线性判别分析

线性降维——主成分分析方法





建设单位：　　大连理工大学
建 设 人：　　刘倩

参考书目：R. O. Duda et. Pattern Classification. Second Edition.

大连理工大学
Dalian University Of Technology

目录

CONTENT




大连理工大学
Dalian University Of Technology

01

PART ONE

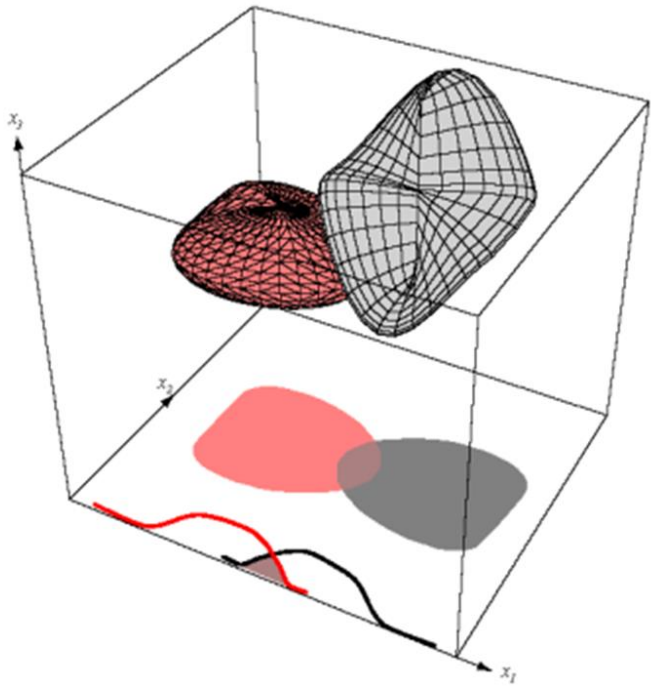
“维数灾难”问题

The Curse of Dimensionality

1 “维数灾难”问题

- 一般情况下，如果使用现有特征获得的分类器性能不佳，则会考虑**添加新的特征**，以提高分类器性能（以**运算复杂度**为代价）

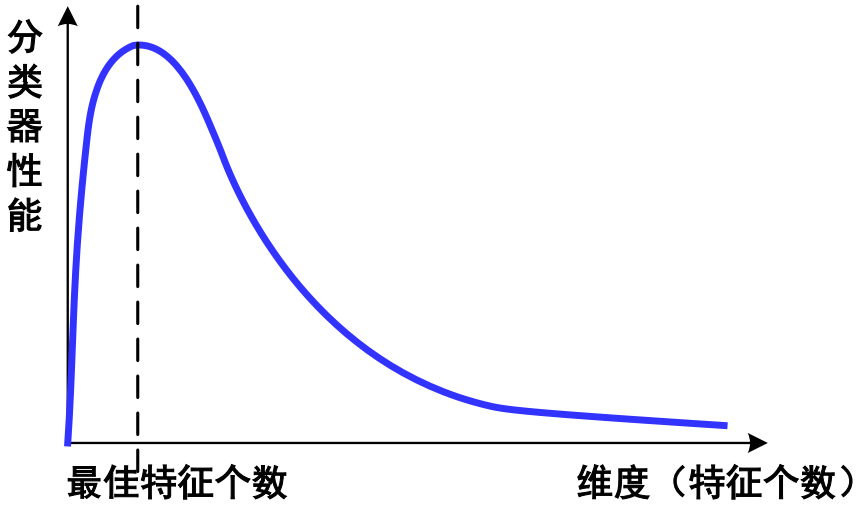
分类器的准确率 ⟺ **运算复杂度**



- 举例

1. 假设两个分布的密度函数在三维空间内无交集（即贝叶斯分类器在三维空间无误差）；
2. 其在一维空间和二维空间的投影存在交集，所以贝叶斯分类器在一维和二维空间存在误差；
3. 理论上，如果**添加新特征提供了额外的信息**，则分类器的性能将得到提高。

