（请大家预习）

第7章第1讲
特征提取与特征变换
Feature Extraction and Feature Transformation

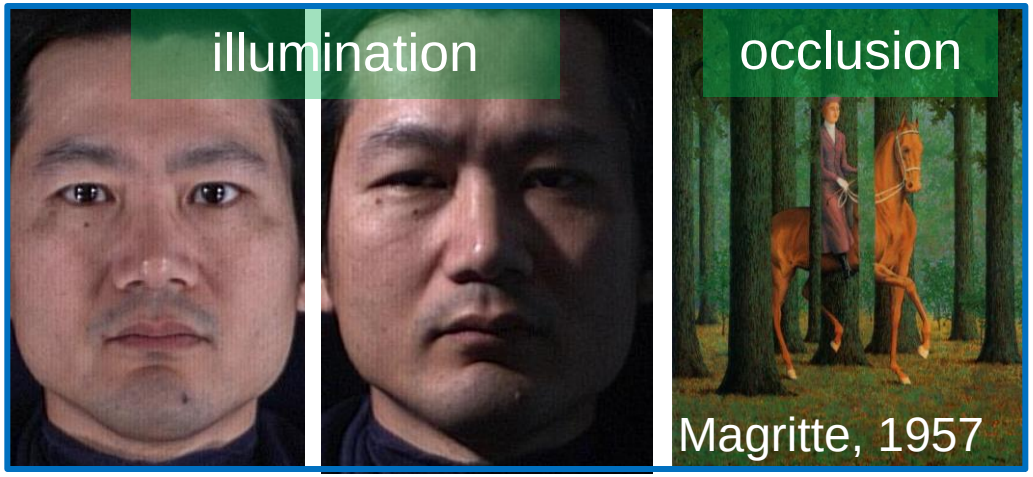
张 燕 明
ymzhang@nlpr.ia.ac.cn
people.ucas.ac.cn/~ymzhang

模式分析与学习课题组(PAL)
多模态人工智能系统实验室 中科院自动化所

助教： 杨 奇 (yangqi2021@ia.ac.cn)
张 涛 (zhangtao2021@ia.ac.cn)

7.1 引言

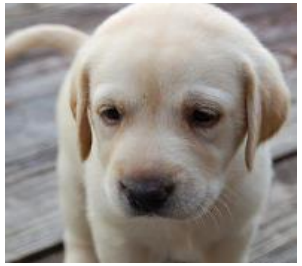
- **自然图像中的挑战**



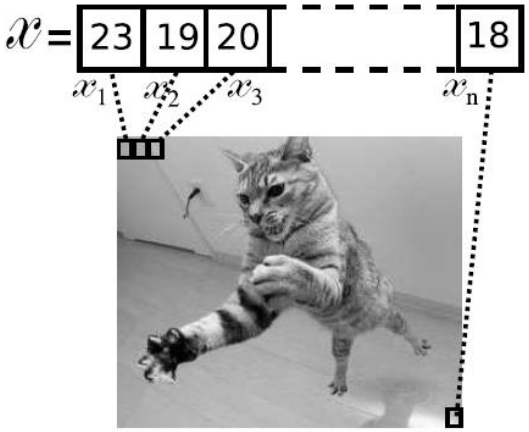
7.1 引言

- 特征表示的重要性

If we want to create an algorithm to distinguish dogs from cats:


Raw input vector representation



L_1 distance: $d_1(I_1 - I_2) = \sum_p \left| I_1^p - I_2^p \right|$

56	32	10	18
90	23	128	133
24	26	178	200
2	0	255	220

Test image

\-

10	20	24	17
8	10	89	100
12	16	178	170
4	32	233	112

Train image

=

46	12	14	1
82	13	39	33
12	10	0	30
2	32	22	108

add → 456

7.1 引言

特征表示的重要性：在深度学习时代

It is well understood that the performance of machine learning methods is heavily dependent on the choice of data representation (or features) on which they are applied. The rapidly developing field of representation learning is concerned with questions surrounding how we can best learn meaningful and useful representations of data. We take a broad view of the field, and include in it topics such as deep learning and feature learning, metric learning, kernel learning, compositional models, non-linear structured prediction, and issues regarding non-convex optimization.

Despite the importance of representation learning to machine learning and to application areas such as vision, speech, audio and NLP, there is currently no common venue for researchers who share a common interest in this topic. The goal of ICLR is to help fill this void.

ICLR 2013 will be a 3-day event from May 2nd to May 4th 2013, co-located with AISTATS2013 in Scottsdale, Arizona. The conference will adopt a novel publication process, which is explained in further detail here: Publication Model.

Regards,

Yoshua Bengio & Yann Lecun,
General Chairs

- 特征表示的重要性

Domain-specific Feature Representation：

- Preprocessing
- Feature extraction
- Reducing within-class variance
- Enlarging Between-class variance

Statistical Pattern Recognition

- Feature Transformation: PCA, LDA, ICA, Isomap, LLE, ...
- Feature Selection: Wrapper, Filter, Embedded,…
- Bayesian Decision Theory: Gaussian, Parzen, KNN, Mixture…
- Neural Network: MLP, RBF, CNN, GNN
- Decision Tree: ID3, C4.5, CART, Random forests
- Kernel Method: SVM
- Ensemble Method: Bagging, Boosting
- Clustering: K-means, Hierarchical , Spectral clustering

7.1 引言

- **特征提取的目的**
  - 提取观测数据的内在特性
  - 减少噪声影响
  - 提高稳定性

- **特征变换的目的**
  - 降低特征空间的维度，增加数据密度、降低过拟合风险
  - 便于分析和减少后续步骤的计算量
  - 减少特征之间可能存在的相关性
  - 有利于分类

7.1 引言

- **根据特征提取对象不同**
  - 语音特征提取
  - 文本特征提取
  - 视觉特征提取
- **根据特征提取的方式不同**
  - 局部特征提取方法： LBP、SIFT等
  - 全局特征提取方法：词袋模型、HoG等

7.1 引言

- **根据特征变换关系不同**
  - 线性特征变换：采用线性映射将原特征变换至一个新的空间（通常维度更低）：
    - PCA、LDA 、ICA
  - 非线性特征变换：采用非线性映射将原特征变换至一个新的空间（通常性能更好）：
    - KPCA、KLDA、Isomap、LLE、 HLLE、 LSTA、…

第一部分：特征提取

7.2 特征提取

7.2.1 语音特征提取

7.2.2 文本特征提取

- 向量空间模型(BOW, TF-IDF)
- 词向量模型(word2vec)

7.2.3 视觉特征提取

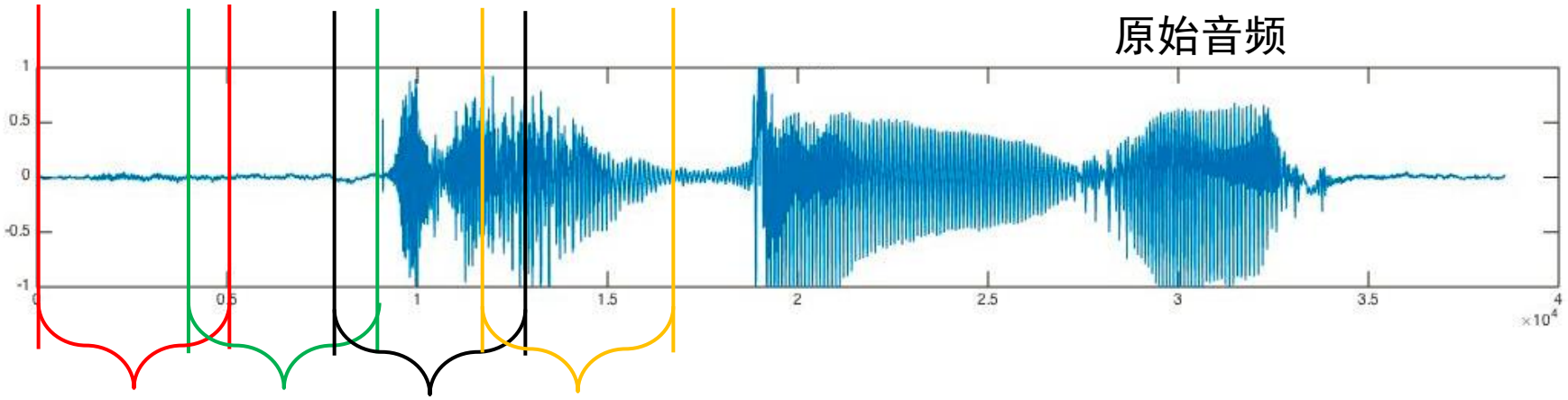
- 局部二值模式（LBP）
- Gabor特征提取
- 尺度不变特征变换（SIFT）
- 视觉词袋（Bag of Visual Words）
- 哈尔特征(Harr)
- 梯度方向直方图（HoG）

7.2.1 语音特征提取 [略]

- **语音特征提取技术路线：**
  1. 对输入的语音信号进行预处理
  2. 对语音信号进行分帧、加窗处理
  3. 对每一帧的波形信号进行一些特定的数学运算，得到低维向量作为提取的特征

- **MFCCs**（Mel Frequency Cepstral Coefficients，梅尔倒谱系数）：
  - 一种在自动语音和说话人识别中广泛使用的特征。
  - 1980年由Davis和Mermelstein提出，语音识别领域人工特征的佼佼者。
  - 符合人的听觉特性。

MFCCs特征提取



语音信号分帧：将一段语音信号，划分成若干帧

- 帧信号要加窗函数（低通滤波），使得帧两端信号平滑过渡到零
- 帧与帧之间有重叠（帧移），以免帧边缘处信号因加窗弱化而丢失

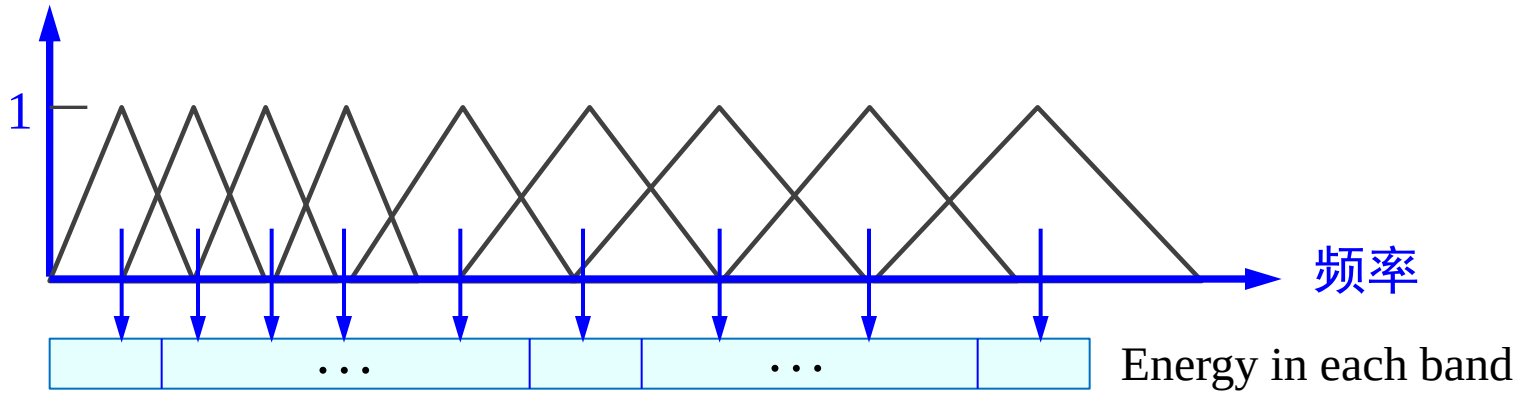
MFCCs特征提取

逐帧计算MFCCs特征：

1. 傅里叶变换：对分帧后的语音信号进行傅里叶变换，保留幅度谱，丢弃相位谱
2. 根据梅尔刻度，利用频域三角窗对傅里叶幅度谱进行求和
3. 对求和后的幅度取对数
4. 离散余弦变换：对取对数后的幅度信号进行离散余弦变换，得到MFCCs特征。

