人工智能概论

第九章：降维与度量学习

丁烨，计算机科学与技术学院
dingye@dgut.edu.cn

目录

基本概念

降维

- ❖ 降维（Dimensionality Reduction）
- ❖ 在某些限定条件下
- ❖ 降低特征数量、得到一组“不相关”的特征的过程
- ❖ 降维可进一步细分为特征选择和特征提取两大方法

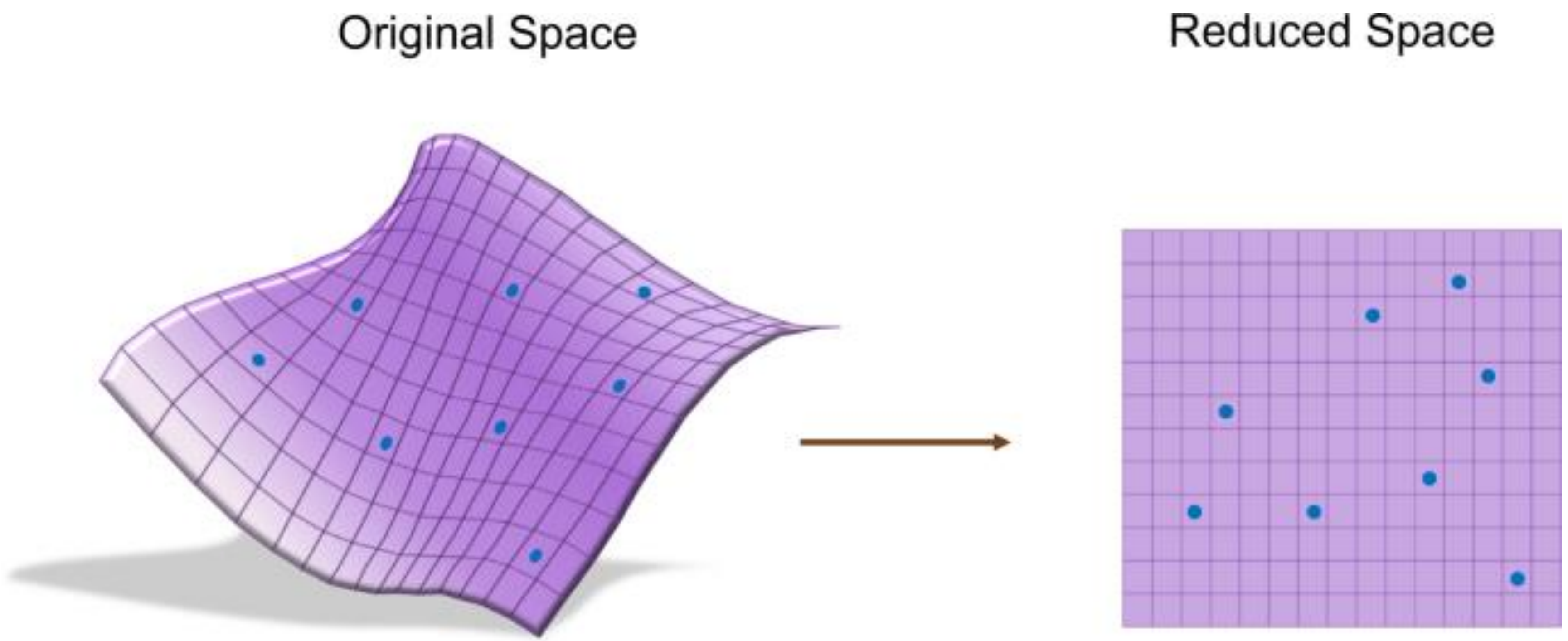
基本概念

降维

- ❖ 降维可以处理数据中存在的两个核心问题：
- ❖ 维度灾难：
- ❖ 特征维度太大，导致训练无法进行或训练过程漫长
- ❖ 特征之间存在着相互关联关系：
- ❖ 导致解的空间不稳定，从而导致模型的泛化能力弱

基本概念

降维



目录

- 基本概念
- 特征选择
- 特征提取
- 度量学习

特征选择

基本概念

- 我们能用很多特征描述一个西瓜
- 例如色泽、根蒂、敲声、纹理、触感等
- 但有经验的人往往只需看看根蒂、听听敲声就知道是否好瓜
- 换言之，对一个学习任务来说，给定特征集
- 其中有些特征可能很关键、很有用
- 另一些特征则可能没什么用

