



LiVE Group
视觉智能与学习研究中心

机器学习 (第8讲)

主讲: 张磊

E-mail: leizhang@cqu.edu.cn
Lab Website: <http://www.leizhang.tk>





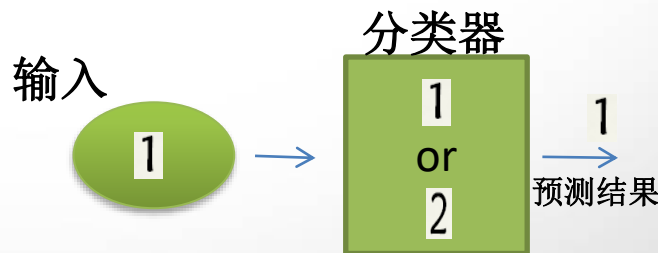
第八章： 分类器

第八章：分类器

□ 定义

- 对于一组训练样本集 x_1, x_2, \dots, x_N ，每个样本由 D 个属性组成，形成 D -维矢量。其中，每个样本分别采用一个标签 y_1, y_2, \dots, y_N 描述所属类别，样本集由 C 个类别组成。对于给定的新样本 x_{new} ，判定其所属的类别，被称为**分类**。
- 利用训练集生成的分类模型/方法/算法，被称为**分类器(Classifier)**。

分类问题是机器学习的核心问题，也是人工智能、模式识别、计算机视觉、图像理解、文本分类、数据挖掘等应用领域要解决的关键问题。





第八章：分类器

□ 分类器种类

机器学习发展至今，衍生出大量的分类器模型，本课程主要介绍以下几类分类器：

➤ 概率分类器（参见前面的章节）

贝叶斯分类器

朴素贝叶斯分类器

Logistic回归

➤ 非概率分类器

K-近邻分类器（欧式距离）

大间隔分类器（支持向量机）

线性判别分类器（投影）

神经网络分类器（多层感知器、BP网络、深度卷积网络—第九章讲）



第八章：分类器

不同于概率分类器，对于每个样本 \mathbf{x} ，计算 $P(y=c|\mathbf{x})$ ，提供类别 $y=c$ 的可能性，非概率分类器输出的是一个指定的类别，即 $y(\mathbf{x})=c$ 。

□ K-近邻分类器（K-nearest neighbors, KNN）

K-近邻分类器是基于欧式距离提出的较为直观的分类器方法，思想极其简单，不需要对训练集进行模型训练，完全无参数。

假设有 N 个训练样本，每个样本分别由属性 \mathbf{x} 和标记 \mathbf{y} 进行描述。对于一个新样本 \mathbf{x}_{new} ，如何利用KNN进行分类？

Step 1: 分别计算 \mathbf{x}_{new} 与训练集中 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ 之间的距离；

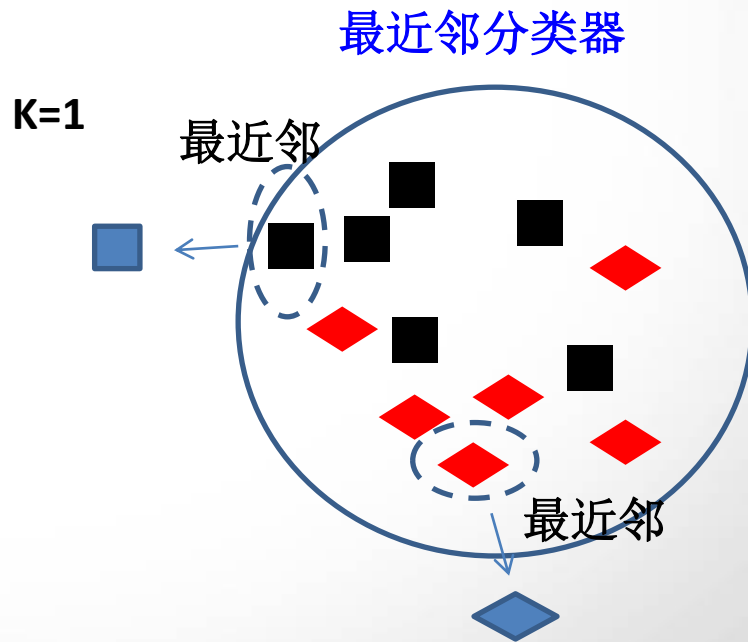
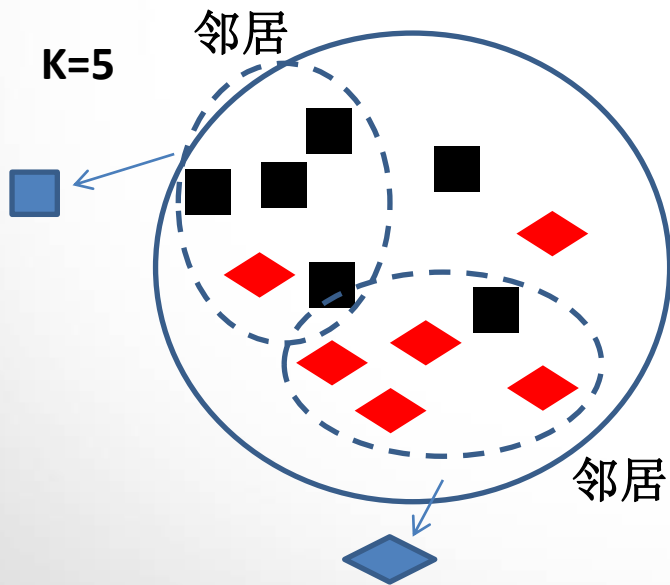
Step 2: 找出与 \mathbf{x}_{new} 最近的 K 个样本 \mathbf{x}_i ；

Step 3: 根据“服从多数”原则， \mathbf{x}_{new} 的类别即为 K 个样本中最多的类。

第八章：分类器

□ K-近邻分类器 (K-nearest neighbors, KNN)

KNN分类器的思想用一个图进行描述





第八章：分类器

□ K-近邻分类器

距离计算

样本 \mathbf{x}_i 和 \mathbf{x}_j 之间的欧式距离表达式

$$d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2 = \sqrt{\sum_{l=1}^d (x_i^l - x_j^l)^2}$$

K值的选择

- K值通常是一个奇数，即选取奇数个邻居点，从而可以通过投票确定类别；
- 如果K太小，分类器容易被噪声干扰；比如K=1时，最近邻点如果被噪声污染，则分类错误。
- 如果K太大，可能会永远无法进行正确的分类（各类样本不均衡问题），比如两类问题，第I类有10个样本，第II类有30个样本，如果K>21，则新样本将会一直被识别为第II类。