MFCCs特征提取

求取频谱在每个三角形区域内的能量总和



- 一个三角形对应一个梅尔频率带
- 低频密、高频疏，模仿人耳听觉特性（梅尔频率刻度）
- 减少数据量、提高稳定性。一般取40个三角形，而傅里叶变换后的频率个数一般几百到上千。

7.2.2 文本特征提取

- 文档/文本：若干词项的有序集合。
- 文本特征提取：将文本内容转化为向量的过程。将一个文档表示成一个向量，向量的相似性反应文档的相似性。
- 主要方法：
  - 向量空间模型（Vector Space Model）：忽略词的顺序，将文档看作词的集合
    - 词袋模型（Bag of Words）
    - **TF-IDF**
  - 词向量：Word2Vec

7.2.2 文本特征提取

词袋模型（Bag of Words, BOW）：一个维度对应于字典中的一个词项。如果一个词项出现在一篇文档中，它在向量中的值是非零的，否则为零。

- 例子

字典：

[‘中’ ‘国’ ‘人’ ‘北’ ‘京’ ‘科’ ‘学’ ‘院’ ‘大’]

文档 d_1 = “中国北京” ⇒ [1, 1, 0, 1, 1, 0, 0, 0, 0]

文档 d_2 = “中国科学院大学” ⇒ [1, 1, 0, 0, 0, 1, 2, 1, 1]

字典包含所有可能的词项，文档特征向量长度等于字典长度

7.2.2 文本特征提取

词频-逆向文档频率（Term Frequency-Inverse Document Frequency, TF-IDF）

- 语料库记为 D，一个由若干文档组成的集合；文档记为 d；词语记为 t
- 词频$TF(t, d)$：词语 t 在文档d 出现的次数 ÷ 文档d 中的总词数
- 文档频率$DF(t, D)$：语料库D中包含词语 t 的文档个数 ÷ 语料库D中的文档个数
- 逆向文档频率$IDF(t, D)$：衡量语料库D中词语 t 提供的信息量

$$IDF(t, D) = \log \frac{1}{DF(t, D)}$$

- 词频-逆向文档频率：

$$\text{TF-IDF}(t, d, D) = TF(t, d) \times IDF(t, D)$$

7.2.2 文本特征提取

词频-逆向文档频率（TF-IDF）：例子

字典：[‘中’ ‘国’ ‘科’ ‘学’ ‘院’ ‘大’]

文档d_1=“中国”，文档d_2= “中国科学院”，文档d_3=“中国科学院大学”

语料库$D = \{d_1, d_2, d_3\}$

d_3的词频：

TF(中，d_3) =TF(国，d_3) =TF(科，d_3) =TF(院，d_3) =TF(大，d_3) = 1/7

TF(学，d_3) = 2/7

文档频率：

DF(中，D) = DF(国，D) = 3/3

DF(科，D) = DF(学，D) = DF(院，D) = 2/3

DF(大，D) = 1/3

d_3的TF-IDF特征表示：

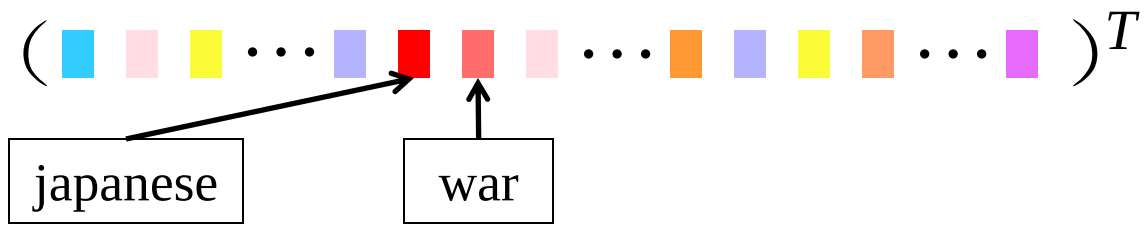
$$[0, \quad 0, \quad \frac{1}{7}\log\frac{3}{2}, \quad \frac{2}{7}\log\frac{3}{2}, \quad \frac{1}{7}\log\frac{3}{2}, \quad \frac{1}{7}\log 3]$$

‘中’ ‘国’ ‘科’ ‘学’ ‘院’ ‘大’