1 “维数灾难”问题



- 分类器的性能随着特征个数的变化不断增加，过了某一个值后，性能不升反降，这种现象称为“维数灾难”

- 原因分析：
  1. 使用了错误的模型（如：高斯分布）
  2. 训练样本的数量有限

- 随着特征空间维数的增加，对样本数量的需求呈指数级增长

举例：深度学习算法对大规模训练数据的需求。

1 “维数灾难”问题

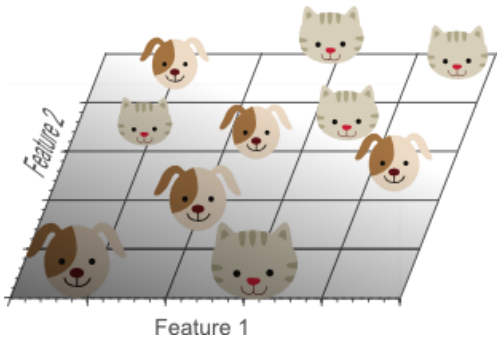
❑ 举例

假设有猫和狗的N张照片，在有限计算能力下，仅选取10张作为训练样本。
目标：训练一个线性分类器，对剩余的照片进行正确分类。

1. 一个特征

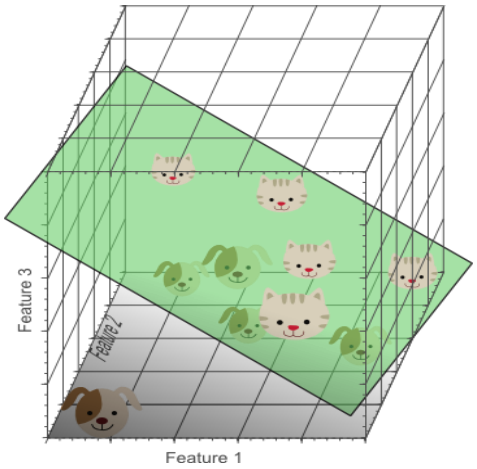
很难线性分类，误差大

2. 两个特征

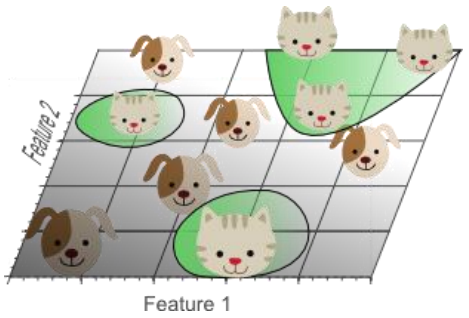


很难线性分类，误差大

3. 三个特征