第八章：分类器

□ 其他距离

距离通常用于度量两个样本之间的相似性，在分类器中经常用到。

- ✓ 马氏距离（协方差）
- ✓ 曼哈顿距离（城市间街道）
- ✓ 切比雪夫距离（ $\max_l |x_i^l - x_j^l|$ ）
- ✓ 余弦距离（向量夹角 $\cos\theta = \mathbf{x}_i \cdot \mathbf{x}_j / |\mathbf{x}_i| |\mathbf{x}_j|$ ）
- ✓ 汉明距离（信息编码）
- ✓ 相关性（统计协方差 $\rho_{\mathbf{x}_i, \mathbf{x}_j}$ ）

□ 度量学习

度量学习不同于上述“已知”的距离表达式，而是通过机器学习建模的方式，学习一个最佳的距离度量，使得对特定的问题，具有最佳的识别性能。

Re-ID要解决的问题:

行人检测或人工标定



目标人物



待搜索视频帧



检测目标提取

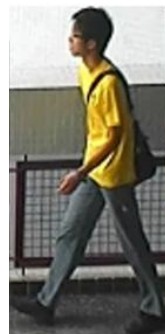
相似性匹配

Re-ID要解决的问题:

Camera #1:



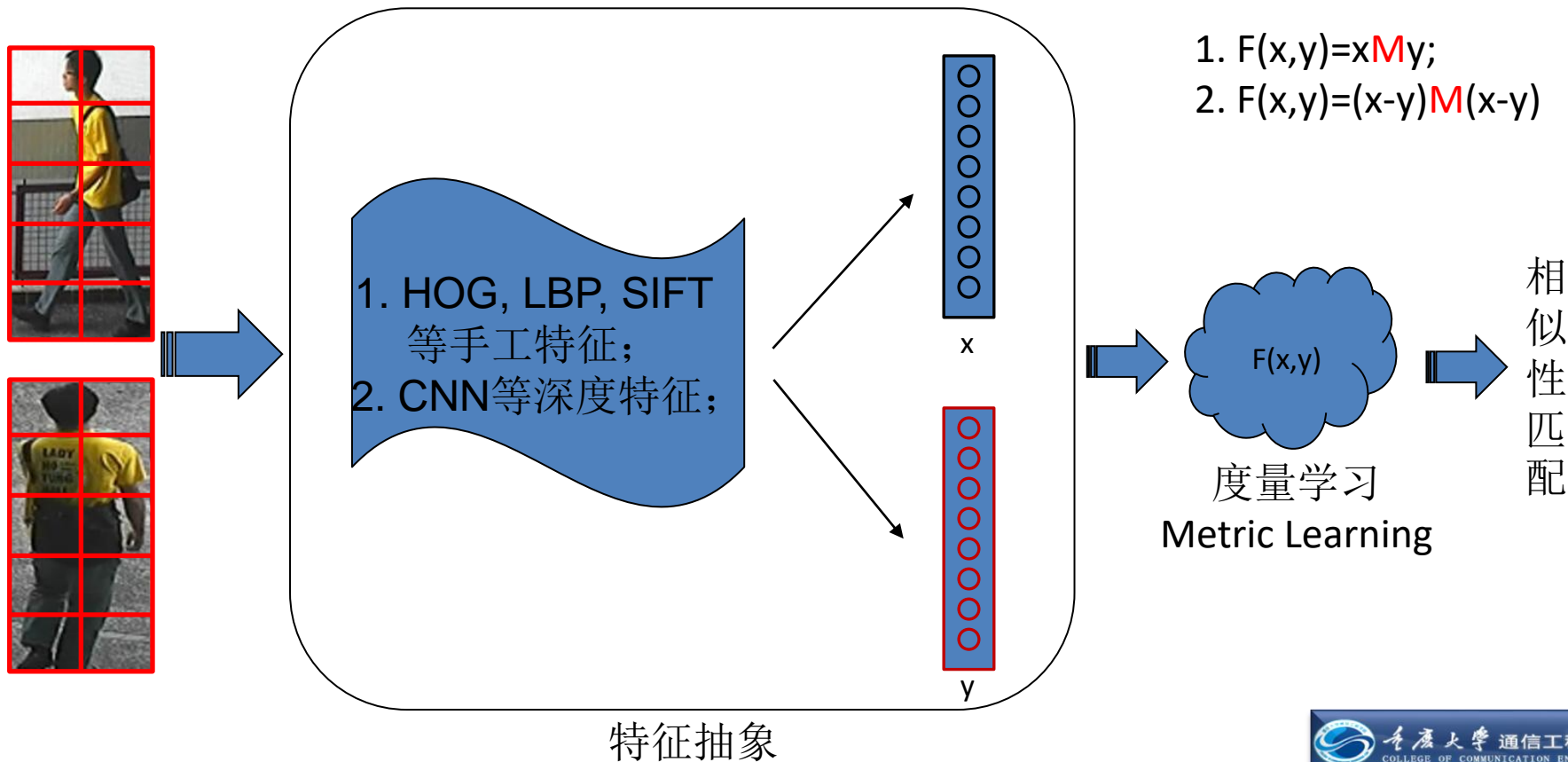
Camera #2:



特征抽象

距离判定

Re-ID要解决的问题:



Re-ID的度量学习原理:

三元组样本集构造: x, x_p, x_n



X

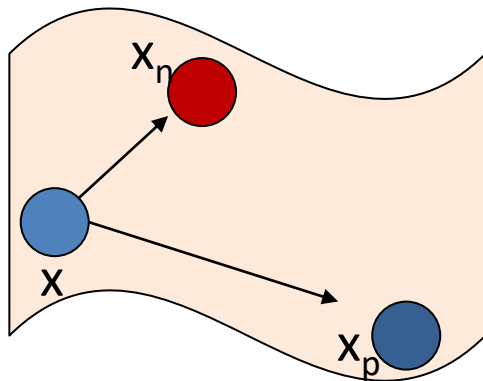


x_p

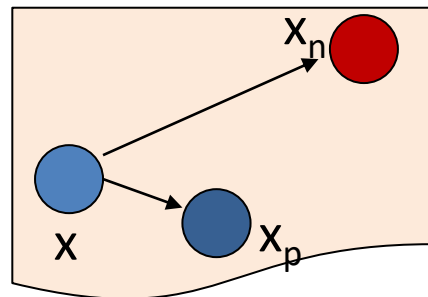


x_n

Metric learning



差的度量



好的度量

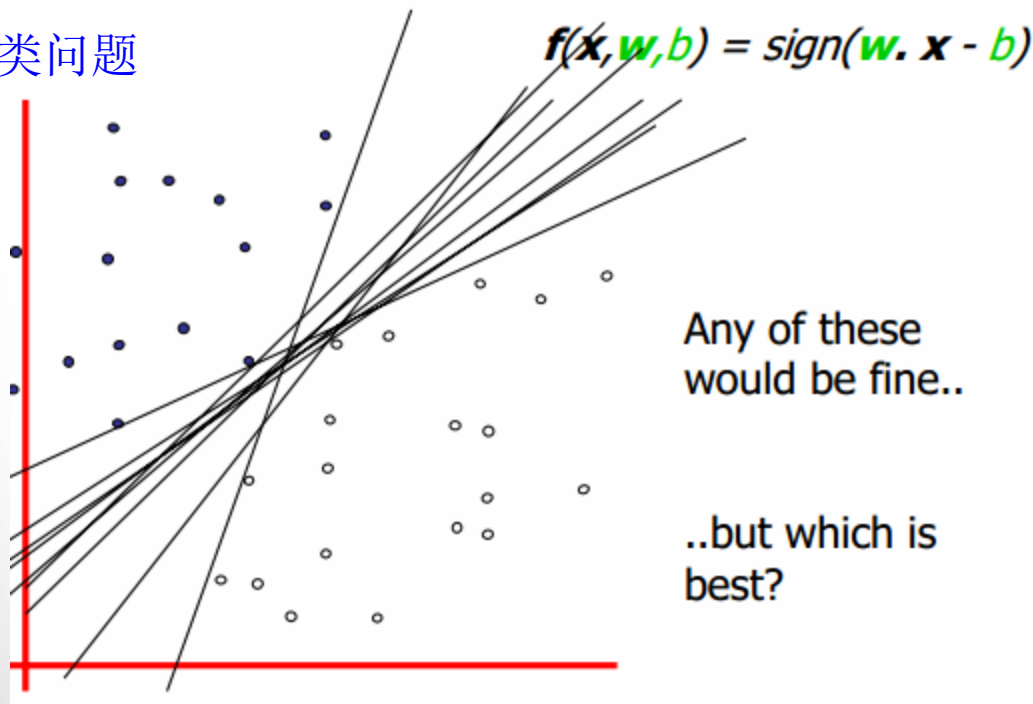
Triplet Loss(三元组):
$$\min \sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$$



第八章：分类器

□ 支持向量机（Support Vector Machine, SVM, Vapnik'95）

考虑线性分类问题



假设给定一个特征空间上的训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ ，其中， $x_i \in \mathbb{R}^n$ ， $y_i \in \{+1, -1\}$ ， $i = 1, 2, \dots, N$ 。 x_i 为第*i*个特征向量，也称为实例， y_i 为 x_i 的类标记；当 $y_i = +1$ 时，称 x_i 为正例；当 $y_i = -1$ 时，称 x_i 为负例。 (x_i, y_i) 称为样本点。



第八章：分类器

□ 支持向量机（Support Vector Machine, SVM）

▶ 线性可分支持向量机

定义： 给定线性可分训练数据集，通过间隔最大化或等价的求解相应的**凸二次规划问题**学习得到的分离超平面为

$$\mathbf{w}\mathbf{x} + \mathbf{b} = 0,$$

相应的分类决策函数

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}\mathbf{x} + \mathbf{b})$$

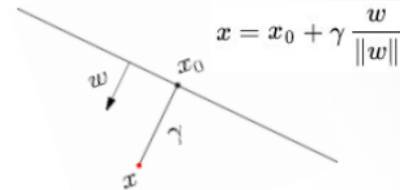
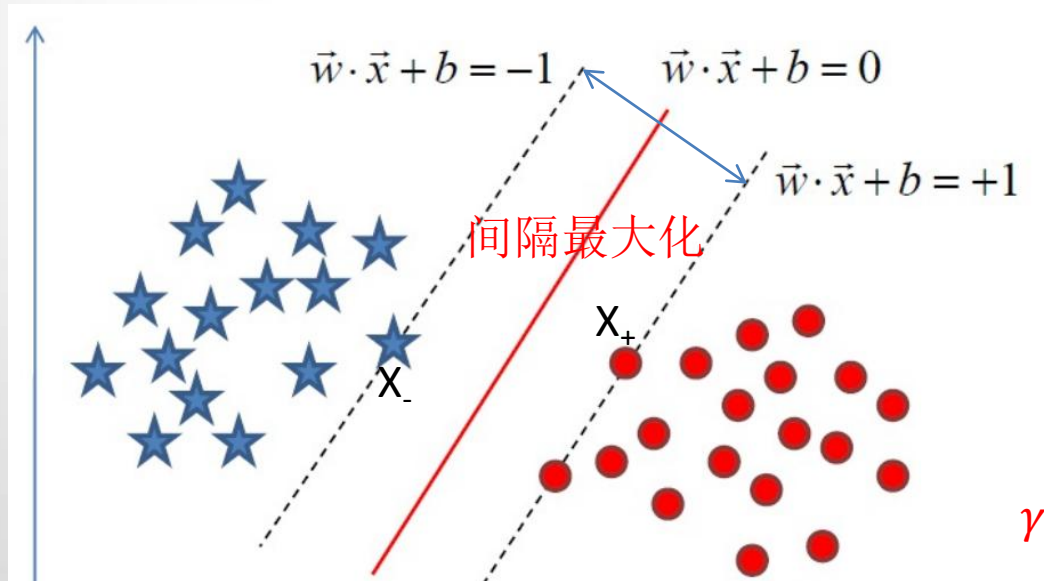
该决策函数称为线性可分支持向量机。

第八章：分类器



支持向量机 (Support Vector Machine, SVM)

线性可分支持向量机



定义：函数间隔和几何间隔

函数间隔：给定的训练数据集 T 和超平面 (w, b) ，定义超平面 (w, b) 关于样本点 (x_i, y_i) 的函数间隔为：

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

几何间隔：两个异类支持向量之差在法向量上的投影，即

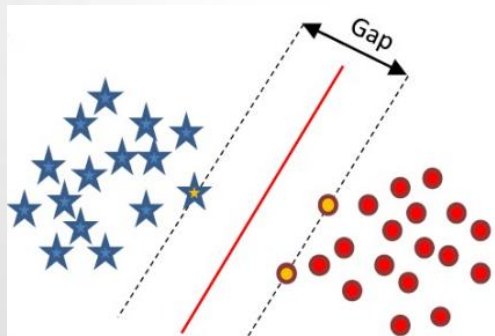
$$\gamma = \frac{|w^T x + b|}{\|w\|} = \frac{y(w^T x + b)}{\|w\|} = \frac{1}{\|w\|}$$

第八章：分类器

支持向量机（Support Vector Machine, SVM）

线性可分支持向量机

最大间隔分类超平面



$$\max_{w,b} \frac{1}{\|w\|}$$

$$s.t. \quad y_i(w \cdot x_i + b) \geq 1, \quad i = 1, 2, \dots, N$$



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

$$s.t. \quad y_i(w \cdot x_i + b) \geq 1, \quad i = 1, 2, \dots, N$$



第八章：分类器

□ 支持向量机（Support Vector Machine, SVM）

➤ 线性可分支持向量机

线性可分支持向量机模型

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} \|w\|^2 \\ \text{s.t.} \quad & y_i (w \cdot x_i + b) \geq 1, \quad i = 1, 2, \dots, N \end{aligned}$$