7.2.2 文本特征提取

词频-逆向文档频率（TF-IDF）：应用

文档特征向量：



US Presidential Speeches Tag Cloud
http://chir.ag/phernalia/preztags/

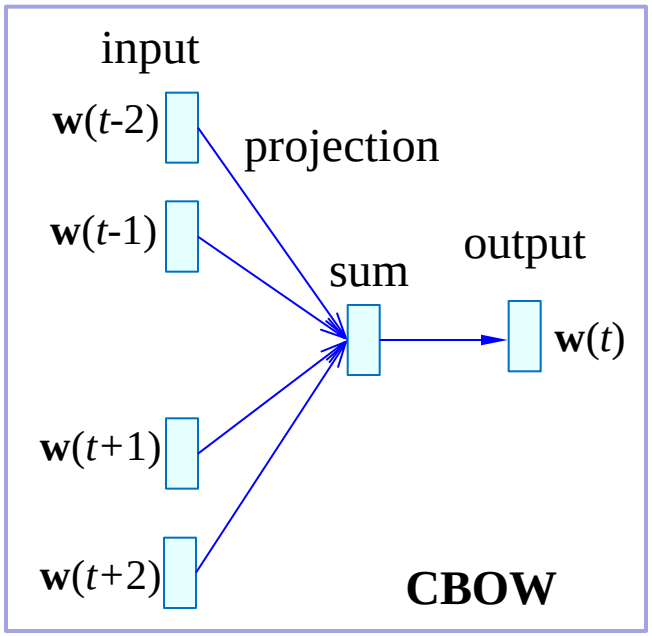
Word2Vec

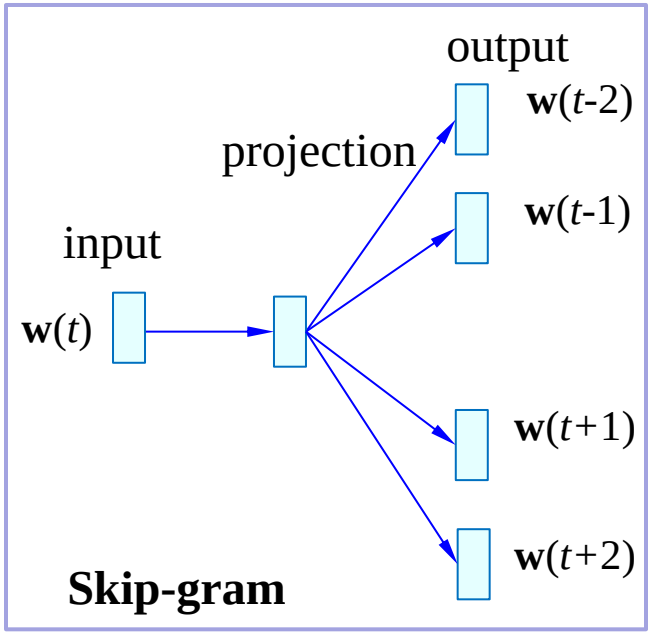
基本任务：利用一个连续向量来表示一个词项

主要目的：相似单词具有相似的向量表示

技术路线：浅层神经网络

主要模型：连续词袋模型（Continuous Bag of Words, CBOW），
跳字模型（Skip Gram）

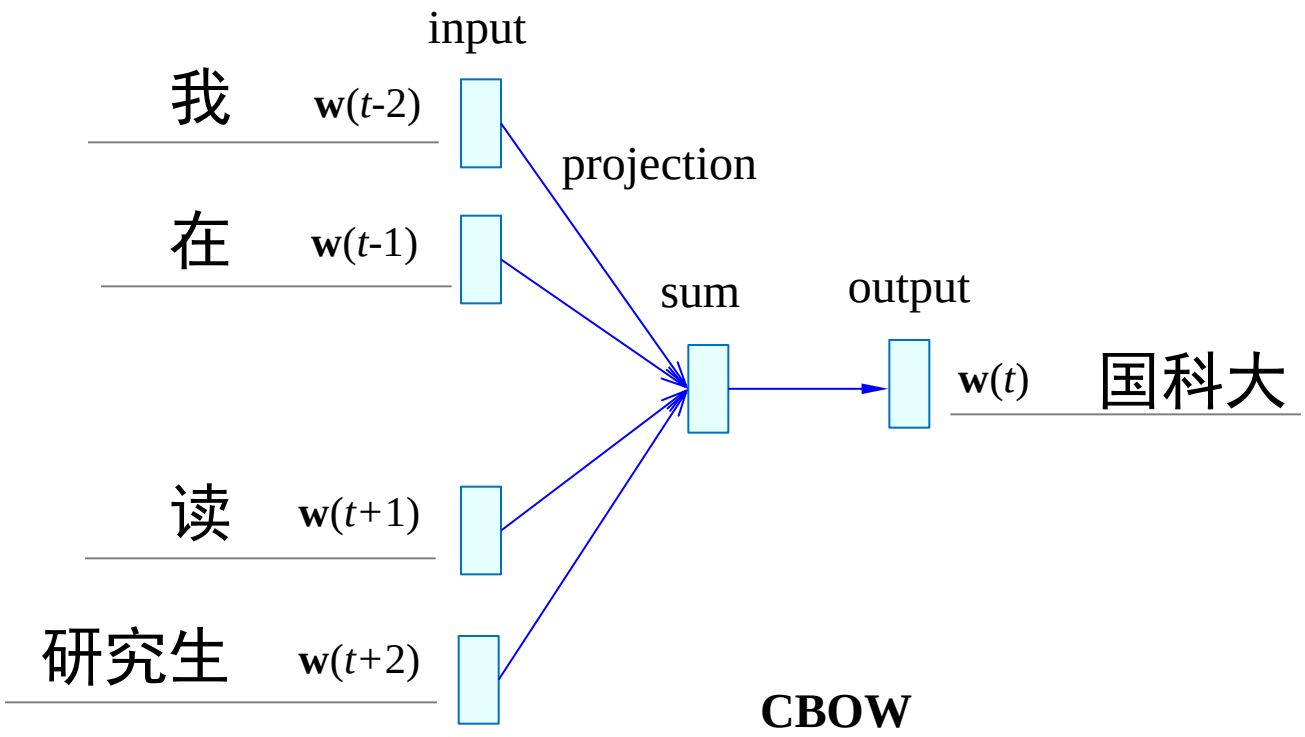




利用上下文单词预测单词　　利用单词预测上下文单词

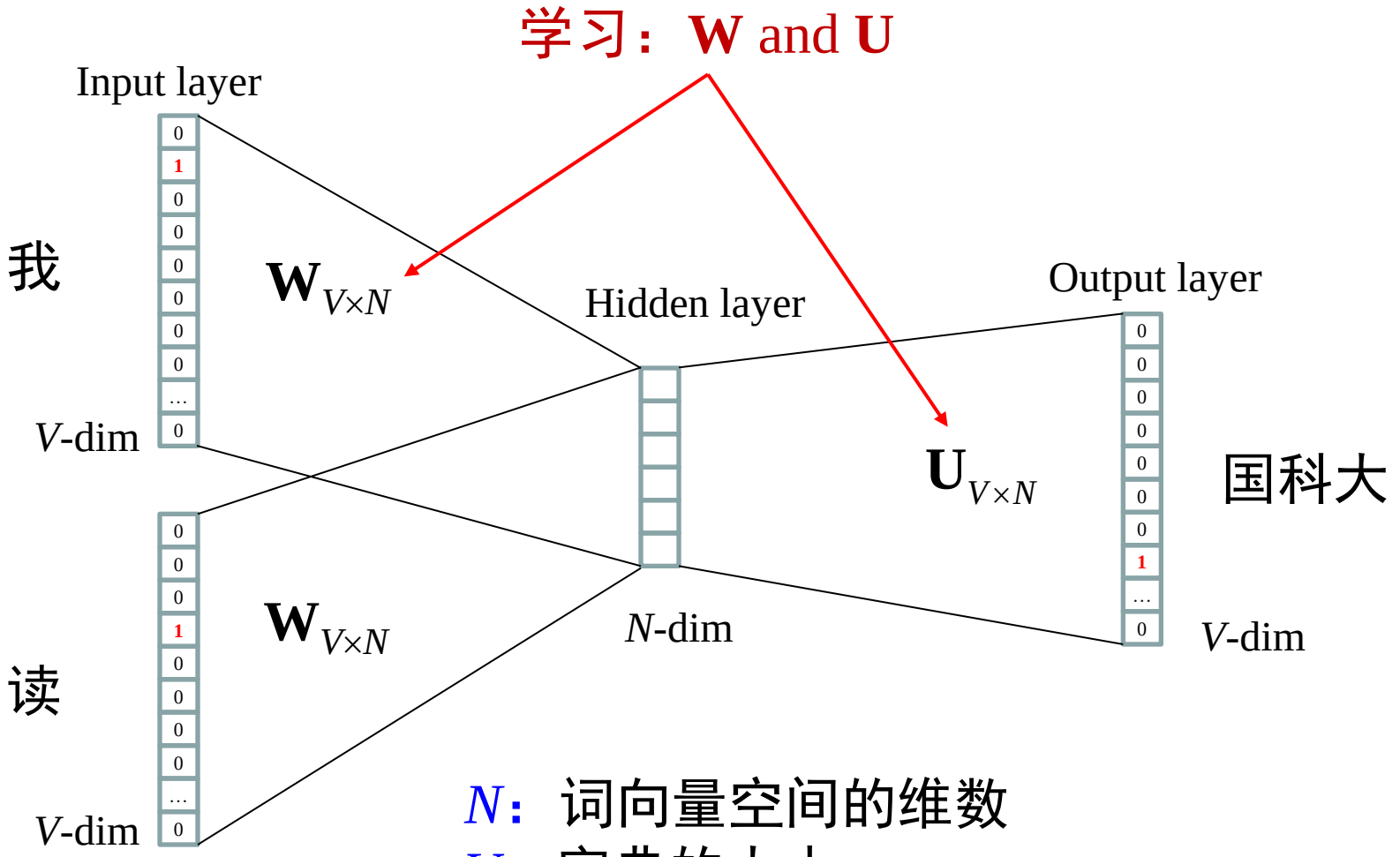
Word2Vec

例子：我在国科大读研究生。



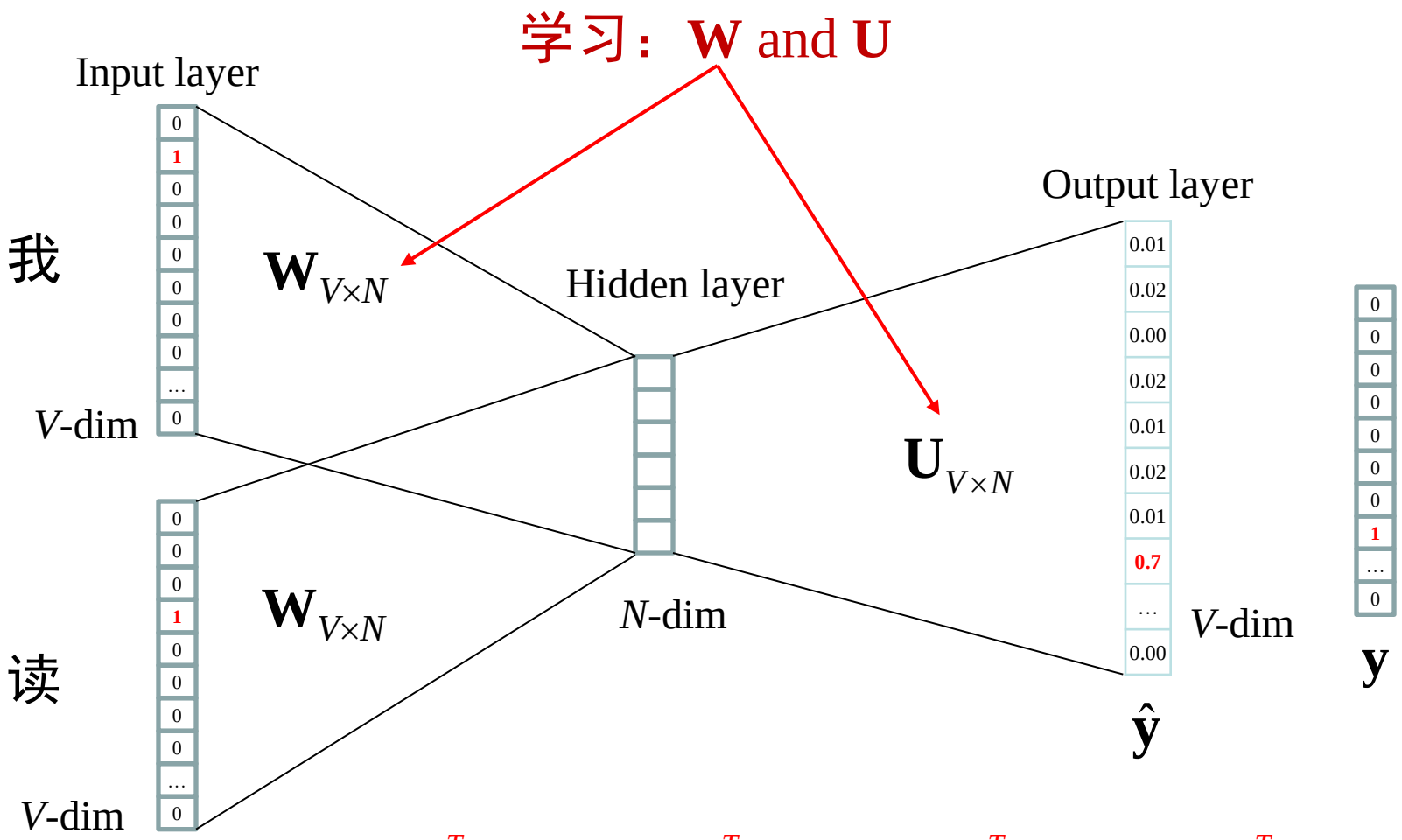
取上下文窗口为2

假定字典的大小为V，每一个词用一个V维one-hot向量表示，即对应的维度元素为1，其余为0。



N：词向量空间的维数
V：字典的大小

学习目标：希望预测的 ŷ 与真实的 y 接近 (softmax)



$$\mathbf{z} = \mathbf{U}(\mathbf{W}^T w(t-2) + \mathbf{W}^T w(t-1) + \mathbf{W}^T w(t+1) + \mathbf{W}^T w(t+2))$$

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{z})$$

Word2Vec

- ✓ **W** 对应了入词汇表的word2vec矩阵，每一行对应一个词的向量表示。
- ✓ **U** 对应了输出词汇表的word2vec矩阵，每一行对应一个词的向量表示。

任取其一，或者取平均作为词汇的特征提取结果。

7.2.3 视觉特征提取

- 局部二值模式（LBP）
- Gabor特征提取
- 尺度不变特征变换（SIFT）
- 视觉词袋（Bag of Visual Words）
- 哈尔特征（Harr）
- 梯度方向直方图（HoG）

7.2.3 视觉特征提取

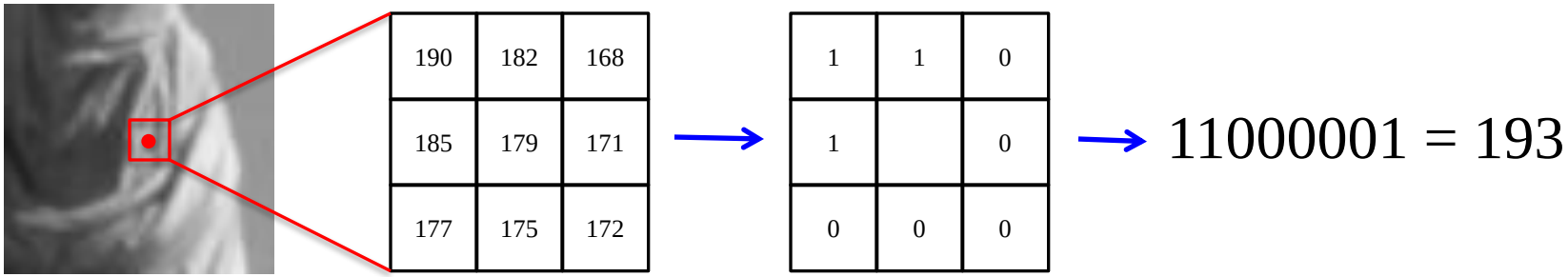
Local Binary Pattern (LBP)

Matti Pietikäinen （芬兰奥卢大学）

T. Ojala, M. Pietikainen, T. Maenpaa.Multiresolution grayscale and rotation invariance texture classification with local binary patterns, IEEE TPAMI, 24(7), 971-987, 2002.

- 局部特征提取方法，针对每个像素点计算
- 计算简单、对于光照变化较稳定
- 广泛用于纹理分析、人脸检测、识别

LBP特征计算过程



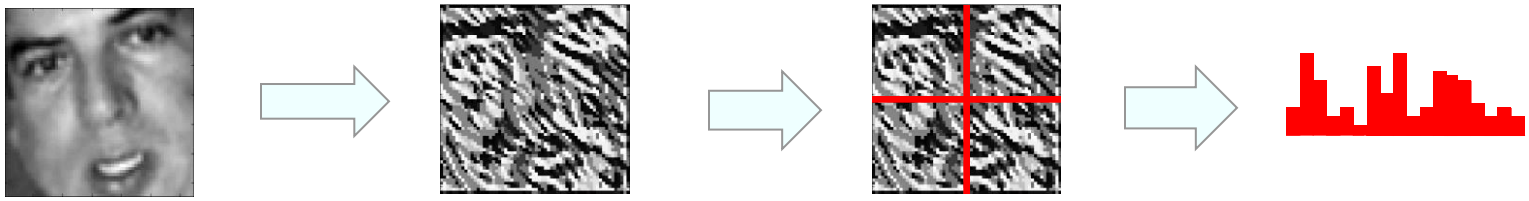
LBP特征的一般性定义：

$$LBP_{R,N}(x)=\sum_{i=0}^{N-1}\mathrm{sign}\left(I(x_i)-I(x)\right)2^i$$

R：像素周围邻域半径；N：像素周围区域采样点个数

若采样点x_i不是整数（不在像素格子上）？双线性插值

LBP特征应用：人脸识别



- 对人脸图像提取人脸特征向量之后，送入人脸分类模型进行人脸识别。
- 人脸分类模型：
  - 特征变换 + *k*NN分类；
  - 特征变换 + 多类SVM；
  - 特征变换 + 神经网络；

第二部分：特征变换

7.3 特征变换

- 特征提取
  - 从数据观测获得原始特征表达的过程，例如：将一幅图像表示成一个向量；将一个文本转化成一个向量；将一段语音表示成一个向量等。
- 特征变换
  - 对已有向量特征进行变换，得到新特征的过程。