特征选择

基本概念

- ❖ 特征选择（Feature Selection）
- ❖ 从给定的特征集合中选择出相关特征子集的过程
- ❖ “数据预处理（Data Preprocessing）”的一部分
- ❖ 对当前学习任务有用的特征称为“相关特征（Relevant Feature）”
- ❖ 没什么用的特征称为“无关特征（Irrelevant Feature）”
- ❖ 核心目标：降低特征数量、去除无关特征

特征选择

基本概念

- ❖ 基础方法：
- ❖ 从初始的特征集合中选取一个包含了所有重要信息的特征子集
- ❖ 评价出它的好坏，基于评价结果产生下一个候选子集，再对其进行评价
- ❖ 这个过程持续进行下去，直至无法找到更好的候选子集为止
- ❖ 关键问题：
- ❖ 如何根据评价结果获取下一个候选特征子集？
- ❖ 如何评价候选特征子集的好坏？

特征选择

基本概念

- 第一个环节是子集搜索（Subset Search）问题
- 给定特征集合 $\{a_1, a_2, \ldots, a_d\}$，将每个特征看作一个候选子集
- 对这 d 个候选单特征子集进行评价
- 假定 $\{a_2\}$ 最优，于是将 $\{a_2\}$ 作为第一轮的选定集
- 然后在上一轮的选定集中加入一个特征，构成包含两个特征的候选子集
- ……假定在第 $k+1$ 轮时，最优的候选 $(k+1)$ 特征子集不如上一轮
- 则停止生成候选子集并将上一轮选定的 k 特征集合作为特征选择结果

特征选择

基本概念

- 这样逐渐增加相关特征的策略称为“前向（Forward）”搜索
- 类似的，若我们从完整的特征集合开始，每次尝试去掉一个无关特征
- 这样逐渐减少特征的策略称为“后向（Backward）”搜索
- 还可将前向与后向搜索结合起来
- 每一轮逐渐增加选定相关特征（这些特征后续将确定不会被去除）
- 同时减少无关特征
- 这样的策略称为“双向（Bidirectional）”搜索

特征选择

基本概念

- 显然，上述策略都是贪心的
- 因为它们仅考虑了使本轮选定集最优
- 例如在第二轮假定选择 a_5 优于 a_6，于是选定集为 $\{a_2, a_4, a_5\}$
- 然而在第四轮却可能是 $\{a_2, a_4, a_6, a_8\}$ 比所有的 $\{a_2, a_4, a_5, a_i\}$ 都更优
- 遗憾的是，若不进行穷举搜索，则这样的问题无法避免

特征选择

基本概念

- 第二个环节是子集评价（Subset Evaluation）问题
- 给定数据集 D，假定 D 中第 i 类样本所占的比例为 p_i
- 对特征子集 A，假定根据其取值将 D 分成了 V 个子集
- 于是我们可计算特征子集 A 的信息增益

$$\mathrm{Gain}(A) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v)$$

特征选择

基本概念

- 其中信息熵定义为：

$$\mathrm{Ent}(D) = -\sum_{i=1}^{|\mathcal{Y}|} p_k \log_2 p_k$$

- 信息增益 Gain(A) 越大
- 意味着特征子集 A 包含的有助于分类的信息越多
- 于是，对每个候选特征子集
- 我们可基于训练数据集 D 来计算其信息增益，以此作为评价准则

特征选择

基本概念

- 将特征子集搜索机制与子集评价机制相结合，即可得到特征选择方法
- 常见的特征选择方法大致可分为三类：
- 过滤式（Filter）
- 包裹式（Wrapper）
- 嵌入式（Embedding）

特征选择

过滤式特征选择

- 过滤式（Filter）特征选择
- 先对数据集进行特征选择，然后再训练模型
- 特征选择过程与后续训练过程无关
- 这相当于先用特征选择过程对初始特征进行“过滤”
- 再用过滤后的特征来训练模型

特征选择

过滤式特征选择

- Relief
- 一种著名的过滤式特征选择方法
- 该方法设计了一个“相关统计量”来度量特征的重要性
- 该统计量是一个向量，其每个分量分别对应于一个初始特征
- 特征子集的重要性由每个特征所对应的相关统计量分量之和来决定
- 最终只需选择比指定阈值大的相关统计量分量所对应的特征即可
- 也可指定欲选取的特征个数，然后选择相关统计量分量最大的若干特征