带有约束的凸二次规划问题。如何求解？



第八章：分类器

支持向量机 (Support Vector Machine, SVM)

线性可分支持向量机 (硬间隔最大化)

拉格朗日乘法

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

原问题的对偶问题：极大极小问题

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

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

先求内层最小值

将拉格朗日函数L(w,b,α)分别对w, b求偏导并令其为0:

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

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

补充:

$$\min f(x)$$

$$\text{s.t. } g(x) \leq 0$$

$$L(x,u) = f(x) + u g(x), u > 0$$

$$\text{所以 } f(x) = \max L(x,u)$$

依据：拉格朗日的对偶性，即广义拉格朗日的极小化问题，可以拆分成两步。



第八章：分类器

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

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

支持向量机 (Support Vector Machine, SVM)

线性可分支持向量机 (硬间隔最大化)

将上式带入拉格朗日函数 $L(\mathbf{w}, b, \alpha)$ 中，得到

$$\begin{aligned} \mathcal{L}(\mathbf{w}, b, \alpha) &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i [y_i (\mathbf{w}^T \mathbf{x}_i + b) - 1] \\ &= \frac{1}{2} \mathbf{w}^T \mathbf{w} - \mathbf{w}^T \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i - b \sum_{i=1}^n \alpha_i y_i + \sum_{i=1}^n \alpha_i \\ &= \frac{1}{2} \mathbf{w}^T \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i - \mathbf{w}^T \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i - b \cdot 0 + \sum_{i=1}^n \alpha_i \\ &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \left(\sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \right)^T \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \\ &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \end{aligned}$$

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



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

$$s.t. \alpha_i \geq 0, i = 1, 2, \dots, n$$

$$\sum_{i=1}^n \alpha_i y_i = 0$$



$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^N \alpha_i$$

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

$$\alpha_i \geq 0, i = 1, 2, \dots, N$$

KKT条件



第八章：分类器

□ 支持向量机（Support Vector Machine, SVM）

➤ 线性可分支持向量机（硬间隔最大化）

KKT条件（Karush[1939], Kuhn和Tucker[1951]），即**确定某点为极值的必要条件**。

$$\begin{aligned} \min f(x) \\ \text{s.t. } g(x) \leq 0 \end{aligned}$$

KKT条件包括**原可行性** $g(x^*) \leq 0$ ，**对偶可行性** $\alpha \geq 0$ 和**互补松弛性** $\alpha \cdot g(x^*) = 0$

SVM中的KKT条件：

$$y_i(wx_i + b) - 1 \geq 0,$$

$$\alpha_i \geq 0,$$

$$\alpha_i [y_i(wx_i + b) - 1] = 0 \longrightarrow \text{支持向量}$$



第八章：分类器

□ 支持向量机 (Support Vector Machine, SVM)

➤ 线性可分支持向量机 (硬间隔最大化)

1.模型:
$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^N \alpha_i$$

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

$$\alpha_i \geq 0, \quad i = 1, 2, \dots, N$$

2.求得最优解 α^*

3. 计算

$$w^* = \sum_{i=1}^N \alpha_i^* y_i x_i$$

$$b^* = y_i - \sum_{i=1}^N \alpha_i^* y_i (x_i \cdot x_j) \xrightarrow{\text{通常计算}} b^* = - \frac{\max_{y_i=-1} w^T x_i + \min_{y_i=1} w^T x_i}{2}$$

4.分离超平面

$$w^* x + b^* = 0$$

5.分类决策函数

$$f(x) = \text{sign}(w^* x + b^*)$$



第八章：分类器

□ 支持向量机（Support Vector Machine, SVM）

➤ 线性可分支持向量机（硬间隔最大化）

举例：给定3个数据点：正样本点 $x_1=(3,3)^T$ ， $x_2=(4,3)^T$ ，负样本点 $x_3=(1,1)^T$ ，利用线性可分支持向量机，求线性分类决策函数？

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^N \alpha_i \\ = \frac{1}{2} \quad & (18\alpha_1^2 + 25\alpha_2^2 + 2\alpha_3^2 + 42\alpha_1\alpha_2 - 12\alpha_1\alpha_3 - 14\alpha_2\alpha_3) - \alpha_1 - \alpha_2 - \alpha_3 \\ \text{s.t.} \quad & \alpha_1 + \alpha_2 - \alpha_3 = 0 \\ & \alpha_i \geq 0, \quad i = 1, 2, 3 \end{aligned}$$



第八章：分类器

□ 支持向量机（Support Vector Machine, SVM）

➤ 线性可分支持向量机（硬间隔最大化）

举例：给定3个数据点：正样本点 $x_1=(3,3)^T$ ， $x_2=(4,3)^T$ ，负样本点 $x_3=(1,1)^T$ ，利用线性可分支持向量机，求线性分类超平面？

✓ 将 $\alpha_1 + \alpha_2 = \alpha_3$ 带入目标函数，得到关于 α_1, α_2 的函数：

$$s(\alpha_1, \alpha_2) = 4\alpha_1^2 + \frac{13}{2}\alpha_2^2 + 10\alpha_1\alpha_2 - 2\alpha_1 - 2\alpha_2$$

✓ 对 α_1, α_2 求偏导并令其为0，易知 $s(\alpha_1, \alpha_2)$ 在点(1.5, -1)处取极值。

✓ 然而，该点不满足条件 $\alpha_2 \geq 0$ 。

当 $\alpha_1=0$ 时，最小值 $s(0, 2/13)=-2/13$

当 $\alpha_2=0$ 时，最小值 $s(1/4, 0)=-1/4$

✓ 于是， $s(\alpha_1, \alpha_2)$ 在 $\alpha_1=1/4$ ， $\alpha_2=0$ 时达到最小，此时， $\alpha_3 = \alpha_1 + \alpha_2 = 1/4$



第八章：分类器

□ 支持向量机（Support Vector Machine, SVM）

➤ 线性可分支持向量机（硬间隔最大化）

举例：给定3个数据点：正样本点 $x_1=(3,3)^T$ ， $x_2=(4,3)^T$ ，负样本点 $x_3=(1,1)^T$ ，利用线性可分支持向量机，求线性分类超平面？

- ✓ 可见， $\alpha_1=\alpha_3=1/4$ 对应的点 x_1, x_3 是支持向量。
- ✓ 带入公式，求 w 和 b ：

$$w^* = \sum_{i=1}^N \alpha_i^* y_i x_i$$

$$b^* = y_i - \sum_{i=1}^N \alpha_i^* y_i (x_i \cdot x_j)$$

得到 $w_1=w_2=0.5$ ， $b=-2$

✓ 分离决策函数为 $f(x) = \text{sign}\left(\frac{1}{2}x_1 + \frac{1}{2}x_2 - 2\right)$

✓ 因此，分离超平面为 $\frac{1}{2}x_1 + \frac{1}{2}x_2 - 2 = 0$

提问：对于 $\alpha > 0$ 时， $y(wx+b)-1=0$?