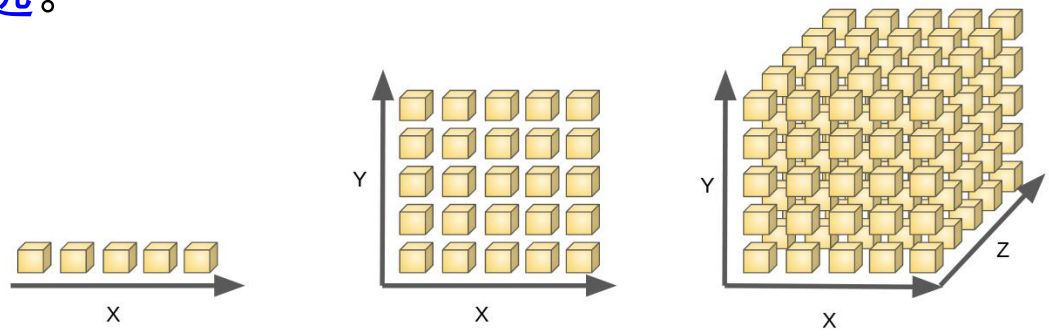
线性变换： $\mathbf{y} = \mathbf{W}^T \mathbf{x}$

非线性变换： $\mathbf{y} = \mathbf{W}(\mathbf{x})$

最终形式都是使用向量来表示数据样本，便于分析。

7.3.1 维数缩减

- **维数灾难**(the curse of dimensionality)
  - 维数灾难最早由理查德·贝尔曼（Richard E. Bellman）在考虑优化问题时提出的，用于描述当空间维度增加时分析和组织数据会遇到各种问题。
    - 当维度增加时，空间的体积增加得很快，可用数据变得稀疏。
    - 稀疏性对于任何要求“具有统计学意义的方法”而言都是一个问题。但是，为了获得在统计学上正确并且可靠的结果，所需要的数据量通常随着维数的增加而呈指数级增长。
    - 随着维数的增加，具有相同距离的两个样本其相似程度可以相差很远。

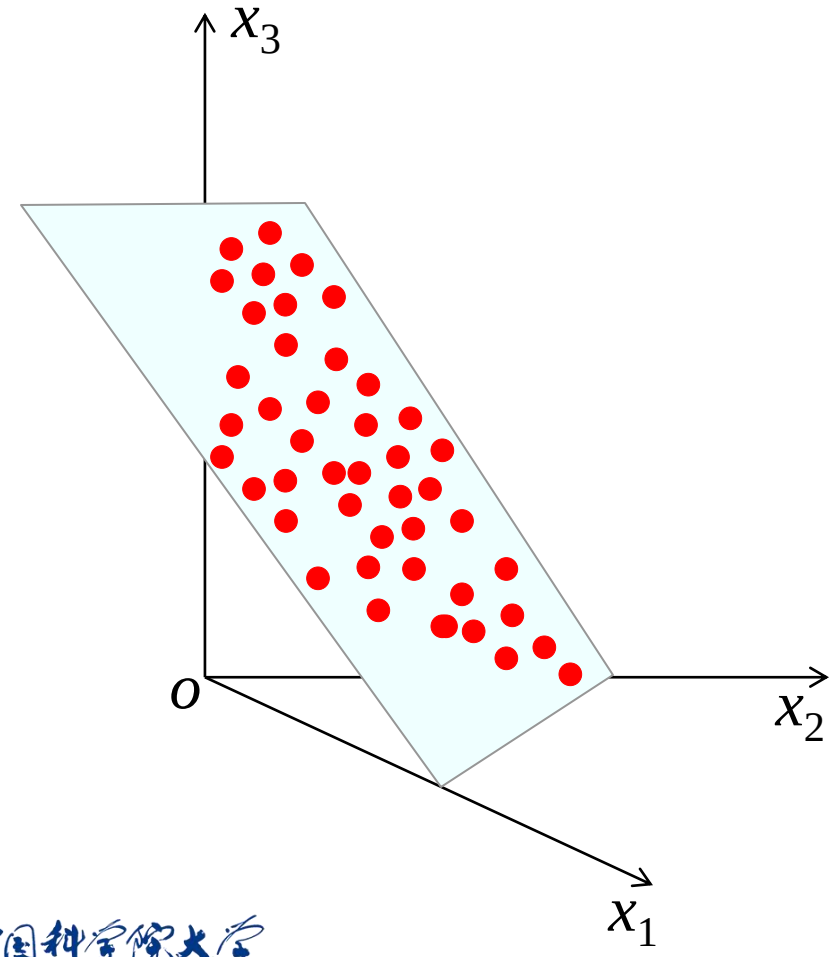


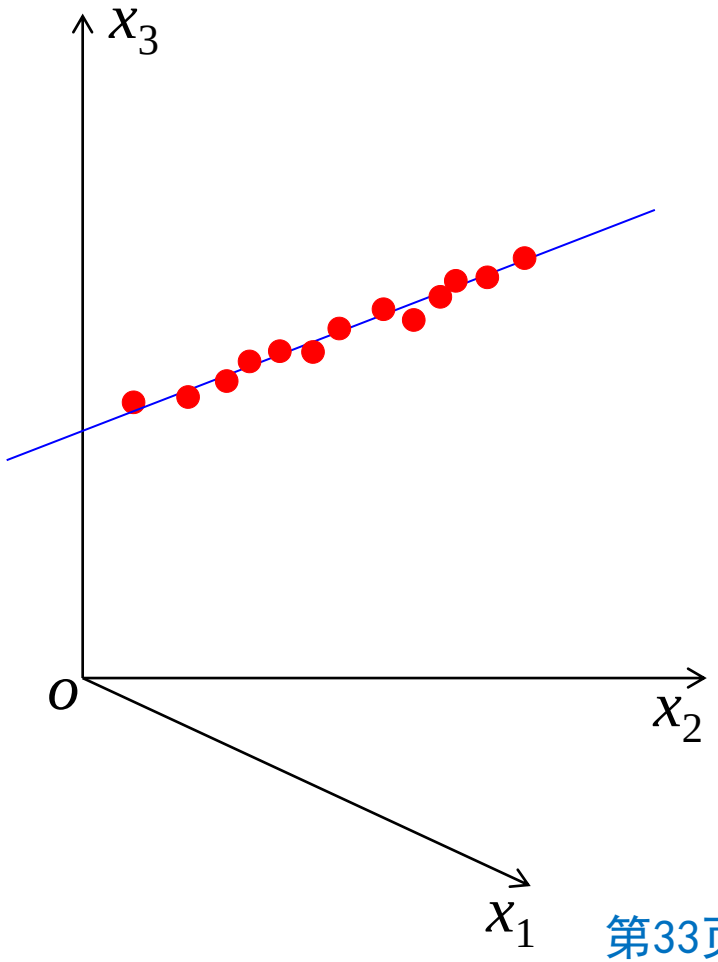
7.3.1 维数缩减

- **维数缩减**(dimensionality reduction)
  - 缓解维数灾难的一个重要途径是降维，即通过某种数学变换将原始高维特征空间变换至某个低维“子空间”。在该子空间中，样本密度大幅度提高，距离计算也变得更为容易。
  - 为什么能降维？
    - 在很多时候，人们观测或收集到的数据虽然是高维的，但与学习任务密切相关的特征通常位于某个低维分布上，即高维空间中的一个低维“嵌入”(embedding)。
    - 感谢非均匀性祝福！

7.3.1 维数缩减

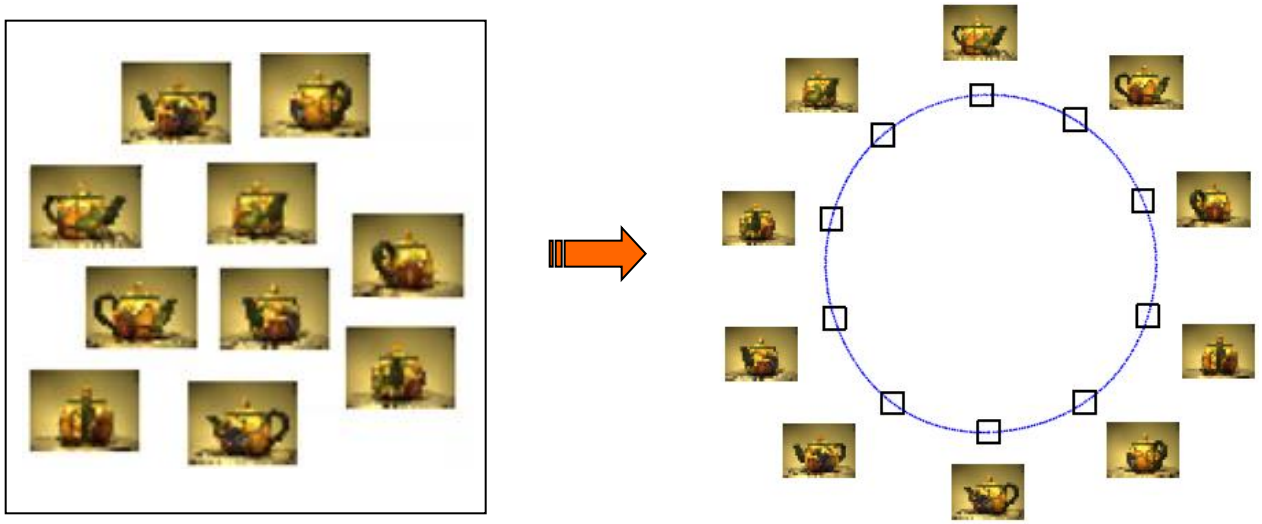
- **线性低维嵌入的例子**

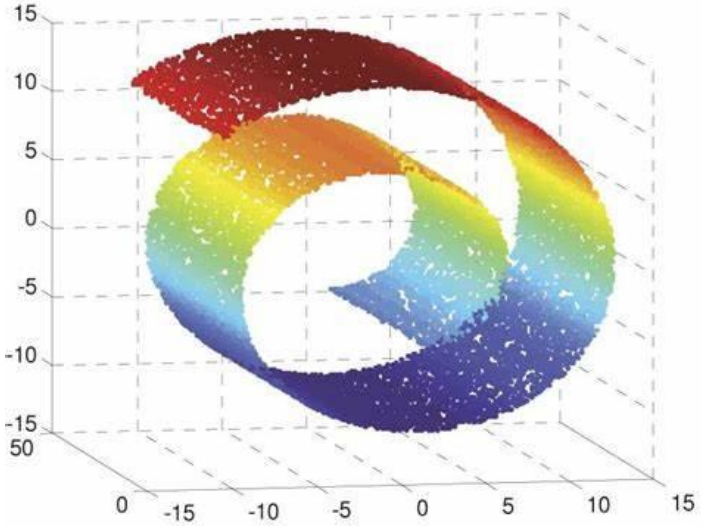




7.3.1 维数缩减

- **非线性低维嵌入的例子**





7.3.1 维数缩减

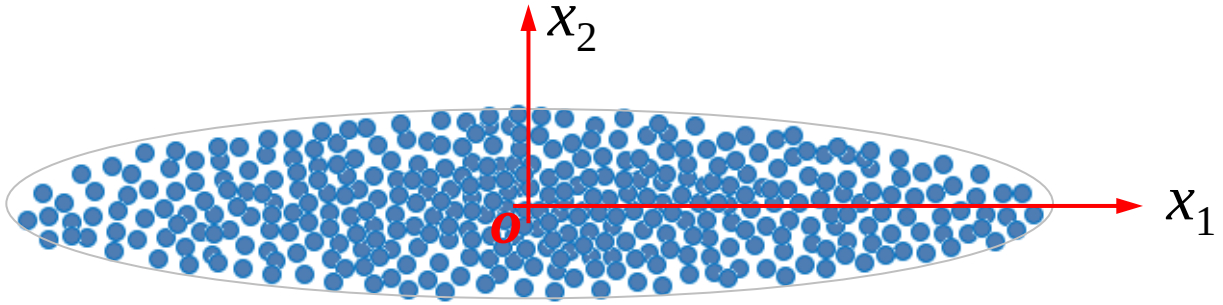
- **线性降维法**
  - 对高维空间中的样本**x**进行线性变换：

$$\mathbf{y} = \mathbf{W}^T\mathbf{x}, \quad \text{where } \mathbf{x} \in R^m, \mathbf{W} \in R^{m\times d}, \mathbf{y} \in R^d, d < m$$

