}(\mathbf{y}-\mathbf{t})$

由此展开

$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - \mathbf{H}^{-1}\nabla E(\mathbf{w})$$

$$\begin{cases}\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n = \mathbf{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t}) \\ \mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi}\end{cases}$$

- **二阶最优化迭代法：Newton-Raphson**

一阶和二阶方法对比	梯度下降	Newton-Raphson
收敛速度	线性收敛	二次收敛
计算成本	低（仅梯度）	高（需Hessian矩阵）
驻点搜索的理论基础	基于函数一阶泰勒展开	基于函数二阶泰勒展开
适用场景	大规模非凸问题	小/中规模凸问题（对非凸问题效果很差）

多变量逻辑回归

- **（K分类情况下）将后验概率假设为Softmax函数的形式：**
  - 后验概率对每个类有：$p(\mathcal{C}_k|\boldsymbol{\phi}) = y_k(\boldsymbol{\phi}) = \dfrac{\exp(a_k)}{\sum_j \exp(a_j)}$　其中激活函数设定为　$a_k = \mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}$

  - 对训练集数据，有似然函数（这里，$y_{nk} = y_k(\boldsymbol{\phi}_n)$）：

$$p(\mathbf{T}|\mathbf{w}_1,\ldots,\mathbf{w}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(\mathcal{C}_k|\boldsymbol{\phi}_n)^{t_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

注意：对K分类情况，使用独热编码（One-hot coding）：
$\mathbf{t_n} = [t_{n1}, t_{n2}, \ldots, t_{nK}]^T$，　$t_{n1} \in \{0,1\}$

  - 将对数似然函数取负，得到如下误差损失函数：

$$E(\mathbf{w}_1,\ldots,\mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{w}_1,\ldots,\mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln y_{nk}$$

“交叉熵函数”

  - 使用梯度下降法（一阶方法）的话，有

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1,\ldots,\mathbf{w}_K) = \sum_{n=1}^{N}(y_{nj} - t_{nj})\,\boldsymbol{\phi}_n \quad \Longrightarrow \quad \mathbf{w}_j^{t+1} = \mathbf{w}_j^{t} - \eta\nabla_{\mathbf{w}_j} E$$

梯度下降法

二分类问题的Bayes逻辑回归

- **贝叶斯原理下的后验概率推断：**
  - 后验概率：现有数据集（代表似然函数）和模型假设（代表先验参数分布）下的参数分布概率[1]：

公式（a） $$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{w})p(\mathbf{t}|\mathbf{w})$$ 此处$\mathbf{t} = [t_1, \ldots, t_n]^T$

- **模型假设**
  - 先验参数分布模型为高斯分布： $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$
  - 似然函数采用二项分布形式：

（回忆）多元（D-元）变量高斯分布：

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma (\mathbf{x}-\boldsymbol{\mu})\right\}$$

此处$\mathbf{t} = [t_1, \ldots, t_n]^T$ 其中 $y_n = \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n)$，使用逻辑sigmoid激活函数

- **最大后验估计（maximum posterior，MAP）**
  - 对公式（a）中的后验概率取对数有

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) + \sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} + \text{const}$$

1. 参见第三讲线性回归模型中的贝叶斯回归部分。

二分类问题的Bayes逻辑回归（续）

- **最大后验估计（Maximum a Posterior，MAP）**
  - 后验概率对数：$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) + \sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} + \text{const}$ 公式(a)
  - 最终目的：通过下述高斯分布近似真实的后验分布$p(\mathbf{w}|\mathbf{t})$

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{S}_N)$$

其中，均值$\mathbf{W}_{\mathrm{MAP}}$通过最大化后验概率（MAP）估计得到（在这里可使用**梯度上升法**）；

协方差 $\mathbf{S}_N$：通过后验分布的曲率（二阶导数）确定。

- 在$\mathbf{W}_{\mathrm{MAP}}$处对$\ln p(\mathbf{w}|\mathbf{t})$进行泰勒展开，

公式（b） $\ln p(\mathbf{w}|\mathbf{t}) \approx \ln p(\mathbf{w}_{\mathrm{MAP}}|\mathbf{t}) - \frac{1}{2}(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}})^{\top}\mathbf{H}(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}})$ 其中 $\mathbf{H} = -\nabla\nabla \ln p(\mathbf{w}|\mathbf{t})\big|_{\mathbf{w}=\mathbf{w}_{\mathrm{MAP}}}$