第八章：分类器

□ 支持向量机（Support Vector Machine, SVM）

➤ 线性支持向量机（软间隔最大化）

若数据近似线性可分，则增加松弛因子 $\xi_i \geq 0$ ，使函数间隔加上松弛变量大于等于1（软间隔）。这样，约束条件变成

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i$$

SVM的模型变成

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i$$

惩罚项

$$s.t. \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \dots, N$$

$$\xi_i \geq 0, \quad i = 1, 2, \dots, N$$

$$\xi_i = \max(0, 1 - y_i(w \cdot x_i + b))$$



第八章：分类器

□ 支持向量机（Support Vector Machine, SVM）

➤ 线性支持向量机（软间隔最大化）

1. 构造SVM模型

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^N \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^N \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, N \end{aligned}$$

2. 求解 α^*

SMO(Sequential Minimal Optimization)

算法求解, 参见文献

John C. Platt, Sequential Minimal Optimization:

A Fast Algorithm for Training Support Vector machines

3. 计算 w 和 b

$$\begin{aligned} w^* &= \sum_{i=1}^N \alpha_i^* y_i x_i \\ b^* &= y_i - \sum_{i=1}^N \alpha_i^* y_i (x_i \cdot x_j) \end{aligned}$$

4. 求得分离超平面

$$w^* x + b^* = 0$$

5. 分类决策函数

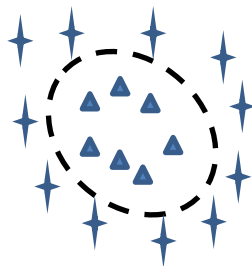
$$\begin{aligned} f(x) &= \text{sign}(w^* x + b^*) \\ &= \text{sign}\left(\sum_{i=1}^N \alpha_i^* y_i x_i x + b^*\right) \end{aligned}$$

第八章：分类器

□ 支持向量机 (Support Vector Machine, SVM)

▶ 非线性支持向量机 (核函数+软间隔最大化)

目的：将线性不可分的数据(信号)，通过一个非线性函数 $\varphi(\cdot)$ ，映射到高维特征空间，使得线性可分。



$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^N \alpha_i$$

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

$$0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, N$$

$$\varphi(x_i) \cdot \varphi(x_j)$$

$$k(x_i, x_j)$$

$$e^{-\sigma \|x_i - x_j\|^2}$$



第八章：分类器

□ 子空间学习（Subspace Learning）

➤ 概念

- ✓ 子空间学习是指对于一给定的分类任务，如何为原数据寻找一个低维子空间，使得在该低维子空间内实现更好的分类，而计算复杂度降低；
- ✓ 子空间学习的方法，一般通过投影的方式，实现高维特征（原数据空间）向低维子空间的映射。这个过程也被称为“降维”；

➤ 种类

- ✓ 线性子空间学习
- ✓ 非线性子空间学习

通常情况下，非线性子空间学习方法是通过对“核函数”进行建模实现。整个过程，是在线性子空间学习的方法上进行扩展。



第八章：分类器

□ 子空间学习（Subspace Learning）

▶ 线性子空间学习

定义原始高维特征集为 \mathbf{x} ，子空间学习从数学上是为了学习一个投影（或变换 \mathbf{W} ），使得特征集 \mathbf{x} 在低维子空间的表达形式为

$$\mathbf{Y} = \mathbf{W}^T \mathbf{X}$$

线性子空间学习的目的是为了获得变换矩阵 \mathbf{W} ，使得低维子空间表达 \mathbf{Y} 能更好的实现分类。

▶ 种类

- ✓ 主成分分析(PCA)；
- ✓ Fisher判别分析（线性判别分析，LDA）；
- ✓ 流形学习（局部保持投影, LPP）；



第八章：分类器

□ 子空间学习 (Subspace Learning)

▶ 主成分分析(PCA)

主成分分析是一种无监督的降维方法，在整个过程中，不需要知道样本集 \mathbf{X} 的标签。数学上，PCA是一种均方误差最小意义下的投影方法。

定义高维特征集 \mathbf{X} ，存在一个投影（或变换 \mathbf{W} ），即特征集 \mathbf{X} 在低维子空间的表达形式为

$$\mathbf{Y} = \mathbf{W}^T \mathbf{X}$$

使得 \mathbf{Y} 保留高维特征集 \mathbf{X} 中的大部分有效信息（能量最大或方差最大），即满足：

$$\max_{\mathbf{W}} \|\mathbf{Y} - \boldsymbol{\mu}\|_F^2 = \max_{\mathbf{W}} \|\mathbf{W}^T \mathbf{X} - \mathbf{W}^T \mathbf{m}\|_F^2$$

$$= \max_{\mathbf{W}} \text{Tr}(\mathbf{W}^T (\mathbf{X} - \mathbf{m})(\mathbf{X} - \mathbf{m})^T \mathbf{W})$$

$$= \max_{\mathbf{W}} \text{Tr}(\mathbf{W}^T \boldsymbol{\Sigma} \mathbf{W}), \text{ s. t. } \mathbf{W}^T \mathbf{W} = \mathbf{I}$$

\mathbf{W} 的求解采用特征值分解

$\boldsymbol{\Sigma} = (\mathbf{X} - \mathbf{m})(\mathbf{X} - \mathbf{m})^T$ 为协方差矩阵



第八章：分类器

□ 子空间学习 (Subspace Learning)

▶ K-L (Karhunen-Loève) 变换

与主成分分析原理相同，KL变换的本质是通过线性变换（正交变换）降低维度，也是均方误差最小意义下的投影方法。

定义一组向量 \mathbf{x} ，如何求一个投影矩阵（或变换矩阵 \mathbf{U} ），满足上述条件？

假设：设 \mathbf{U} 是一个完备正交基的集合，则向量 \mathbf{x} 可以由该**完备正交基**集合线性表示

$$\mathbf{x} = \sum_{i=1}^{\infty} \alpha_i \mathbf{u}_i, \text{ 则 } \alpha_i = \mathbf{u}_i^T \mathbf{x}$$

此时，如果我们只考虑有限估计，即采用有限的 d 个正交基近似表示 \mathbf{x} ，有

$$\hat{\mathbf{x}} = \sum_{i=1}^d \alpha_i \mathbf{u}_i$$



第八章：分类器

□ 完备正交基（知识回顾）

1. 连续函数集合的正交性：

设正交函数集合 $U = \{u_0(t), u_1(t), \dots\}$ ，则有

$$\int_{t_0}^{t_0+T} u_m(t)u_n(t)dt = \begin{cases} C, & \text{if } m = n \\ 0, & \text{otherwise} \end{cases}$$