  - 变换矩阵$\mathbf{W}=[\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_d]$可视为$m$维空间中由$d$ 个基向量组成的矩阵。
  - $\mathbf{y}=\mathbf{W}^T\mathbf{x}$可视为样本**x**与$d$个基向量分别做内积而得到，即**x**在新坐标系下的坐标。新空间中的特征是原空间中特征的线性组合。这就是线性降维法。
  - 不同降维方法的差异：对低维子空间的性质有不同的要求，即对**W**施加不同的约束。

7.3.2 主成分分析 (Principal Component Analysis, PCA)

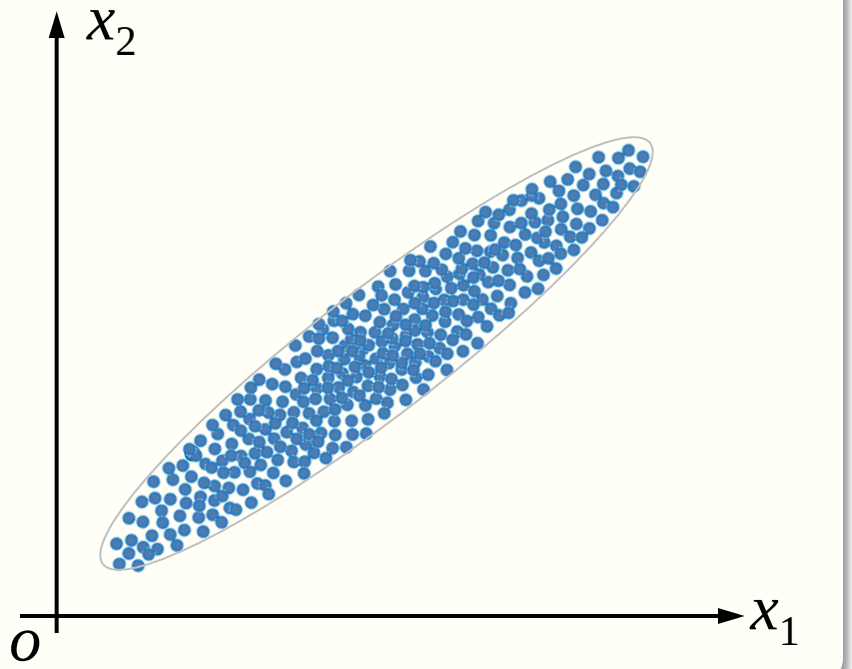
- **一个例子**



- ✓ 向x_1轴投影：忽略每个样本的第二维，只保留$\{x_1\}$
- ✓ 向x_2轴投影：忽略每个样本的第一维，只保留$\{x_2\}$
- ✓ 问题：
  - 对上述两种投影操作，哪一种将更多地保留原始数据集的信息？（沿x_1轴方差较大，沿x_2轴方差较小）

7.3.2 主成分分析(PCA)

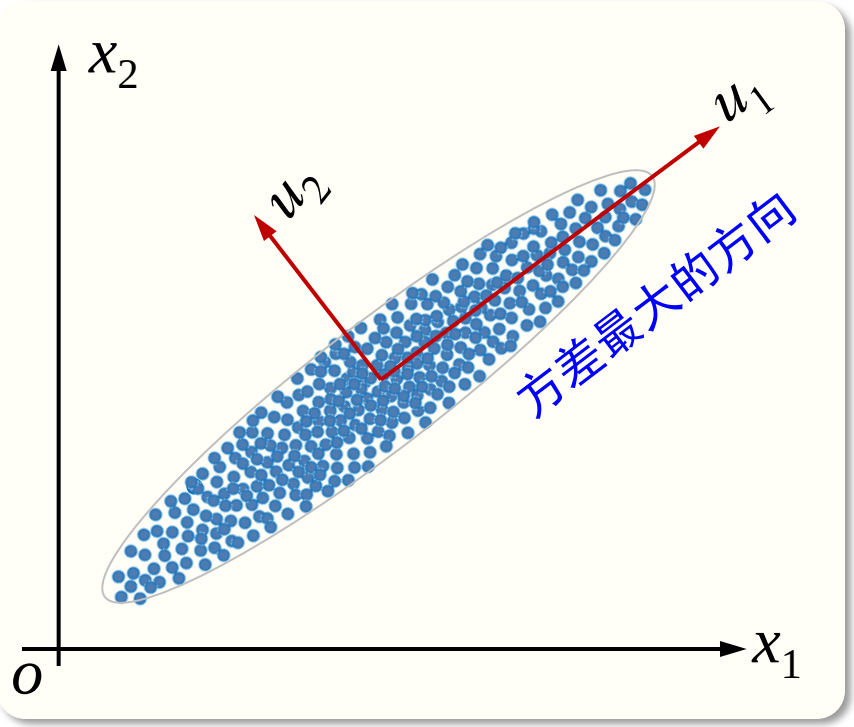
- **一个例子**



- ✓ 向x_1轴投影：忽略每个样本的第二维，只保留$\{x_1\}$
- ✓ 向x_2轴投影：忽略每个样本的第一维，只保留$\{x_2\}$
- ✓ 对上述两种投影操作均不能很好地保留原始数据集的信息。

7.3.2 主成分分析(PCA)

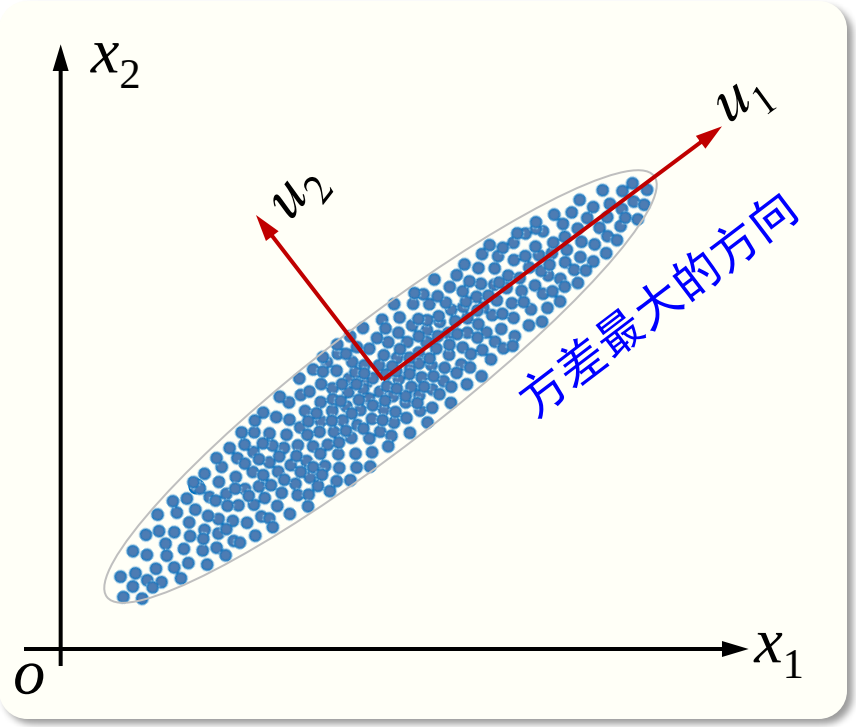
- **一个例子**



- ✓ 向u_1轴投影：忽略每个样本的第二维，只保留$\{u_1\}$
- ✓ 向u_2轴投影：忽略每个样本的第一维，只保留$\{u_2\}$
- ✓ 在坐标系得到变换后（等价地，各个样本得到变换后），第一种投影操作仍然能够很好地保留原始数据集的信息。

7.3.2 主成分分析(PCA)

- **动机**



- ✓ 动机：寻找一组方差较大的方向，将原始数据（样本）在这些方向上进行投影。即将数据在新坐标系下进行表示，保留少数在方差最大方向上的投影，达到尽可能地保留原始数据信息和降维的目的。
- ✓ 方差较大的方向称为主成分(Principal Components)。其中，方差最大的方向称为第一主成分，其次为第二主成分，依次类推。

PCA: Finding Principal Components

- **Given:** n examples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, each example $\mathbf{x}_n \in \mathrm{R}^d$.
- **Goal:** we want to **capture the maximum possible variance in the projected data**, namely, project the data from d dimensions to m dimensions with $m < d$.
- Technically and algorithmically, let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m \in \mathrm{R}^d$ be the principal components, assumed to be:
  - **Orthogonal:** $(\mathbf{w}_i)^T \mathbf{w}_j = 0, \forall\ i \neq j$, and $(\mathbf{w}_i)^T \mathbf{w}_i = 1$, $i, j = 1, 2, \ldots, m$.
- ***We want only the first m principal components.***

PCA: Finding Principal Components

- Projection $\mathbf{y}_i$ of a data point $\mathbf{x}_i$ along $\mathbf{w}_1$: $\mathbf{w}_1^T \mathbf{x}_i$
- Projection $\bar{\mathbf{y}}$ of the mean $\bar{\mathbf{x}}$ along $\mathbf{w}_1$:

$$\bar{\mathbf{y}} = \mathbf{w}_1^T \bar{\mathbf{x}}, \quad \text{where} \quad \bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

- Variance of the projected data (along projection direction $\mathbf{w}_1$):

$$\text{var} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{w}_1^T \mathbf{x}_i - \mathbf{w}_1^T \bar{\mathbf{x}} \right)^2$$

- Want to obtain direction $\mathbf{w}_1$ that maximizes the projected data variance:

$$\max \quad \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{w}_1^T \mathbf{x}_i - \mathbf{w}_1^T \bar{\mathbf{x}} \right)^2$$

$$\text{s.t.} \quad \mathbf{w}_1^T \mathbf{w}_1 = 1$$

PCA: Finding Principal Components

- Note that:

$$\begin{aligned}
\text{var} &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{w}_1^T\mathbf{x}_i - \mathbf{w}_1^T\bar{\mathbf{x}}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{w}_1^T\mathbf{x}_i - \mathbf{w}_1^T\bar{\mathbf{x}}\right)\left(\mathbf{w}_1^T\mathbf{x}_i - \mathbf{w}_1^T\bar{\mathbf{x}}\right)^T \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_1^T\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^T\mathbf{w}_1 \\
&= \mathbf{w}_1^T\left(\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^T\right)\mathbf{w}_1 \\
&= \mathbf{w}_1^T\mathbf{C}\mathbf{w}_1
\end{aligned}$$

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^T$$

covariance matrix of data

- Then we have:

$$\max \quad \mathbf{w}_1^T\mathbf{C}\mathbf{w}_1,$$

$$\text{s.t.} \quad \mathbf{w}_1^T\mathbf{w}_1 = 1$$

PCA: Finding Principal Components

- Now we introduce a Lagrange multiplier λ to this subject, obtaining the following objective function:

$$obj = \mathbf{w}_1^T \mathbf{C} \mathbf{w}_1 - \lambda(\mathbf{w}_1^T \mathbf{w}_1 - 1)$$

- Taking the derivative w.r.t. $\mathbf{w}_1$ and setting it to zero gives:

$$\frac{\partial obj}{\partial \mathbf{w}_1} = 2\mathbf{C}\mathbf{w}_1 - 2\lambda \mathbf{w}_1 = 0$$

$$\mathbf{C}\mathbf{w}_1 = \lambda \mathbf{w}_1$$

- This is just an eigenvalue equation. Thus, $\mathbf{w}_1$ must be an eigenvector of $\mathbf{C}$.