- 对比公式（a）与（b）则有近似高斯分布的协方差矩阵由 Hessian 矩阵的逆给出

$$\mathbf{S}_N = -\nabla\nabla \ln p(\mathbf{w}|\mathbf{t}) = \mathbf{S}_0^{-1} + \sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}.$$

由先验部分的Hessian和似然部分的Hessian组成

二分类问题的Bayes逻辑回归（续）

- **基于最大后验估计（MAP）预测新样本x的类别概率（施加特征变换$\phi(\mathbf{x})$后）**

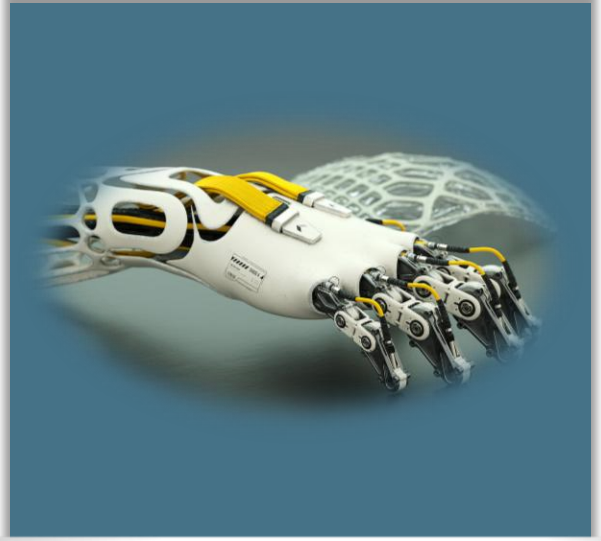
$$\begin{cases} p(\mathcal{C}_1|\boldsymbol{\phi},\mathbf{t}) = \int p(\mathcal{C}_1|\boldsymbol{\phi},\mathbf{w})p(\mathbf{w}|\mathbf{t})\,\mathrm{d}\mathbf{w} \simeq \int \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})q(\mathbf{w})\,\mathrm{d}\mathbf{w} \\ p(\mathcal{C}_2|\boldsymbol{\phi},\mathbf{t}) = 1 - p(\mathcal{C}_1|\boldsymbol{\phi},\mathbf{t}) \end{cases}$$

  - 由于积分困难，可采取点估计近似：$p(C_1|\phi,\mathbf{t}) \approx \sigma(\mathbf{w}_{\mathrm{MAP}}^{T}\phi_{\mathrm{new}})$　　其中 $\phi_{\mathrm{new}} = \phi(\mathbf{x})$

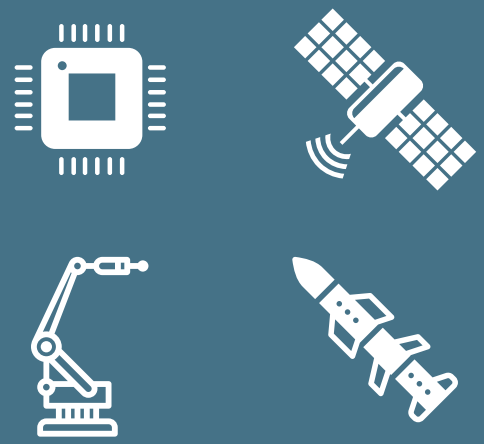
- **Probit近似**

  - 利用高斯分布的性质，将 Sigmoid 函数替换为 Probit 函数，得到解析近似：

$$p(t_{\mathrm{new}} = 1) \approx \sigma\left(\frac{\mathbf{w}_{\mathrm{MAP}}^{\top}\boldsymbol{\phi}_{\mathrm{new}}}{\sqrt{1 + \boldsymbol{\phi}_{\mathrm{new}}^{\top}\mathbf{S}_N\boldsymbol{\phi}_{\mathrm{new}}}}\right)$$