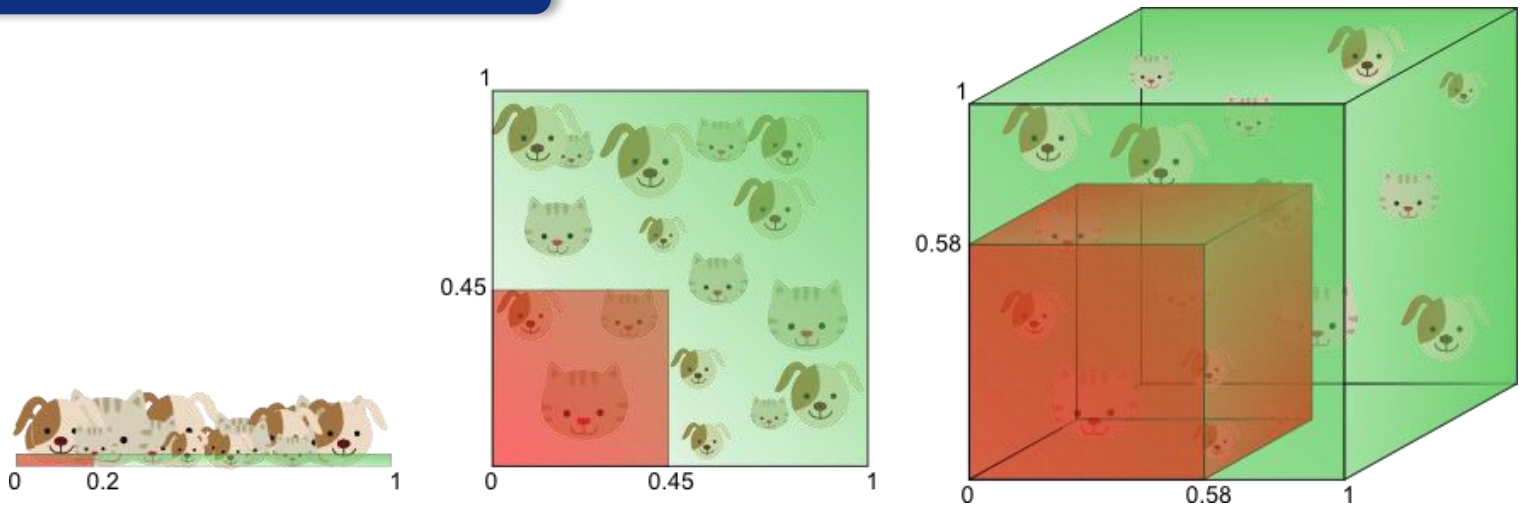
映射



过拟合问题

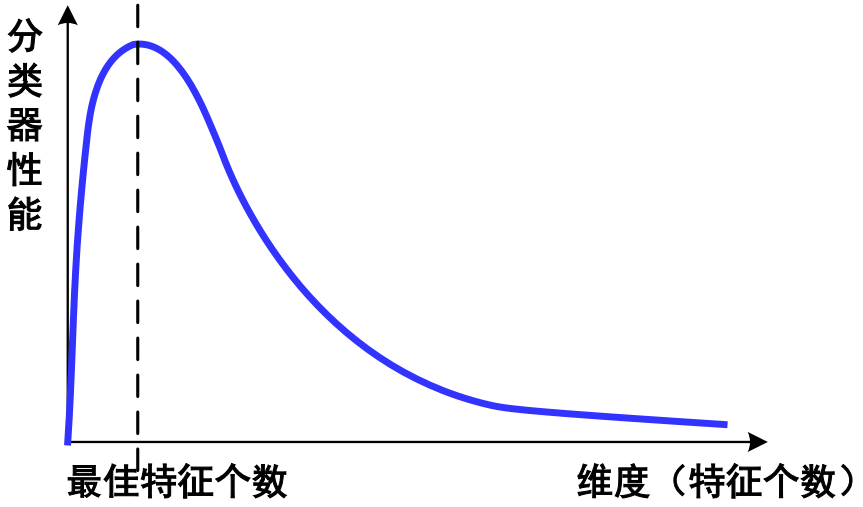
增加特征数量使得高维空间线性可分，相当于在低维空间内训练一个复杂的非线性分类器

1 “维数灾难”问题



- 假设只有一个特征时，特征的值域是0到1，每一只猫和狗的特征值都是唯一的。如果希望训练样本覆盖特征值值域的20%，那么就需要猫和狗总数的20%；
- 增加一个特征后，覆盖特征值值域的20%需要猫和狗总数的45% ($0.45^2=0.2$)
- 继续增加一个特征后，覆盖特征值值域的20%需要猫和狗总数的58% ($0.58^3=0.2$)
- 随着特征数量的增加，为了覆盖特征值值域的20%，就需要更多的训练样本。
- 如果没有足够的训练样本，就可能会出现过拟合问题。

1 “维数灾难”问题



- 分类器的性能随着特征个数的变化不断增加，过了某一个值后，性能不升反降，这种现象称为“维数灾难”