PCA: Finding Principal Components

- We know that $\mathbf{w}_1$ must be an eigenvector of $\mathbf{C}$, and λ is the corresponding eigenvalue.
- But, there are multiple eigenvectors of $\mathbf{C}$, which one is $\mathbf{w}_1$?
- Consider

$$\mathbf{w}_1^T \mathbf{C} \mathbf{w}_1 = \mathbf{w}_1^T \lambda \mathbf{w}_1 = \lambda \mathbf{w}_1^T \mathbf{w}_1 = \lambda$$

- We want to maximize the projected data variance $\mathbf{w}_1^T \mathbf{C} \mathbf{w}_1 = \lambda$
  - Thus λ should be the largest eigenvalue, and $\mathbf{w}_1$ is the first eigenvector of $\mathbf{C}$ (with eigenvalue λ) .
  - This is the first principal component (direction of highest variance in the data)

PCA: The Algorithm

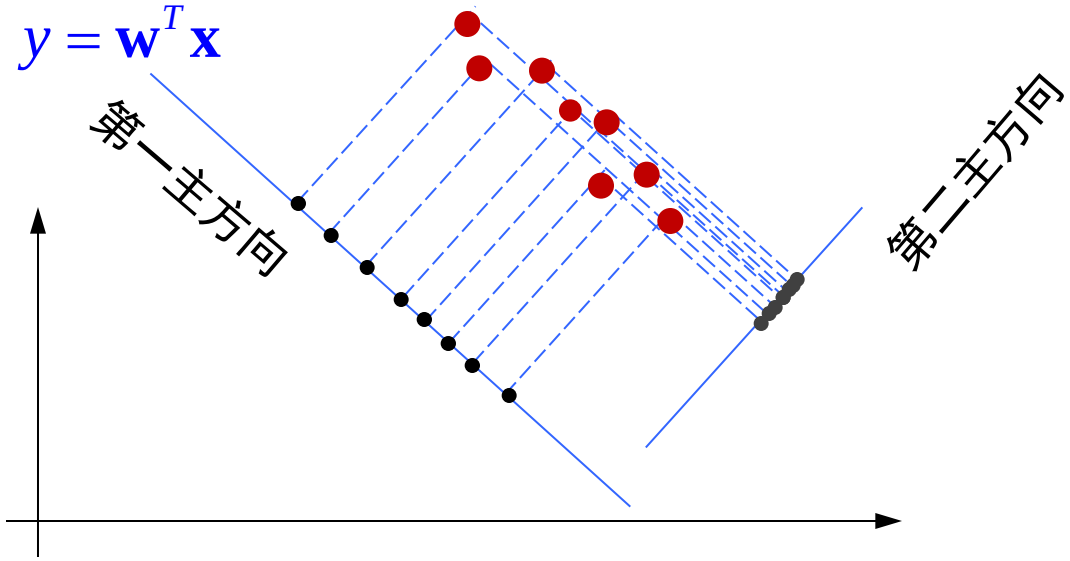
- 计算数据均值： $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$
- 计算数据的协方差矩阵：$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$
- 对矩阵C进行特征值分解，并取最大的m个特征值($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$)对应的特征向量$\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m$，组成投影矩阵 $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m] \in R^{d\times m}$
- 将每一个数据进行投影：$\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i \in R^m,\ i = 1, 2, \ldots, n$

7.3.3 主成分分析--进一步的分析

- **PCA的基本思想**
  - 如何仅用一个超平面从整体上对所有样本$\{\mathbf{x}_1, \mathbf{x}_2,\ldots,\mathbf{x}_n\}\subset R^m$进行恰当的表示？
  - 通常有如下两种思路：
    - **可区分性**：样本点在这个超平面上的投影能够尽可能地分开。（方差最大化）
    - **可重构性**：样本到这个超平面的距离都足够近。（重构误差最小化）

7.3.3 主成分分析--进一步的分析

- **PCA 一采用最大可分性观点**
  - 使所有样本点的投影尽可能地分开，则需最大化投影点的方差：



- **PCA—采用重构的观点**
  - 由**W**定义新坐标系：假定投影变换是正交变换，即新坐标系由$\mathbf{W}=[\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_d]\in R^{m\times d}$来表示 $(d<m)$，$\mathbf{w}_i$的模等于1，$\mathbf{w}_i$与$\mathbf{w}_j$两两正交。
  - 设样本点$\mathbf{x}_i$在新坐标系下的坐标为：

$$\mathbf{y}_i=[y_{i1},y_{i2},\ldots,y_{id}]^T\in R^d$$

$$y_{ij}=\mathbf{w}_j^T\mathbf{x}_i,\quad \mathbf{w}_j\in R^m,\quad j=1,2,\ldots,d \quad\Longleftrightarrow\quad \mathbf{y}_i=\mathbf{W}^T\mathbf{x}_i$$

  - 在旧坐标系下，可得$\mathbf{x}_i$的重构表示：

$$\hat{\mathbf{x}}_i=\sum_{j=1}^{d}y_{ij}\mathbf{w}_j=\mathbf{W}\mathbf{y}_i,\quad i=1,2,\ldots,n$$

- **PCA —采用重构的观点**
  - 重构误差：

$$\sum_{i=1}^{n}\left\|\mathbf{x}_i-\hat{\mathbf{x}}_i\right\|_2^2=\sum_{i=1}^{n}\left\|\mathbf{x}_i-\mathbf{W}\mathbf{y}_i\right\|_2^2$$

$$=\sum_{i=1}^{n}\left(\left(\mathbf{W}\mathbf{y}_i\right)^T\mathbf{W}\mathbf{y}_i-2\mathbf{x}_i^T\mathbf{W}\mathbf{y}_i+\mathbf{x}_i^T\mathbf{x}_i\right)$$

$$\left(\because \mathbf{W}^T\mathbf{W}=\mathbf{I}\right)=\sum_{i=1}^{n}\left(\mathbf{y}_i^T\mathbf{y}_i-2\mathbf{y}_i^T\mathbf{y}_i+\mathbf{x}_i^T\mathbf{x}_i\right)$$

$$\left(\because \mathbf{y}_i=\mathbf{W}^T\mathbf{x}_i\right)=-\sum_{i=1}^{n}\mathbf{y}_i^T\mathbf{y}_i+const=-\sum_{i=1}^{n}\left(\mathbf{W}^T\mathbf{x}_i\right)^T\left(\mathbf{W}^T\mathbf{x}_i\right)+const$$

$$=-tr\left(\sum_{i=1}^{n}\left(\mathbf{W}^T\mathbf{x}_i\right)^T\left(\mathbf{W}^T\mathbf{x}_i\right)\right)+const$$

$$\left(\because tr(\mathbf{AB})=tr(\mathbf{BA})\right)\left(\because tr(\mathbf{A})+tr(\mathbf{B})=tr(\mathbf{A}+\mathbf{B})\right)=-tr\left(\mathbf{W}^T\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\mathbf{W}\right)+const$$

对非零均值化数据的完整推导见
C. M. Bishop的教材PRML第12章

- **PCA —采用重构的观点**

令 $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n]\in R^{m\times n}$

$$\sum_{i=1}^{n}\left\|\mathbf{x}_i-\hat{\mathbf{x}}_i\right\|_2^2=-tr\left(\mathbf{W}^T\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\mathbf{W}\right)+const$$

$$=-tr\left(\mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W}\right)+const$$

– 假定数据已经零均值化，即 $\bar{\mathbf{x}}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i=\mathbf{0}$

$$\mathbf{X}\mathbf{X}^T=\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T=\sum_{i=1}^{n}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)^T=n\mathbf{C}$$

于是，获得主成分分析的最优化模型：

$$\max_{\mathbf{W}\in R^{m\times d}}\ tr\left(\mathbf{W}^T\mathbf{C}\mathbf{W}\right)=\sum_{i=1}^{d}\mathbf{w}_i^T\mathbf{C}\mathbf{w}_i,\quad \text{s.t. } \mathbf{W}^T\mathbf{W}=\mathbf{I}$$

7.3.3 主成分分析--进一步的分析

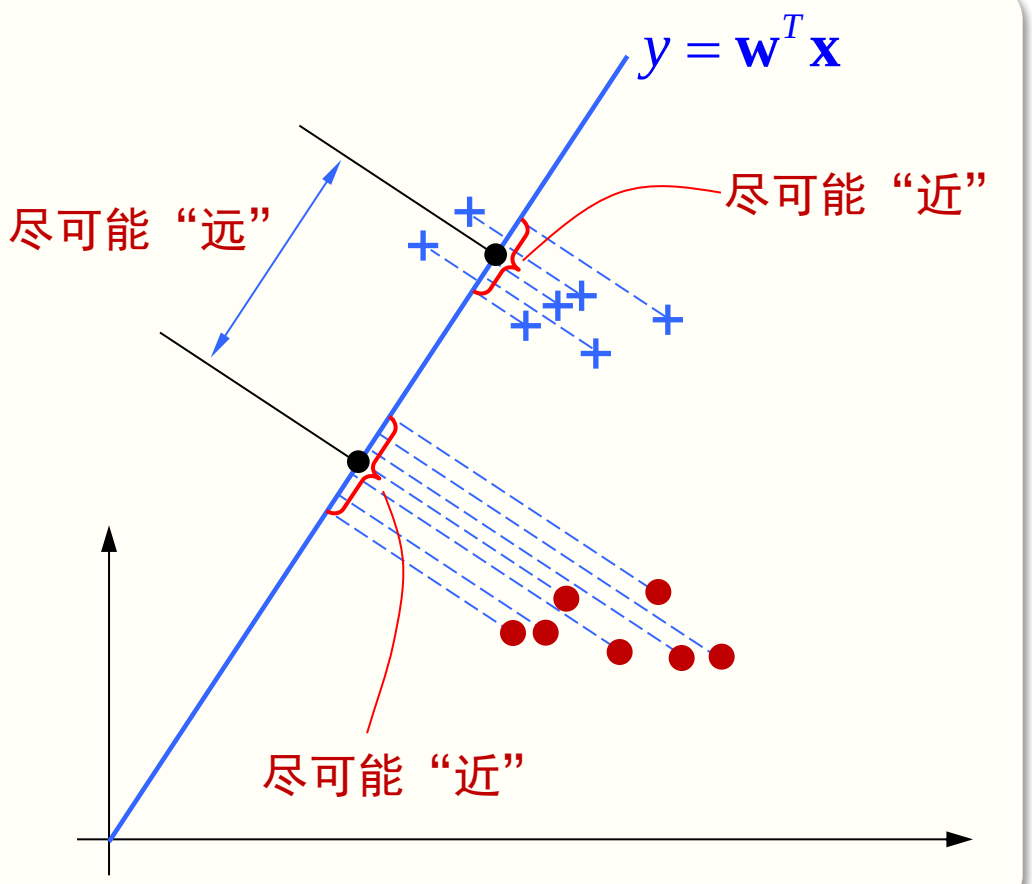
- **讨论：**
  - 降低至多少维：$\frac{\sum_{i=1}^{d}\lambda_i}{\sum_{i=1}^{m}\lambda_i} \geq t$ (比如，t=95%)
    - 也可以采用交叉验证，结合最近邻分类器来选择合适的维度d。
  - 舍弃$m-d$个特征值对应的特征向量导致了维数缩减。
    - 舍弃这些信息之后能使样本的采样密度增大，这正是降维的重要动机。
    - 另外，当数据受到噪声影响时，最小的特征值所对应的特征向量往往与噪声有关，将它们舍弃可在一定程度上起到去噪的效果。

7.4 线性判别分析

- **算法思想**
  - 线性判别分析（Linear Discriminant Analysis, LDA）是一种经典的线性判别学习方法。
  - LDA的思想较直观：对于两类分类问题，给定训练集，设法将样本投影到一条直线上，使得同类样本的投影点尽可能接近，不同类样本的投影点尽可能相互远离。
  - 对新样本分类时，将其投影到这条直线上，再根据投影点的位置来判断其类别。

7.4 线性判别分析

- **算法思想**



- ✓ 动机：寻找一组投影方向，使样本在投影之后（即在新坐标系下）类内样本点尽可能靠近，类间样本点尽可能远离，提升样本表示的分类鉴别能力。
- ✓ 投影方向数小于原始数据的维度，因此投影样本即相当于将样本在子空间内进行表示，从而达到降维的目的。