统计机器学习中的分类问题



1. 判别函数
2. 概率生成模型和判别模型
3. 贝叶斯逻辑回归（Logistic Regression）
4. 分类模型的评价指标

准确率（Accuracy）

- **准确率表示模型正确分类的样本数占总样本数的比例。**
- **混淆矩阵**
  - 混淆矩阵也称误差矩阵，用于评价算法或者分类器的结果。
  - 混淆矩阵式一个方阵：每一列代表了预测类别，每一列的总数表示预测为该类别的数据的数目；每一行代表了数据的真实归属类别，每一行的总数表示该类别的数据实例的数目；每一列中的数值表示真实数据被预测为该类的数目
- **准确率**
  - 二分类问题的准确率：$\frac{\text{正确分类的样本数}}{\text{总样本数}}$ $\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$
  - K分类问题的准确率=$\frac{\sum_{i=1}^{K} \text{TP}_i}{\text{总样本数}}$

精确率（查准率）和召回率（查全率）

- **二分类问题中数据类别分为阳性（1：Positive）和阴性（-1：Negative）**

	真实值为+1	真实值为-1
预测值为+1	真阳性：True Positive （TP）	假阳性：False Positive （FP）
预测值为-1	假阴性：False Negative（FN）	真阴性：True Negative（TN）

  - **精确率（Precision）：TP / (TP + FP)，即预测类别=+1时，真实类别=+1的条件概率；**
  - **召回率（Recall）：TP/ (TP + FN)，即真实类别=+1时，预测类别=+1的条件概率。**

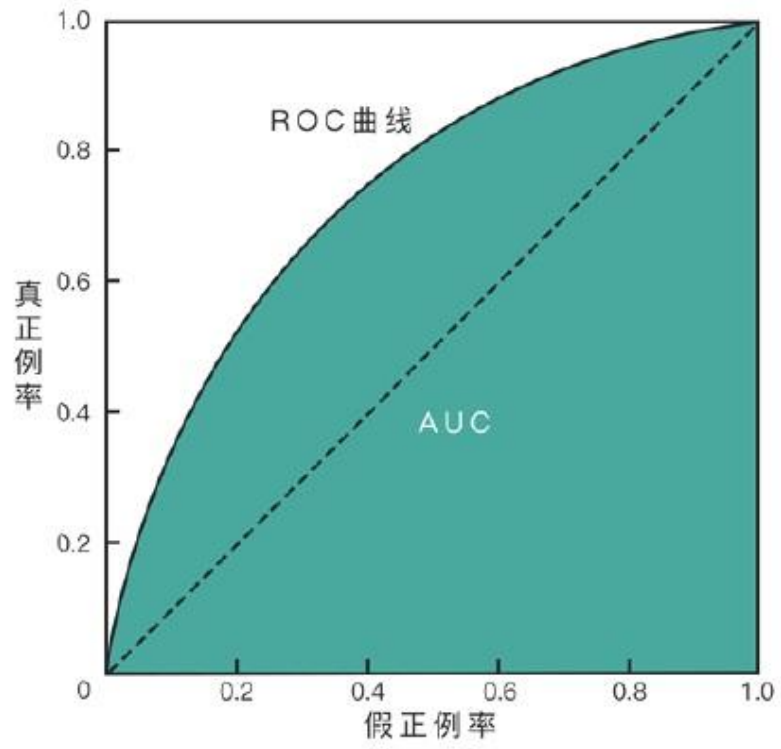
- **多分类问题的查准率**
  - 对某一类别而言：$\text{Precision}_{\text{i}} = \frac{TP_i}{TP_i+FP_i} = \frac{\text{正确预测为该类的样本数}}{\text{所有预测为该类的样本数}}$，
  - 对所有类别而言：平均查准率$=\frac{1}{K}\sum_{\text{i}=1}^{K} \text{Precision}n_i$