- 原因分析：
  1. 使用了错误的模型（如：高斯分布）
  2. 训练样本的数量有限

- 随着特征空间维数的增加，对样本数量的需求呈指数级增长

如何应对高维特征？ → 特征降维（保留主要信息量） → 主成分分析方法

大连理工大学
Dalian University Of Technology

PART TWO

主成分分析（PCA）方法

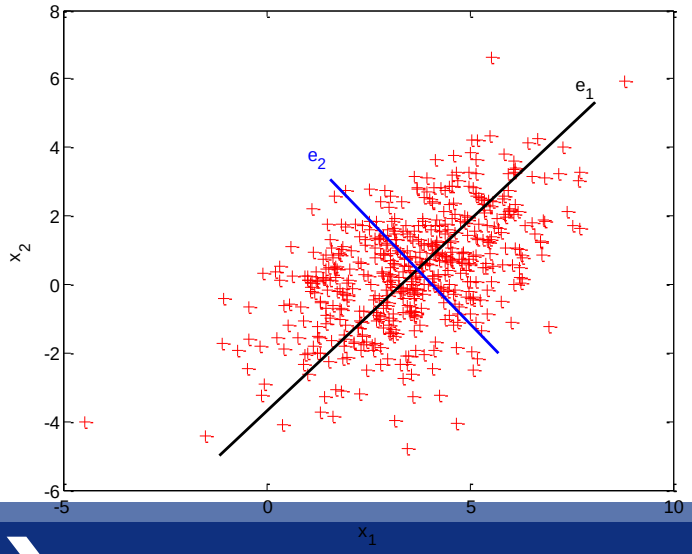
Principle Component Analysis

2 主成分分析方法(PCA)

❑ 主成分分析（PCA）

- 从原始高维空间投影到低维子空间，称为**主成分子空间**，且投影点包含了原始数据的绝大部分信息
- 从最小二乘的角度，PCA是寻求一个低维空间，使**投影误差平方和（即数据点与投影点的距离平方和）最小**

❑ 举例——从二维降到一维（投影/映射）



1. 为了尽可能的保留最多的原始信息，一种直观的看法：**希望投影点尽可能分散**
2. 投影点的分散程度数学上用**方差**描述
3. **最佳投影直线的方向为投影点方差最大方向**
4. e_1比e_2更可能成为主成分方向

2 主成分分析方法

❑ 问题描述

如何将N个d-维样本点 $\mathbf{x}_1, \ldots \mathbf{x}_N \in \mathbb{R}^d$, 投影到$d'$-维子空间（$d' < d$），并使原始数据点与投影点的距离平方和最小?

❑ 数学推导过程

- 主要思路：寻找数据的主轴方向，由主轴构成一个新的坐标系（维度可以比原维度低），然后数据由原坐标系向新的坐标系投影，投影误差最小

1. 寻找数据的主轴方向

首先，找到一个样本点$\mathbf{x}_0$表示所有N个样本点 $\mathbf{x}_1, \ldots \mathbf{x}_N \in \mathbb{R}^d$，使$\mathbf{x}_0$ 与 $\mathbf{x}_k$之间的距离平方和最小

误差准则函数为： $\mathrm{J}_0(\mathbf{x}_0) = \sum_{k=1}^{N} \|\mathbf{x}_0 - \mathbf{x}_k\|^2$

目标函数： $J_0(\mathbf{x}_0) = \sum_{k=1}^{N} \|\mathbf{x}_0 - \mathbf{x}_k\|^2$

❑ 如何找到$\mathbf{x}_0$使$J_0(\mathbf{x}_0)$最小?

$$J_0(\mathbf{x}_0) = \sum_{k=1}^{N} \|(\mathbf{x}_0 - \mathbf{m}) - (\mathbf{x}_k - \mathbf{m})\|^2$$

⇒引入样本均值$\mathbf{m}$，并将$+\mathbf{m} - \mathbf{m}$添加到$J_0(\mathbf{x}_0)$中

$$= \sum_{k=1}^{N} \|\mathbf{x}_0 - \mathbf{m}\|^2 - 2\sum_{k=1}^{n} (\mathbf{x}_0 - \mathbf{m})^T(\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2$$

$$= \sum_{k=1}^{N} \|\mathbf{x}_0 - \mathbf{m}\|^2 - 2(\mathbf{x}_0 - \mathbf{m})^T \underbrace{\sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m})}_{=0} + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2$$