7.4 线性判别分析

- **算法思想**
  - 样本集$D=\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2),\ldots,(\mathbf{x}_n, y_n)\}$，$y_i\in\{0,1\}$，令$\mathbf{X}_i$、$\boldsymbol{\mu}_i$、$\boldsymbol{\Sigma}_i$分别表示第$i\in\{0,1\}$类的样本集合、均值向量、协方差矩阵。
  - 若将数据投影到方向$\mathbf{w}$上，则两类样本的中心在直线上的投影分别为$\mathbf{w}^T\boldsymbol{\mu}_0$和$\mathbf{w}^T\boldsymbol{\mu}_1$；两类样本的协方差分别为$\mathbf{w}^T\boldsymbol{\Sigma}_0\mathbf{w}$和$\mathbf{w}^T\boldsymbol{\Sigma}_1\mathbf{w}$。
  - 欲使同类样本的投影点尽可能接近，可让同类样本投影点的方差尽可能小，即$\mathbf{w}^T\boldsymbol{\Sigma}_0\mathbf{w}+\mathbf{w}^T\boldsymbol{\Sigma}_1\mathbf{w}$尽可能小。
  - 欲使异类样本的投影点尽可能远离，可让类中心点之间的距离尽可能大，即$(\mathbf{w}^T\boldsymbol{\mu}_0-\mathbf{w}^T\boldsymbol{\mu}_1)^2$尽可能大。

7.4 线性判别分析

- **算法思想**

最大化如下目标函数：

两个类的中心距离尽可能远

两类的类内方差尽可能小

$$J = \frac{\left\| \mathbf{w}^T \boldsymbol{\mu}_0 - \mathbf{w}^T \boldsymbol{\mu}_1 \right\|_2^2}{\mathbf{w}^T \boldsymbol{\Sigma}_0 \mathbf{w} + \mathbf{w}^T \boldsymbol{\Sigma}_1 \mathbf{w}}$$

$$= \frac{\mathbf{w}^T (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \mathbf{w}}{\mathbf{w}^T (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1) \mathbf{w}}$$

- 根据上述算法思想，我们可以定义一些量，从而将算法进行推广。

7.4 线性判别分析

- **LDA算法**
  - 类内散度矩阵：
$$\mathbf{S}_w = \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1$$
$$= \sum_{\mathbf{x}\in X_0} (\mathbf{x}-\boldsymbol{\mu}_0)(\mathbf{x}-\boldsymbol{\mu}_0)^T + \sum_{\mathbf{x}\in X_1} (\mathbf{x}-\boldsymbol{\mu}_1)(\mathbf{x}-\boldsymbol{\mu}_1)^T$$
  - 类间散度矩阵：
$$\mathbf{S}_b = (\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^T$$
  - 目标函数重写为（**广义Rayleigh商**）：
$$J = \frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}}$$

注意：J 的值与向量的长度无关，只与其方向有关，不失一般性可令**w**为单位长度的向量。

7.4 线性判别分析

- **LDA算法**
  - 由于目标函数值与长度无关（只与方向有关），因此可采用一种更直观的方法：令 $\mathbf{w}^T \mathbf{S}_w \mathbf{w} = 1$:

$$\max \ \mathbf{w}^T \mathbf{S}_b \mathbf{w}, \quad \text{s.t.} \quad \mathbf{w}^T \mathbf{S}_w \mathbf{w} = 1$$

  - 根据拉格良日乘子法，于是有：

$$\mathbf{S}_b \mathbf{w} = \lambda \mathbf{S}_w \mathbf{w} \quad \Rightarrow \quad \mathbf{S}_w^{-1} \mathbf{S}_b \mathbf{w} = \lambda \mathbf{w}$$

上式表明： $\mathbf{w}$为是矩阵 $\mathbf{S}_w^{-1} \mathbf{S}_b$ 的特征向量。

7.4 线性判别分析

- **LDA算法：构造性求解方法**

$$\mathbf{S}_b = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T$$

$$\mathbf{S}_b\mathbf{w} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T\mathbf{w} = s\cdot(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1), \quad s = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T\mathbf{w} \in R$$

标量

上式表明： $\mathbf{S}_b\mathbf{w}$ 方向与 $\boldsymbol{\mu}_0-\boldsymbol{\mu}_1$ 的方向相同。 不妨令：

$$\mathbf{S}_b\mathbf{w} = \lambda(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$$

$\mathbf{S}_b\mathbf{w}=\lambda\mathbf{S}_w\mathbf{w}$ ⇒ $\mathbf{w} = \mathbf{S}_w^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$

- **多类LDA算法 (设类别数为c)**
  - 全局散度矩阵：$\mathbf{S}_t = \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T$，$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$
  - 类内散度矩阵：$\mathbf{S}_w = \sum_{j=1}^{c} \mathbf{S}_{wj}$，

    其中，$\mathbf{S}_{wj} = \sum_{\mathbf{x}\in X_j} (\mathbf{x} - \boldsymbol{\mu}_j)(\mathbf{x} - \boldsymbol{\mu}_j)^T$，$\boldsymbol{\mu}_j = \frac{1}{n_j}\sum_{\mathbf{x}\in X_j} \mathbf{x}$
  - 类间散度矩阵：

    $$\mathbf{S}_b = \mathbf{S}_t - \mathbf{S}_w = \sum_{j=1}^{c} n_j (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T$$

    其中，n_j为属于第j类的样本个数。

思考题：试证明矩阵$\mathbf{S}_b$的秩小于等于$d-1$！

7.4 线性判别分析

- **多类LDA算法**

最大化投影后的距离

Problem 1:

（迹比值最大化）

$$\max \frac{\mathrm{tr}\left(\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right)}{\mathrm{tr}\left(\mathbf{W}^T\mathbf{S}_w\mathbf{W}\right)} = \frac{\sum_i \mathbf{w}_i^T\mathbf{S}_b\mathbf{w}_i}{\sum_i \mathbf{w}_i^T\mathbf{S}_w\mathbf{w}_i}, \quad \text{s.t.} \quad \mathbf{W}^T\mathbf{W} = \mathbf{I}$$

最大化投影后数据分布的体积

Problem 2:

（行列式比值最大化）

$$\max \frac{\left|\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right|}{\left|\mathbf{W}^T\mathbf{S}_w\mathbf{W}\right|}, \quad \text{s.t.} \quad \mathbf{W}^T\mathbf{W} = \mathbf{I}$$

行列式

7.4 线性判别分析

- **多类LDA算法**
  - Problem 1与 Problem 2的解是不同的，Problem 2的解可以通过求解如下广义特征值问题得到：

$$\mathbf{S}_b\mathbf{w}=\lambda\mathbf{S}_w\mathbf{w}$$

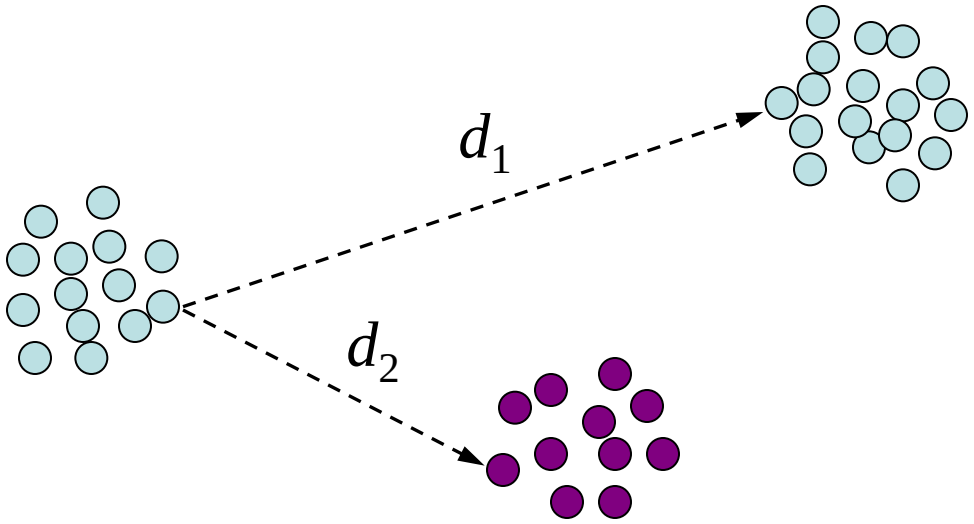
  - **Problem 1的求解较复杂**，可以参考如下文献：

Shiming Xiang, Feiping Nie, Changshui Zhang. Learning a Mahalanobis distance metric for data clustering and classification. Pattern Recognition, 41(12), Pages 3600 - 3612, 2008

$$\max \frac{\mathrm{tr}(\mathbf{W}^T\mathbf{S}_b\mathbf{W})}{\mathrm{tr}(\mathbf{W}^T\mathbf{S}_w\mathbf{W})}, \quad s.t. \quad \mathbf{W}^T\mathbf{W}=\mathbf{I}$$

7.5 局部线性判别分析

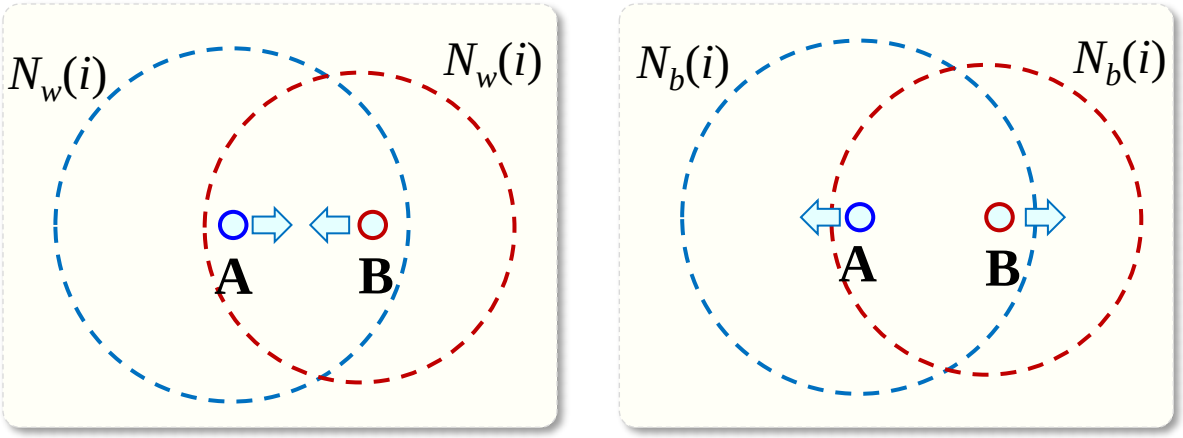
- **局限性**
  - In each class, the distribution of data is Gaussian



LDA hopes $d_1 < d_2$. **It is difficult, and sometimes impossible!**

7.5 局部线性判别分析

- **Techniques of Local Analysis**
  - Neighborhood constraints （方法一）
  - Locally weighting （方法二）
    - Weighting for 1-NN
    - Local Fisher discriminant analysis
- Basic Motivation



within class　　between classes

7.5 局部线性判别分析

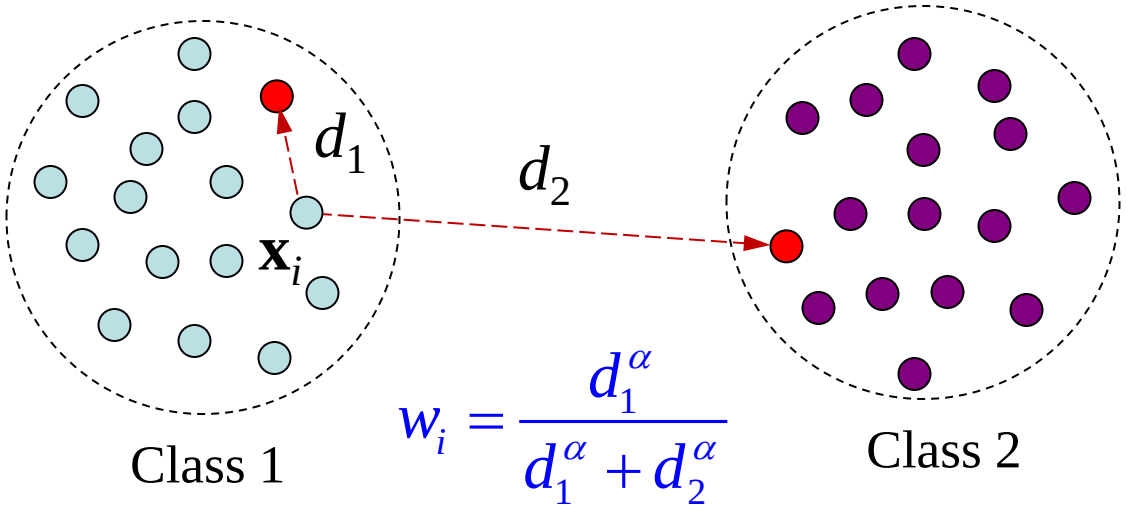
- **Modify S_w and S_b**
  - Neighborhood Constraints:

类内散度矩阵 $$\mathbf{S}_w = \sum_{\substack{y_i = y_j \\ \mathbf{x}_i \in N(\mathbf{x}_j),\, \mathbf{x}_j \in N(\mathbf{x}_i)}} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$$

类间散度矩阵 $$\mathbf{S}_b = \sum_{\substack{y_i \neq y_j \\ \mathbf{x}_i \in N(\mathbf{x}_j),\mathbf{x}_j \in N(\mathbf{x}_i)}} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$$

7.5 局部线性判别分析

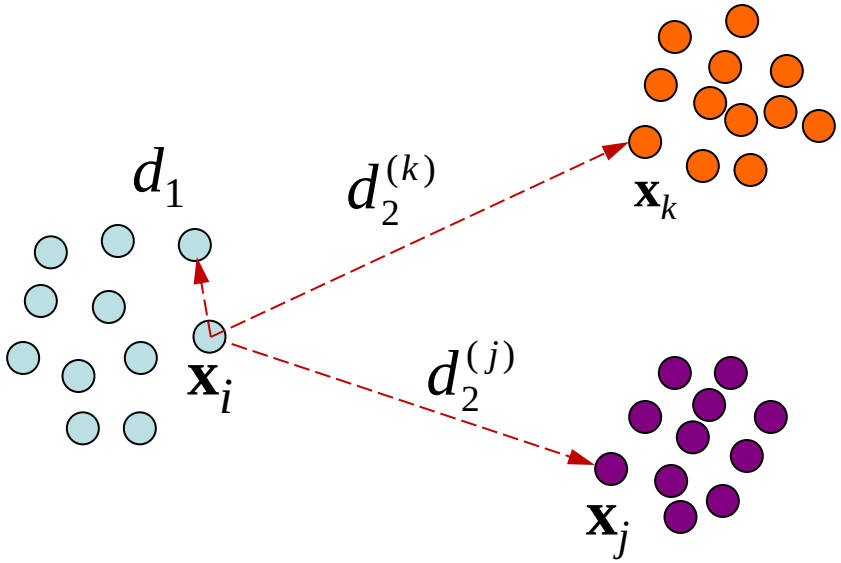
- **Nearest Neighbor Discriminant Analysis, NNDA**
  - **近邻加权**
    - A problem in neighborhood constraints is the selection of the number of nearest neighbors (k)



两类最近邻加权

7.5 局部线性判别分析

- **Nearest Neighbor Discriminant Analysis, NNDA**



$$w_i = \frac{d_1^{\alpha}}{d_1^{\alpha} + (\min\{d_2^{(j)}\})^{\alpha}}, \quad (0<\alpha<1)$$

多类最近邻加权

7.5 局部线性判别分析

- **Nearest Neighbor Discriminant Analysis, NNDA**

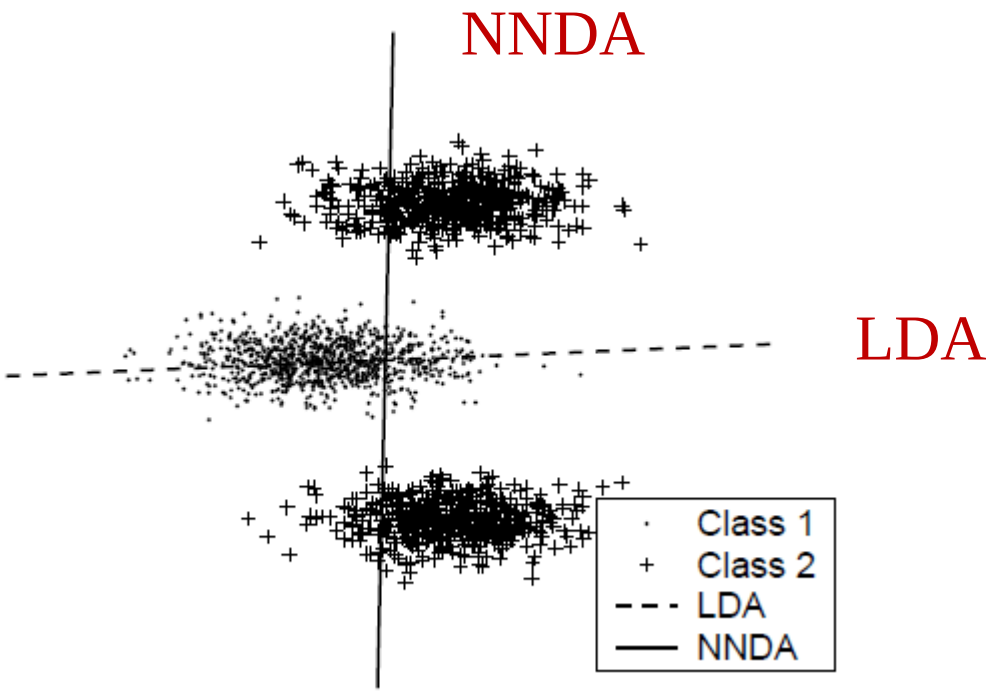
$$\mathbf{S}_w = \sum_{\substack{y_i = y_j \\ \mathbf{x}_j \in N_{1-nn}(\mathbf{x}_i),\, i=1,2,\cdots,n}} w_i (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$$

$$\mathbf{S}_b = \sum_{\substack{y_i \neq y_j \\ \mathbf{x}_j \in N_{1-nn}(\mathbf{x}_i),\, i=1,2,\cdots,n}} w_i (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$$

Xipeng Qiu, Lide Wu: Stepwise Nearest Neighbor Discriminant Analysis. IJCAI 2005: 829-834

7.5 局部线性判别分析

- **Nearest Neighbor Discriminant Analysis, NNDA**



NNDA finds the correct projection direction, but LDA fails !

7.5 局部线性判别分析

- **Local Fisher Discriminant Analysis, LFDA**
  - Motivation
    - LFDA does not impose far-apart data pairs of the same class to be close, by which local structure of the data tends to be preserved.
    - 邻域加权 (Locally Weighting)

Masashi Sugiyama, Local Fisher Discriminant Analysis for Supervised Dimensionality Reduction, ICML, 2006

7.5 局部线性判别分析

- Step1: Construct an affine matrix for *n* data points:

$$\mathbf{A}=\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{pmatrix}$$

$$A_{ij}=\begin{cases} 1, & \text{if } \mathbf{x}_j \text{ is a neighbor of } \mathbf{x}_i \\ 0, & \text{otherwise} \end{cases}, \text{ or}$$

$$A_{ij}=\begin{cases} \exp(-\|\mathbf{x}_j-\mathbf{x}_i\|^2/(2\sigma^2)), & \text{if } \mathbf{x}_j \text{ is a neighbor of } \mathbf{x}_i \\ 0, & \text{otherwise} \end{cases}$$

7.5 局部线性判别分析

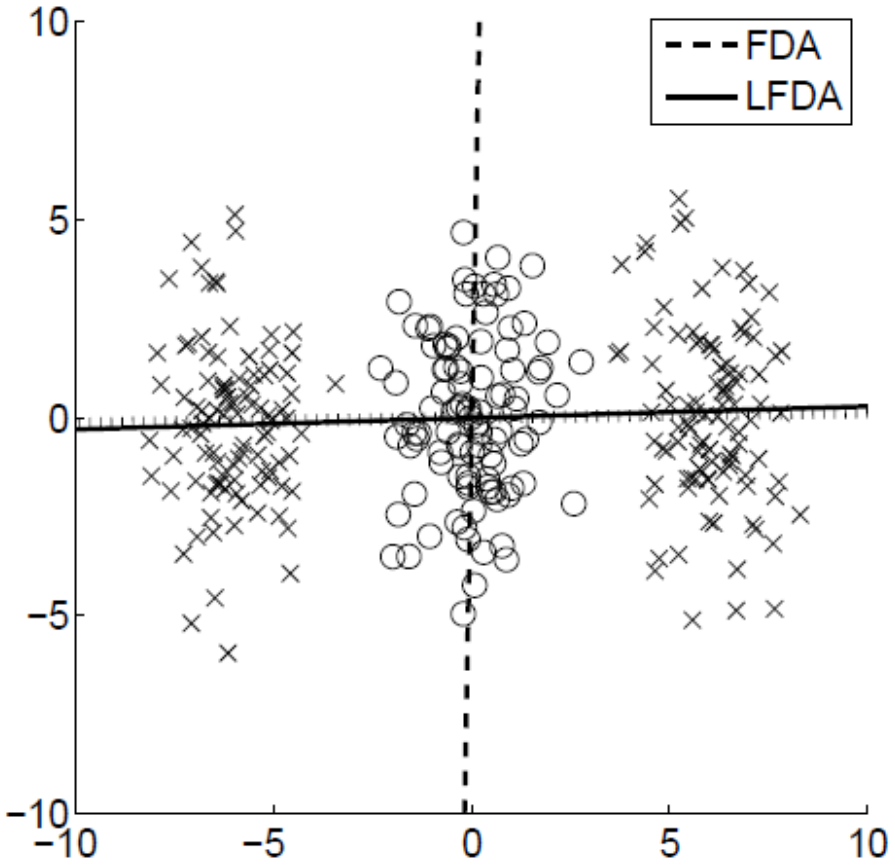
- Step2: Modify $\mathbf{S}_w$ and $\mathbf{S}_b$:

$$\mathbf{S}_w = \frac{1}{2}\sum_{i,j} \bar{A}_{ij}^{(w)} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \qquad \mathbf{S}_b = \frac{1}{2}\sum_{i,j=1}^{n} \bar{A}_{ij}^{(b)} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$$

$$\bar{A}_{ij}^{(w)} = \begin{cases} A_{ij}/n_c, & \text{if } y_i = y_j = c \\ 0, & \text{if } y_i \neq y_j \end{cases} \qquad \bar{A}_{ij}^{(b)} = \begin{cases} A_{ij}(1/n - 1/n_c), & \text{if } y_i = y_j = c \\ 1/n, & \text{if } y_i \neq y_j \end{cases}$$

7.5 局部线性判别分析

- 例子



7.6 其它维数缩减方法

- **经典方法**
  - 独立成分分析（Independent Component Analysis , ICA)
  - 典型关联分析(Canonical Correlation Analysis, CCA)
  - 2D-PCA, 2D-LDA
  - KPCA
- **流形学习方法：**LLE, Isomap, LE, LTSA,… (下次课)
- **深度学习方法**
  - PCANet
  - RBM, DBN, DBM, AutoEncoder
  - Deep CCA
  - Learning understanding Neural networks with Non-negative matrix factorization
- **应用：**Eigenface, PCA-SIFT (CVPR 2004),…

致谢

- PPT由向世明老师提供

Thank All of You!

(Questions?)

张燕明

ymzhang@nlpr.ia.ac.cn

people.ucas.ac.cn/~ymzhang

模式分析与学习课题组(PAL)

中科院自动化研究所· 模式识别国家重点实验室