当 $C=1$ 时，则该正交函数集合为**归一化**的正交函数集合。

多维空间坐标系的基轴方向互相正交。



第八章：分类器

□ 完备正交基（知识回顾）

2. 离散情况

对于n个正交向量集合 $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_n]$ ，其中，

$$\begin{aligned}\mathbf{a}_1 &= [a_{11}, a_{21}, \dots, a_{n1}]^T \\ \mathbf{a}_n &= [a_{1n}, a_{2n}, \dots, a_{nn}]^T\end{aligned}$$

$$\text{则有 } \sum_{k=1}^n a_{k,i} a_{k,j} = \begin{cases} C, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

若 $C=1$ ，则该集合为归一化的正交向量集合。

那么，矩阵 \mathbf{A} 是一个正交矩阵，并满足 $\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I}$



第八章：分类器

□ 完备正交基（知识回顾）

正交函数集合U的完备性

若 $f(x)$ 是定义在 t_0 和 $t_0 + T$ 区间的实值信号，平方可积。可以表示为

$$f(x) = \sum_{n=0}^{\infty} a_n u_n(x)$$

对任意小的 $\varepsilon > 0$ ，存在充分大的 N ，用 N 个有限展开式估计 $f(x)$ ，

$$\tilde{f}(x) = \sum_{n=0}^{N-1} a_n u_n(x)$$

并有 $\int_{t_0}^{t_0+T} |f(x) - \tilde{f}(x)|^2 dx < \varepsilon$ ，

那么称函数集合U是完备的。（集合足够大）

傅里叶变换即是一种正交变换

$$F(u) = \frac{1}{N} \sum_{x=0}^{N-1} f(x) e^{-\frac{j2\pi ux}{N}}, u = 0, 1, \dots, N-1$$
$$f(x) = \sum_{u=0}^{N-1} F(u) e^{\frac{j2\pi ux}{N}}, x = 0, 1, \dots, N-1$$



第八章：分类器

□ 子空间学习（Subspace Learning）

均方误差：

$$\begin{aligned}\varepsilon &= E[(x - \hat{x})^T(x - \hat{x})] = E\left[\sum_{i=d+1}^{\infty} \alpha_i^2\right] = E\left[\sum_{i=d+1}^{\infty} \mathbf{u}_i^T x x^T \mathbf{u}_i\right] \\ &= \sum_{i=d+1}^{\infty} \mathbf{u}_i^T E(x x^T) \mathbf{u}_i = \sum_{i=d+1}^{\infty} \mathbf{u}_i^T \Sigma_x \mathbf{u}_i\end{aligned}$$

均方误差最小化原则： $\min_{\mathbf{U}} \varepsilon = \sum_{i=d+1}^{\infty} \mathbf{u}_i^T \Sigma_x \mathbf{u}_i$

仅当 $\Sigma_x \mathbf{u}_i = \lambda_i \mathbf{u}_i$ 时，取得最小值： $\varepsilon = \sum_{i=d+1}^{\infty} \lambda_i$

结论：前 d 个最大特征值所对应的特征向量 \mathbf{u}_i 所构成的矩阵 \mathbf{U} 即是最终解。



第八章：分类器

□ 子空间学习 (Subspace Learning)

▶ Fisher判别分析(LDA)

与PCA不同，LDA是一种有监督的降维方法，需要利用样本集的标签信息，从而使得获得的低维子空间表达，能够更好地实现分类任务。

两者区别在于，PCA是为了找到一种原始数据的最佳表达，而LDA是为了找到一种有利于分类的最佳表达。

比如：在字符识别中，区分字母O和字母Q。PCA方法会保留两个字母最相似的部分即O，而抛弃字母Q的“尾巴”，然而对于LDA则会尽力保住这个“尾巴”，因为更易于区分。

定义高维特征集 \mathbf{x} ，存在一个投影（或变换 \mathbf{W} ），即特征集 \mathbf{x} 在低维子空间的表达形式为

$$\mathbf{Y} = \mathbf{W}^T \mathbf{X}$$

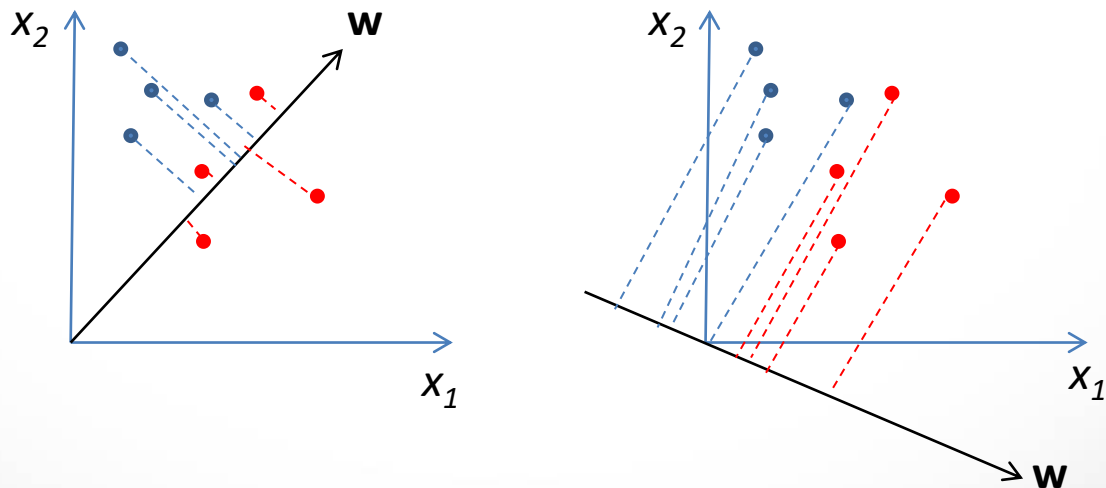
使得 \mathbf{Y} 更加可分类。



第八章：分类器

□ 子空间学习 (Subspace Learning)

➤ Fisher判别分析(LDA)



不同方向的投影，投影后的数据点可分能力不同；对于给定的一组特征集，我们是有可能找到能够最大限度的区分各类数据点的投影方向。



第八章：分类器

□ 子空间学习 (Subspace Learning)

▶ Fisher判别分析(LDA)

- ✓ 定义高维特征集 $X=\{x_1, x_2, \dots, x_n\}$ ，其中 n_1 个样本属于第一类 ω_1 ，剩余的 n_2 个样本属于第二类 ω_2 。那么存在一个投影（或变换 W ），即特征集 X 在低维子空间的表达形式为

$$Y = W^T X$$

使得 Y 更加可分类。

那么该如何建模，进而求出我们期望的 W 呢？

✓ LDA建模过程：

思想：我们希望经过 W 映射后，两类数据的中心间的距离越远越好（异类分离），即类间散度 J_1 ；同时，相同类的样本越往中心靠拢越好（同类紧凑），即类内散度 J_2 。