独立于$\mathbf{x}_0$

样本均值

$\min J_0(\mathbf{x}_0)$ ⟹ $\mathbf{x}_0 = \mathbf{m} = \frac{1}{N}\sum_{k=1}^{N} \mathbf{x}_k$

大连理工大学 Dalian University Of Technology

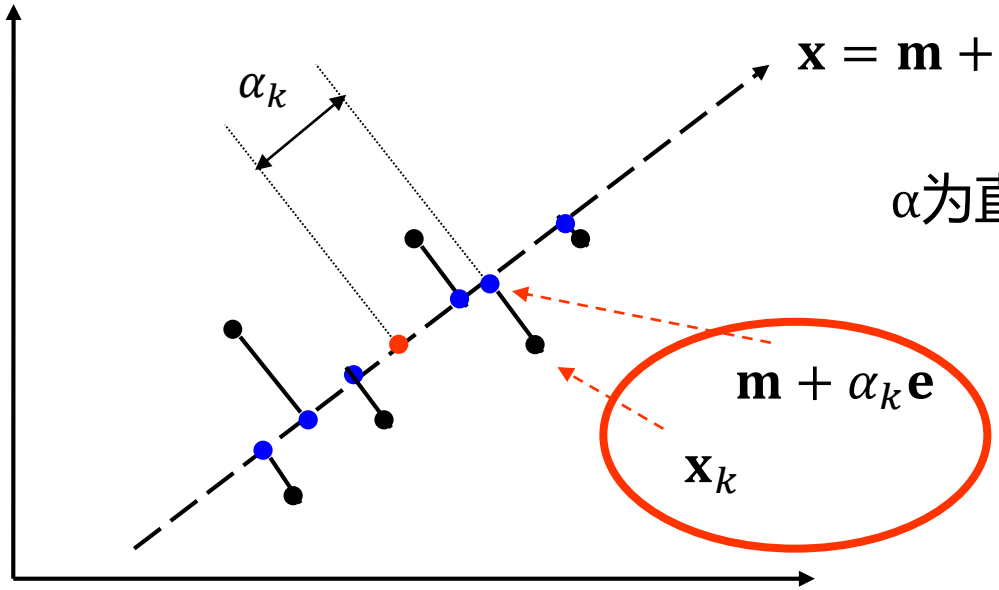
❑零维

- 样本均值是样本数据集的零维表达
- 将样本数据集的空间分布，压缩为一个均值点

简单，但不能反映样本间的差异

❑一维 ——将数据集空间，压缩为一条过样本均值点的线

每个样本在直线上存在不同的投影，可以反映样本间的差异



$\mathbf{x}=\mathbf{m}+\alpha\mathbf{e}$

$\mathbf{e}$为直线的单位向量

α为直线上的点到$\mathbf{m}$的距离

$$J_0(\mathbf{x}_0) \Rightarrow J_1(\alpha_1,\cdots,\alpha_n,\mathbf{e})$$

$$=\sum_{k=1}^{n}\|(\mathbf{m}+\alpha_k\mathbf{e})-\mathbf{x}_k\|^2$$

❑ 一维平方误差（目标函数）可表示为：

$$J_1(\alpha_1,\cdots,\alpha_n,\mathbf{e})=\sum_{k=1}^{N}\|(\mathbf{m}+\alpha_k\mathbf{e})-\mathbf{x}_k\|^2=\sum_{k=1}^{N}\|\alpha_k\mathbf{e}-(\mathbf{x}_k-\mathbf{m})\|^2$$

$$=\sum_{k=1}^{N}{\alpha_k}^2\|\mathbf{e}\|^2-2\sum_{k=1}^{N}\alpha_k\mathbf{e}^T(\mathbf{x}_k-\mathbf{m})+\sum_{k=1}^{N}\|\mathbf{x}_k-\mathbf{m}\|^2$$