F1 Score

- **F1分数：调和查准率和查全率的加权平均，更适合类别不平衡的任务。**
  - 变化范围在0-1 之间。
- **定义式**
  - 二分类问题：$\mathrm{F1} = \frac{2TP}{2TP+FN+FP} = \frac{2 \bullet \mathrm{Precision} \bullet \mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}$
  - 多分类问题：
    - 按类别计算F1 Score：$\mathrm{F1_i} = \frac{2 \bullet \mathrm{Precision_i} \bullet \mathrm{Recall_i}}{\mathrm{Precision_i}+\mathrm{Recall_i}}$
    - 按宏平均计算：$\mathrm{F1} = \frac{1}{\mathrm{K}} \sum_{\mathrm{i}=1}^{K} \mathrm{F1}_i$

ROC曲线（Receiver Operating Characteristic Curve）

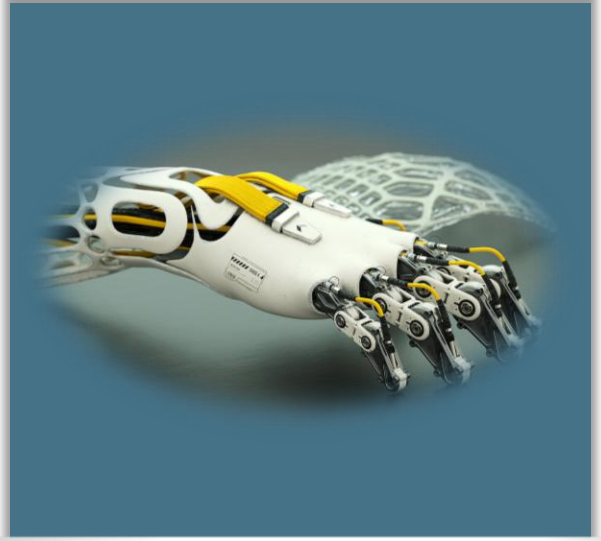
- **ROC全称是“受试者工作特征”（Receiver Operating Characteristic）曲线**
  - 用于描述混淆矩阵中FPR（False Positive Rate）-TPR（True Positive Rate）两个量之间的相对变化情况，即，用于描述样本的真实类别和预测概率相对关系。
  - 横轴： FPR，False Positiv $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = \frac{\text{误判为负的负类样本}}{\text{所有负类样本}}$
  - 纵轴：TPR，True Positive $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \frac{\text{正确预测的正类样本}}{\text{所有正类样本}}$



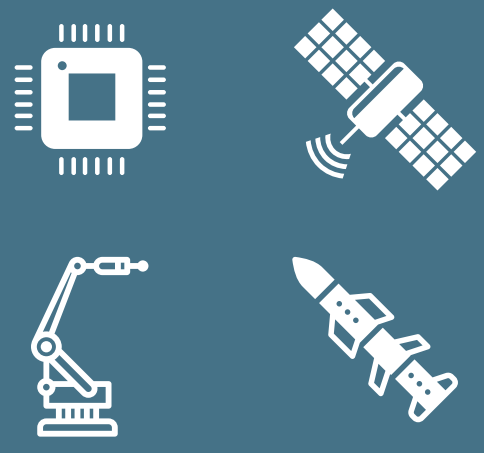
AUC（Area Under the ROC Curve） 面积

- **AUC是 ROC 曲线下方的面积，用于量化二分类模型的整体分类性能**
  - 回顾：ROC曲线：以 假正类率（FPR） 为横轴，真正类率（TPR） 为纵轴绘制的曲线。
  - 定义：ROC曲线与横轴围成的面积，取值范围在 [0, 1] 之间。
    - AUC=1：模型完美区分正负类（ROC曲线为左上角直角）。
    - AUC=0.5：模型无区分能力（等同于随机猜测，ROC曲线为对角线）。
    - AUC>0.5：模型优于随机猜测，值越大性能越好。
  - 注意：AUC只能用于评价二分类。

指标	AUC	准确率（ACC）
关注点	正负类排序能力	整体预测正确率
阈值	与阈值无关	依赖特定阈值（与任务相关）
类别均衡	不受类别分布影响	易被多数类主导，导致虚高
适用场景	需全面评估模型	类别均衡且阈值明确