Fisher准则：满足 $\frac{J_1}{J_2}$ 最大化！



第八章：分类器

$$Y = W^T X$$

□ 子空间学习 (Subspace Learning)

➤ Fisher判别分析(LDA)

✓ LDA建模过程:

思想：我们希望经过 W 映射后，两类数据的中心间的距离越远越好（异类分离），即类间散度 J_1 ；同时，相同类的样本越往中心靠拢越好（同类紧凑），即类内散度 J_2 。

Fisher准则：满足 $\frac{J_1}{J_2}$ 最大化！

1. 投影后，类间散度 J_1 的构造

$$\text{第1类均值 (中心): } \mu_1 = \frac{1}{n_1} \sum_{y \in \omega_1} y = \frac{1}{n_1} \sum_{x \in \omega_1} W^T x = W^T \frac{1}{n_1} \sum_{x \in \omega_1} x = W^T m_1$$

$$\text{第2类均值 (中心): } \mu_2 = \frac{1}{n_2} \sum_{y \in \omega_2} y = \frac{1}{n_2} \sum_{x \in \omega_2} W^T x = W^T \frac{1}{n_2} \sum_{x \in \omega_2} x = W^T m_2$$

$$\text{中心间的距离: } J_1 = \|\mu_1 - \mu\|^2 + \|\mu_2 - \mu\|^2 = \|W^T m_1 - W^T m\|^2 + \|W^T m_2 - W^T m\|^2$$



第八章：分类器

$$Y = W^T X$$

□ 子空间学习 (Subspace Learning)

➤ Fisher判别分析(LDA)

✓ LDA建模过程:

2. 投影后，类内散度J2的构造

$$\text{第1类均值 (中心): } \mu_1 = \frac{1}{n_1} \sum_{y \in \omega_1} y = \frac{1}{n_1} \sum_{x \in \omega_1} W^T x = W^T \frac{1}{n_1} \sum_{x \in \omega_1} x = W^T m_1$$

$$\text{第2类均值 (中心): } \mu_2 = \frac{1}{n_2} \sum_{y \in \omega_2} y = \frac{1}{n_2} \sum_{x \in \omega_2} W^T x = W^T \frac{1}{n_2} \sum_{x \in \omega_2} x = W^T m_2$$

$$\text{第1类类样本点与其中心间的距离: } J2_1 = \sum_{y \in \omega_1} \|y - \mu_1\|^2 = \sum_{x \in \omega_1} \|W^T x - W^T m_1\|^2$$

$$\text{第2类类样本点与其中心间的距离: } J2_2 = \sum_{y \in \omega_2} \|y - \mu_2\|^2 = \sum_{x \in \omega_2} \|W^T x - W^T m_2\|^2$$

$$\text{因此, } J2 = J2_1 + J2_2 = \sum_{x \in \omega_1} \|W^T x - W^T m_1\|^2 + \sum_{x \in \omega_2} \|W^T x - W^T m_2\|^2$$



第八章：分类器

子空间学习 (Subspace Learning)

Fisher判别分析(LDA)

✓ LDA建模过程:

3. 利用Fisher准则 (即满足 $\frac{J1}{J2}$ 最大化), 构造LDA模型

$$\max_W \frac{J1}{J2} = \max_W \frac{\|W^T m_1 - W^T m\|^2 + \|W^T m_2 - W^T m\|^2}{\sum_{x \in \omega_1} \|W^T x - W^T m_1\|^2 + \sum_{x \in \omega_2} \|W^T x - W^T m_2\|^2}$$

模型化简:

$$\max_W \frac{\text{Tr} \left(W^T \left((m_1 - m)(m_1 - m)^T + (m_2 - m)(m_2 - m)^T \right) W \right)}{\text{Tr} \left(W^T \left(\sum_{x \in \omega_1} (x - m_1)(x - m_1)^T + \sum_{x \in \omega_2} (x - m_2)(x - m_2)^T \right) W \right)}$$

$$\max_W \frac{\text{Tr}(W^T S_B W)}{\text{Tr}(W^T S_W W)}$$

特征值分解法求该优化问题

其中, $S_B = (m_1 - m)(m_1 - m)^T + (m_2 - m)(m_2 - m)^T$ 为类间散度矩阵 (between-class scatter); $S_W = \sum_{x \in \omega_1} (x - m_1)(x - m_1)^T + \sum_{x \in \omega_2} (x - m_2)(x - m_2)^T$ 为类内散度矩阵 (within-class scatter)



第八章：分类器

□ 子空间学习 (Subspace Learning)

➤ 多重判别分析(MDA)

MDA实质上是LDA的多类推广：

A. 对于多类问题(假设C个类)， J_2 （类内散度）的构造可以由LDA直接推广，即

$$\begin{aligned} J_2 &= J_{2_1} + J_{2_2} + \dots + J_{2_c} \\ &= \sum_{\mathbf{x} \in \omega_1} \|\mathbf{W}^T \mathbf{x} - \mathbf{W}^T \mathbf{m}_1\|^2 + \dots + \sum_{\mathbf{x} \in \omega_c} \|\mathbf{W}^T \mathbf{x} - \mathbf{W}^T \mathbf{m}_c\|^2 \\ &= \sum_{c=1}^C \sum_{\mathbf{x} \in \omega_c} \|\mathbf{W}^T \mathbf{x} - \mathbf{W}^T \mathbf{m}_c\|^2 = \text{Tr}(\mathbf{W}^T \mathbf{S}_W \mathbf{W}) \end{aligned}$$

其中， $\mathbf{S}_W = \sum_{c=1}^C \sum_{\mathbf{x} \in \omega_c} (\mathbf{x} - \mathbf{m}_c)(\mathbf{x} - \mathbf{m}_c)^T$ 为类内散度矩阵（within-class scatter）



第八章：分类器

□ 子空间学习 (Subspace Learning)

▶ 多重判别分析(MDA)

MDA实质上是LDA的多类推广：

B. 对于多类问题(假设C个类)，J1（类间散度）的构造：

$$\text{第1类均值（中心）： } \mu_1 = \frac{1}{n_1} \sum_{y \in \omega_1} y = \frac{1}{n_1} \sum_{x \in \omega_1} W^T x = W^T \frac{1}{n_1} \sum_{x \in \omega_1} x = W^T m_1$$

$$\text{第2类均值（中心）： } \mu_2 = \frac{1}{n_2} \sum_{y \in \omega_2} y = \frac{1}{n_2} \sum_{x \in \omega_2} W^T x = W^T \frac{1}{n_2} \sum_{x \in \omega_2} x = W^T m_2$$

.