对α_k求偏导，得到：

$$\frac{\partial \boldsymbol{J}_1}{\partial\alpha_k}=2\alpha_k\|\mathbf{e}\|^2-2\mathbf{e}^T(\mathbf{x}_k-\mathbf{m})=0$$

由于$\|\mathbf{e}\|=1$，得到：

$$\alpha_k=\mathbf{e}^T(\mathbf{x}_k-\mathbf{m})$$

从几何学来讲，这个结果说明，只需要将$\mathbf{x}_k$垂直投影到通过样本均值的直线e上，就能够获得最小方差

如何找到e的最优方向?

$$J_1(\alpha_1,\cdots,\alpha_n,\mathbf{e})=\sum_{k=1}^{N}{\alpha_k}^2\|\mathbf{e}\|^2-2\sum_{k=1}^{N}\alpha_k\mathbf{e}^T(\mathbf{x}_k-\mathbf{m})+\sum_{k=1}^{N}\|\mathbf{x}_k-\mathbf{m}\|^2$$

- 将得到的$\alpha_k=\mathbf{e}^T(\mathbf{x}_k-\mathbf{m})$代入上式，得到

$$J_1(\mathbf{e})=\sum_{k=1}^{n}{\alpha_k}^2-2\sum_{k=1}^{n}{\alpha_k}^2+\sum_{k=1}^{n}\|\mathbf{x}_k-\mathbf{m}\|^2$$

$$=-\sum_{k=1}^{n}[\mathbf{e}^T(\mathbf{x}_k-\mathbf{m})]^2+\sum_{k=1}^{n}\|\mathbf{x}_k-\mathbf{m}\|^2$$

$$=-\mathbf{e}^T\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m})(\mathbf{x}_k-\mathbf{m})^T\mathbf{e}+\sum_{k=1}^{n}\|\mathbf{x}_k-\mathbf{m}\|^2$$

散布矩阵

$$\mathbf{S}=\sum_{k=1}^{N}(\mathbf{x}_k-\mathbf{m})(\mathbf{x}_k-\mathbf{m})^T$$

$$J_1(\mathbf{e})=-\mathbf{e}^T\mathbf{S}\mathbf{e}+\sum_{k=1}^{n}\|\mathbf{x}_k-\mathbf{m}\|^2$$

$$J_1(\mathbf{e}) = -\mathbf{e}^T\mathbf{S}\mathbf{e} + \sum_{k=1}^{n}\|\mathbf{x}_k - \mathbf{m}\|^2$$

使J_1最小的向量e，能够使$\mathbf{e}^T\mathbf{S}\mathbf{e}$最大

- 因此，目标函数可以表示为：

$$\max\{\mathbf{e}^T\mathbf{S}\mathbf{e}\}$$
$$\text{s.t. } \|\mathbf{e}\| = 1$$

- 拉格朗日乘子法，得到：

$$\mathbf{u} = \mathbf{e}^T\mathbf{S}\mathbf{e} - \lambda(\mathbf{e}^T\mathbf{e} - 1)$$

- 对e求偏导数，得到：

$$\frac{\partial \mathbf{u}}{\partial \mathbf{e}} = 2\mathbf{S}\mathbf{e} - 2\lambda\mathbf{e} = 0$$

$\mathbf{S}\mathbf{e} = \lambda\mathbf{e}$ ⇒ $\mathbf{e}^T\mathbf{S}\mathbf{e} = \lambda\mathbf{e}^T\mathbf{e} = \lambda$ ⇒ 最大化$\mathbf{e}^T\mathbf{S}\mathbf{e}$ 等价于最大化λ