统计机器学习中的分类问题



附录：数据处理

数据预处理

- **识别和处理数据集中的缺失值**
  - 1. 删除有缺失值的训练样本，或特征列（维度）；
  - 2. 填补缺失值，例如使用均值插补（Mean Imputation），即用整个特征列的平均值替换缺失值
    - 例子：使用scikit-learn的函数：SimpleImputer(missing_values=np.nan, strategy='mean')。
- **为分类标签编码**
  - 整数编码：将原始训练集（字典）中的分类标签转换成整数
    - 注意：分类标签并不是有序的，具体匹配哪个整数无关紧要。但是整数的有序性会影响机器学习模型的训练结果。
  - 为名义特征做独热编码（one-hot encoding）：
    - 为标签值的每一个唯一值创建一个新的二元虚拟特征
      - 例子：对‘color’特征，转换成blue、green和red三个虚拟特征，然后用特定二进制值表示每个特征，如一个blue样本可以表达为：blue=1，green=0，red=0。

特征缩放

- **目的：避免某一个特征维度在训练过程中起主导作用**
  - 原理：消除量纲差异，从而提升模型训练效率；
  - 注意：只有决策树和随机森林两种算法是机器学习方法中不需要特征缩放的；
  - 对以梯度下降为基础的方法，特征缩较不缩放放表现更佳。
- **方法1：归一化**
  - 强制特征缩放到[0,1]区间，对每个特征列i，应用最大-最小缩放

$$x_i^{norm} = {x_i - x_i^{\min}} \Big/ {x_i^{\max} - x_i^{\min}}$$

- **方法2：标准化**
  - 保留分布形状，把特征的中心点设在均值为0且标准差为1的位置

$$x_i^{\mathrm{std}} = {x_i - \mu_{\mathrm{i}}} \Big/ {\sigma_{x_i}}$$

训练样本集的划分

- **数据集的划分**
  - **训练集**（Training Set）：基于训练集的数据，确定模型的参数；
  - **验证集**（Validation Set）：也叫做开发集（Dev Set），用来做模型选择（Model Selection），用于模型的最终优化及确定，用来辅助模型的构建，即训练超参数，可选；
  - **测试集**（Test Set）： 用于测试已经训练好的模型的精确度；

- **训练集-测试集的随机划分**
  - 在data_loader模块，通过调用或实现划分函数，完成对样本数据集的划分
    - 举例：在scikit-learn库中，调用model_selection中的train_test_split()函数，进行分层抽样（stratify=y）：

```
train_test_split(X, y, test_size=0.3， stratify=y) ##确保划分后的数据集同原始数据集有相同的类别比例；
```

模型训练中的正则化

- **回顾：过拟合的原因在于，与给定的训练数据相比，模型太过复杂，因此可以通过**
  - 收集更多数据；
  - 引入正则化对复杂性的惩罚项；
  - 选择参数较小的模型；
  - 降低数据的维数。

- **L1和L2正则化**
  - 对参数添加惩罚项，抑制过大权重

特性	L1正则化（Lasso）	L2正则化（Ridge）
数学形式	$\lambda\sum\|w\|$	$\lambda\sum w^{T}w$
解的性质	稀疏解（自动特征选择）	非稀疏解（平滑权重分布）
抗噪声能力	较强（抑制无关特征）	较弱（保留所有特征但缩小权重）
适用场景	高维特征选择（如基因数据）	防止过拟合（通用场景）