.

$$\text{第C类均值（中心）： } \mu_C = \frac{1}{n_C} \sum_{y \in \omega_C} y = \frac{1}{n_C} \sum_{x \in \omega_C} W^T x = W^T \frac{1}{n_C} \sum_{x \in \omega_C} x = W^T m_C$$

$$\text{总体均值 } \mu = \frac{1}{N} \sum y = \frac{1}{N} \sum W^T x = W^T \frac{1}{N} \sum x = W^T m$$



第八章：分类器

□ 子空间学习 (Subspace Learning)

▶ 多重判别分析(MDA)

MDA实质上是LDA的多类推广：

B. 对于多类问题(假设C个类)，J1（类间散度）的构造：

$$J1 = \sum_{c=1}^C n_c \|\mu_c - \mu\|^2 = \sum_{c=1}^C n_c \|\mathbf{W}^T \mathbf{m}_c - \mathbf{W}^T \mathbf{m}\|^2 = \text{Tr}(\mathbf{W}^T \mathbf{S}_B \mathbf{W})$$

其中， $\mathbf{S}_B = \sum_{c=1}^C n_c (\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^T$

利用Fisher 准则，MDA模型如下：

$$\max_W \frac{\text{Tr}(\mathbf{W}^T \mathbf{S}_B \mathbf{W})}{\text{Tr}(\mathbf{W}^T \mathbf{S}_W \mathbf{W})}$$

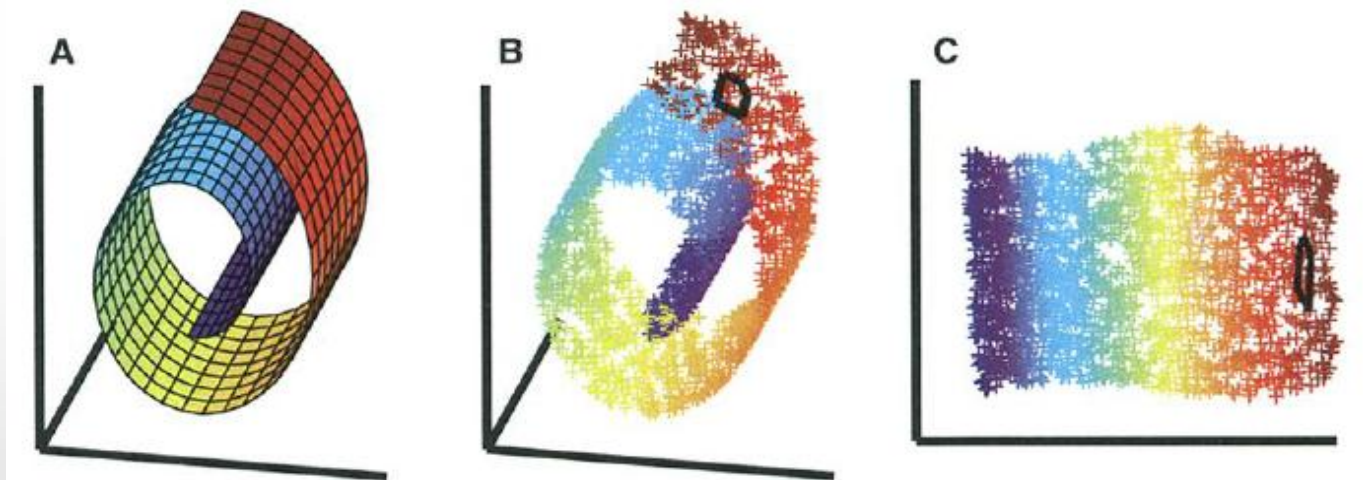
与LDA相同，特征值分解法求该优化问题

第八章：分类器

□ 子空间学习 (Subspace Learning)

➤ 局部保持投影(LPP)

LPP是一种流形思想下的子空间降维方法，前提假设了高维特征数据 \mathbf{X} ，实际上是一种低维的流形结构 \mathbf{Y} 嵌入在高维空间。流形学习的目的映射到低维中，揭示高维数据的本质（局部特性）。





第八章：分类器

□ 子空间学习（Subspace Learning）

➤ 局部保持投影(LPP)

LPP是一种流形思想下的子空间降维方法，前提假设了高维特征数据 \mathbf{x} ，实际上是一种低维的流形结构 \mathbf{Y} 嵌入在高维空间。流形学习的目的映射到低维中，揭示高维数据的本质（局部特性）。

局部保持特性是指，在高维空间相近的 k 个样本点簇，经过 \mathbf{W} 投影以后，这个 k 个样本点簇依然保持相近！

数学模型为

$$\min_W \sum_i \sum_{j, \forall \mathbf{x}_j \in N_k(\mathbf{x}_i)} A_{i,j} \|\mathbf{y}_i - \mathbf{y}_j\|^2$$

$$\text{其中, } A_{i,j} = \begin{cases} \mathbf{1}, & \text{if } \mathbf{x}_j \in N_k(\mathbf{x}_i) \parallel \mathbf{x}_i \in N_k(\mathbf{x}_j) \\ \mathbf{0}, & \text{otherwise} \end{cases}$$



第八章：分类器

□ 子空间学习（Subspace Learning）

▶ 局部保持投影(LPP)

LPP是一种流形思想下的子空间降维方法，前提假设了高维特征数据 \mathbf{x} ，实际上是一种低维的流形结构 \mathbf{Y} 嵌入在高维空间。流形学习的目的映射到低维中，揭示高维数据的本质（局部特性）。

局部保持特性是指，在高维空间相近的 k 个样本点簇，经过 \mathbf{W} 投影以后，这个 k 个样本点簇依然保持相近！

数学模型为

$$\begin{aligned} \min_W \sum_i \sum_{j, \forall \mathbf{x}_j \in N_k(\mathbf{x}_i)} A_{i,j} \|\mathbf{y}_i - \mathbf{y}_j\|^2 &= \min_W \sum_i \sum_{j, \forall \mathbf{x}_j \in N_k(\mathbf{x}_i)} A_{i,j} \|\mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_j\|^2 \\ &= \min_W \text{Tr}(\mathbf{W}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{W}) \end{aligned}$$

其中， \mathbf{L} 是拉普拉斯矩阵， $\mathbf{L}=\mathbf{D}-\mathbf{A}$ ， \mathbf{D} 是一个对角阵， $D_{ii} = \sum_i A_{i,:}$



第八章：分类器

□ 子空间学习与分类器

前面介绍了几种不同的子空间学习方法，其目的是在低维空间内实现分类任务，从而降低计算复杂度，提升分类效率和准确率。

如何实现分类？ 结合本章的前两节，讲述的KNN和SVM方法，可以实现分类，其实现方法基本形式为：

A. PCA+KNN & PCA+SVM

B. LDA+KNN & LDA+SVM

C. MDA+KNN & MDA+SVM

D. LPP+KNN & LPP+SVM

由于子空间模型的灵活性，可提出不同的子空间改进方案，实现高效的分类/识别任务。