λ为散布矩阵S的最大特征值，e 为最大特征值对应的特征向量

2 主成分分析方法

❑一维空间映射推广至d'维空间映射

$$\mathbf{x} = \mathbf{m} + \alpha\mathbf{e} \quad \Rightarrow \quad \mathbf{x} = \mathbf{m} + \sum_{k=1}^{d'} \alpha_i \mathbf{e}_i$$

d'维投影的目标函数 $J_{d'}$可表示为：

$$J_{d'} = \sum_{k=1}^{n} \left\| \left(\mathbf{m} + \sum_{k=1}^{d'} \alpha_{ki} \mathbf{e}_i \right) - \mathbf{x}_k \right\|^2$$

要使$J_{d'}$ 最小，只需将散布矩阵S的特征值由大到小排序，选择最大的前d'个特征值对应的特征向量构成一个新的d'维坐标系，将样本向新的坐标系的各个轴上投影，计算出新的特征矢量

大连理工大学
Dalian University Of Technology

PART THREE

使用PCA进行特征降维步骤

Implementation Procedure of Feature Reduction with PCA

3 使用PCA进行特征降维步骤

目标: 使用PCA将具有d-维特征的N个样本，$\mathbf{x}_1, \ldots \mathbf{x}_N \in \mathbb{R}^d$, 进行特征降维，即求出 $\mathbf{y}_1, \ldots \mathbf{y}_N \in \mathbb{R}^{d'}$ $(d' < d)$.

1. 计算样本均值 $\mathbf{m} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$;

2. 计算散布矩阵 $\mathbf{S} = \sum_{i=1}^{N} \mathbf{z}_i \mathbf{z}_i^t$， 其中 $\mathbf{z}_i = \mathbf{x}_i - \mathbf{m}$;

3. 计算散布矩阵 $\mathbf{S}$ 的特征值和特征向量，得到最大的d'个特征值对应的特征向量 $\mathbf{e}_1, \ldots, \mathbf{e}_{d'}$ ；

4. 使用 $\mathbf{e}_1, \ldots, \mathbf{e}_{d'}$ 构建矩阵 $\mathbf{M} = [\mathbf{e}_1 \ \ldots \ \mathbf{e}_{d'}]$;

5. 降维后的向量 $\mathbf{y}_i$ 可以表示为： $\mathbf{y}_i = \mathbf{M}^T \mathbf{z}_i$

大连理工大学
Dalian University Of Technology

04

PART FOUR

总结

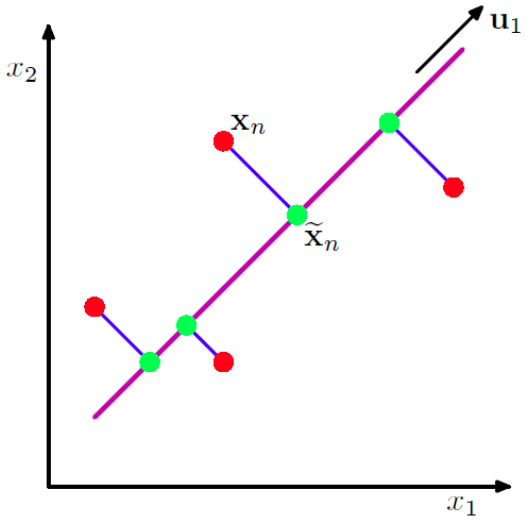
Conclusions

4 总结

- PCA的两种解释

  1. 寻求一个低维空间，称为主成分子空间，用紫色线表示，使得数据点（红点）在这个子空间上的正交投影点（绿点）的方差最大化；

  2. 最小化由蓝线表示的投影误差平方和（即数据点与投影点的距离平方和）。



- 如何选择合适的 d'？

$$\frac{\sum_{i=1}^{d'}\lambda_i}{\sum_{i=1}^{d}\lambda_i} > \text{门限值} \quad (\text{例如：}0.9, 0.95)$$

- PCA将所有的样本作为一个整体对待，去寻找一个均方误差最小意义下的最优线性映射，而**忽略了类别属性**，即**方差最大的方向不一定可用于分类**