费舍线性判别函数（Fisher's Discriminator）

- **基本思想：找到将高维样本投影到低维空间权重向量w（投影方向），并最大化不同类的样本投影集合的类间距离，最小化同类样本投影的集合内差异**
  - 举例：对二分类问题，在未做降维投影的情况下，C_1和C_2两类样本子集各有N_1，N_2个样本；
  - 两个样本子集的平均向量为 $\mathbf{m}_1 = \frac{1}{N_1}\sum_{n\in C_1}\mathbf{x}_n$ $\mathbf{m}_2 = \frac{1}{N_2}\sum_{n\in C_2}\mathbf{x}_n$
  - 最大化平均向量的低维投影距离： $m_2 - m_1 = \mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1)$ ➡ $\mathbf{w} \propto (\mathbf{m}_2 - \mathbf{m}_1)$
  - 类别 C_k的类内投影方差为： $s_k^2 = \sum_{n\in C_k}(y_n - m_k)^2$ 其中 $y_n = \mathbf{w}^{\mathrm{T}}\mathbf{x}_n$
  - Fisher判别准则（目标函数）：类间方差（between-class variance）：类内方差（within-class variance）
  - 对二分类问题有： $J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$ 写成矩阵形式 ➡ $J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}}$ 最大广义瑞利商
  - 类间协方差矩阵（between-class covariance matrix）： $\mathbf{S}_{\mathrm{B}} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}}$
  - 总类内协方差矩阵： $\mathbf{S}_{\mathrm{W}} = \sum_{n\in C_1}(\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathrm{T}} + \sum_{n\in C_2}(\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathrm{T}}$

费舍线性判别函数（续）

- **（二分类问题）基于最大化Fisher判别准则的权重向量的解：**
  - 对目标函数 $J(\mathbf{w}) = \dfrac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}}$ 求导，得

$$(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w})\mathbf{S}_{\mathrm{W}}\mathbf{w} = (\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w})\mathbf{S}_{\mathrm{B}}\mathbf{w}.$$

  - 由类间协方差矩阵： $\mathbf{S}_{\mathrm{B}} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}}$ 知$\boldsymbol{S}_{\mathrm{B}}\mathbf{w}$总是沿（$\mathrm{m}_2$- m_1）的方向
  - 由于w是投影方向的法向量，我们只关心w的方向，不关心w的向量长度，则**上式二项式部分可以忽略**：

$$\mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

Fisher's linear discriminant

- **K类别情况下Fisher判别的解：**
  - 总类内协方差矩阵： $\mathbf{S}_{\mathrm{W}} = \sum_{k=1}^{K}\mathbf{S}_k$ 其中 $\mathbf{S}_k = \sum_{n\in\mathcal{C}_k}(\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^{\mathrm{T}}$ $\mathbf{m}_k = \dfrac{1}{N_k}\sum_{n\in\mathcal{C}_k}\mathbf{x}_n$
  - 类间协方差矩阵：

$$\mathbf{S}_{\mathrm{B}} = \sum_{k=1}^{K} N_k(\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^{\mathrm{T}}$$

  - K类别Fisher判别准则（目标函数）：$J(\mathbf{w}) = \mathrm{Tr}\left\{(\mathbf{W}\mathbf{S}_{\mathrm{W}}\mathbf{W}^{\mathrm{T}})^{-1}(\mathbf{W}\mathbf{S}_{\mathrm{B}}\mathbf{W}^{\mathrm{T}})\right\}$ → $S_W^{-1}S_BW = \lambda W$

取最大特征值对应的特征向量

费舍线性判别函数的另一种解释：

- 由于我们不关心w的长短，令$\mathbf{w}^T\mathbf{S}_\mathrm{w}\mathbf{w}=1$，最大化广义瑞利商等价于

$$\min_{w} \ -\mathbf{w}^T\mathbf{S}_B\mathbf{w}$$
$$s.t. \quad \mathbf{w}^T\mathbf{S}_W\mathbf{w}=1$$

- 引入拉格朗日乘子，有如下拉格朗日函数

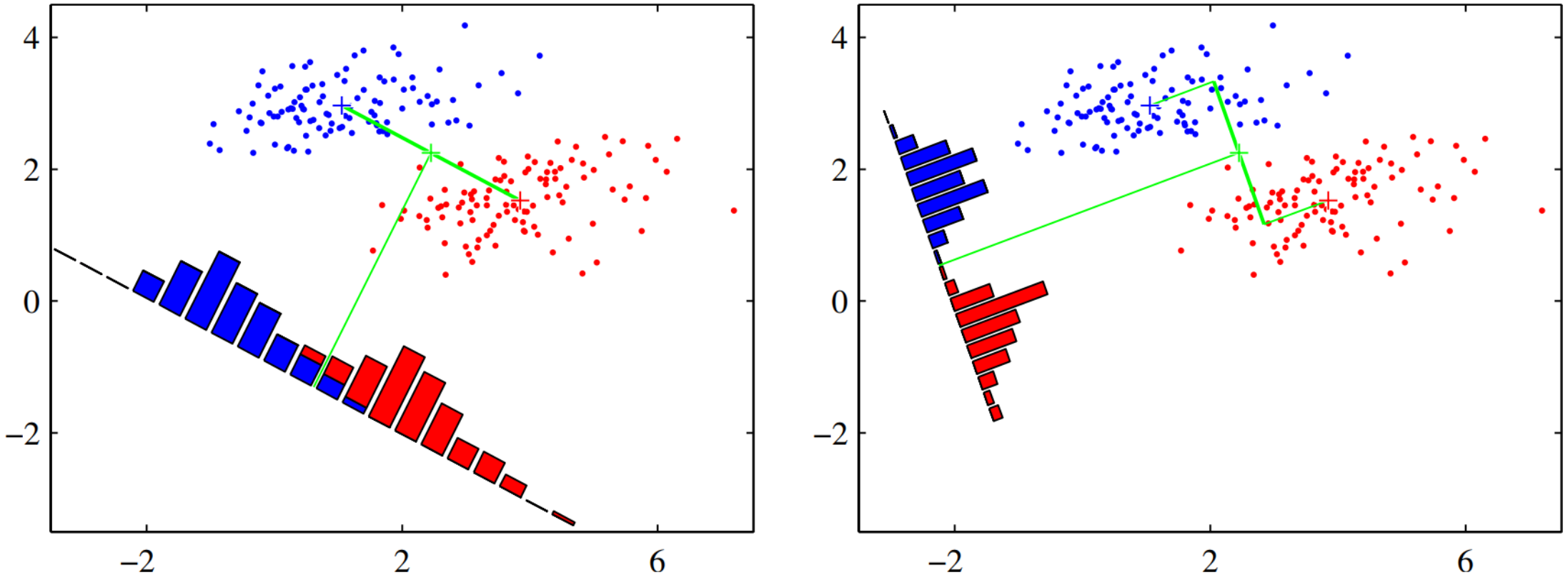
$$L(\mathbf{w},\lambda)=-\mathbf{w}^\mathbf{T}\mathbf{S}_\mathrm{B}\mathbf{w}+\lambda(\mathbf{w}^\mathbf{T}\mathbf{S}_\mathbf{w}\mathbf{w}-1)$$

- 对w求导，有 $0=\frac{\partial L(\mathbf{w},\lambda)}{\partial \mathbf{w}}=-(\mathbf{S}_\mathrm{B}+\mathbf{S}_B{}^T)\mathbf{w}+\lambda(\mathbf{S}_\mathbf{w}+\mathbf{S}_\mathbf{w}^\mathbf{T})\mathbf{w}$ ⟶ $\mathbf{S}_\mathbf{B}\mathbf{w}=\lambda\mathbf{S}_\mathbf{w}\mathbf{w}$

由$\mathbf{S}_\mathbf{B}\mathbf{w}$定义，有 $\mathbf{S}_\mathbf{B}\mathbf{w}=(\mathbf{m_2}-\mathbf{m_1})\underbrace{(\mathbf{m_2}-\mathbf{m_1})^T\mathbf{w}}$ ⟶ $\mathbf{w}=\mathbf{S}_\mathbf{w}^{-1}(\mathbf{m_2}-\mathbf{m_1})$

标量，令其等于λ

费舍线性判别函数（续）



- 左图：两个类别的样本（分别用红色和蓝色表示），以及将样本投影到类均值连线上的直方图结果。
  - 注意：在投影空间中存在显著的类重叠。
- 右图：基于Fisher线性判别的投影结果，显示出类间分离程度得到了显著改善。