特征选择

过滤式特征选择

- ❖ 相关统计量
- ❖ 先从每个样本的同类样本中寻找其最近邻
- ❖ 称为“猜中近邻（Near-hit）”
- ❖ 再从每个样本的异类样本中寻找其最近邻
- ❖ 称为“猜错近邻（Near-miss）”
- ❖ 如果“猜中近邻”

特征选择

过滤式特征选择

- 若真相与其猜中近邻在某个特征上的距离小于与其猜错近邻的距离
- 则说明该特征对区分同类与异类样本是有益的
- 于是增大该特征所对应的统计量分量
- 反之，则说明该特征起负面作用
- 于是减小该特征所对应的统计量分量
- 最后，对基于不同样本得到的估计结果进行平均
- 就得到各特征的相关统计量分量
- 分量值越大，则对应特征的分类能力就越强

特征选择

包裹式特征选择

- 包裹式（Wrapper）特征选择
- 与过滤式特征选择不考虑后续模型训练过程不同
- 包裹式特征选择直接把模型的性能作为特征子集的评价准则
- 目的就是为指定算法选择最有利于其性能、“量身定做”的特征子集

特征选择

包裹式特征选择

- 一般而言，由于包裹式特征选择方法直接针对指定算法进行优化
- 因此从最终模型性能来看
- 包裹式特征选择比过滤式特征选择更好
- 但另一方面，由于在特征选择过程中需多次训练模型
- 因此包裹式特征选择的计算开销通常比过滤式特征选择大得多

特征选择

包裹式特征选择

- LVW（Las Vegas Wrapper）
- 一个典型的包裹式特征选择方法
- LVW 在拉斯维加斯方法（Las Vegas Method）框架下使用随机策略来进行子集搜索
- 并以最终模型的误差为特征子集评价准则

输入： 数据集 D;
特征集 A;
学习算法 $\mathfrak{L}$;
停止条件控制参数 T.

过程：

$E = \infty$;
$d = |A|$;
$A^* = A$;
$t = 0$;
while $t < T$ **do**
    随机产生特征子集 A';
    $d' = |A'|$;
    $E' = \text{CrossValidation}(\mathfrak{L}(D^{A'}))$;
    if $(E' < E) \vee ((E' = E) \wedge (d' < d))$ **then**
        $t = 0$;
        $E = E'$;
        $d = d'$;
        $A^* = A'$
    else
        $t = t + 1$
    end if
end while

输出： 特征子集 A^*

图 11.1 LVW 算法描述

特征选择

嵌入式特征选择

- ❖ 嵌入式（Embedding）特征选择
- ❖ 在过滤式和包裹式特征选择方法中
- ❖ 特征选择过程与模型训练过程有明显的分别
- ❖ 与此不同，嵌入式特征选择是将特征选择过程与模型训练过程融为一体
- ❖ 两者在同一个优化过程中完成
- ❖ 即在模型训练过程中自动地进行了特征选择

特征选择

嵌入式特征选择

- 给定数据集 $D=\{(x_1,y_1),(x_2,y_2),\ldots(x_m,y_m)\}$
- 其中 $x\in\mathbb{R}^d$， $y\in\mathbb{R}$
- 我们考虑最简单的线性回归模型，以平方误差为损失函数
- 则优化目标为：

$$\min_{\boldsymbol{w}}\sum_{i=1}^{m}(y_i-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i)^2$$

特征选择

嵌入式特征选择

- 当样本特征很多，而样本数相对较少时
- 上式很容易陷入过拟合
- 为了缓解过拟合问题，可对上式引入正则化项
- 若使用 L_2 范数正则化，则有：

$$\min_{\boldsymbol{w}} \sum_{i=1}^{m} (y_i - \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i)^2 + \lambda \|\boldsymbol{w}\|_2^2$$

特征选择

嵌入式特征选择

- 上式被称为“岭回归（Ridge Regression）”
- 通过引入 L_2 范数正则化，的确能显著降低过拟合的风险
- 那么，能否将正则化项中的 L_2 范数替换为 L_p 范数呢？
- 答案是肯定的
- 若令 $p=1$，即采用 L_1 范数，则有：

$$\min_{\boldsymbol{w}} \sum_{i=1}^{m} (y_i - \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i)^2 + \lambda \|\boldsymbol{w}\|_1$$

特征选择

嵌入式特征选择

$$\min_{\boldsymbol{w}} \sum_{i=1}^{m} (y_i - \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i)^2 + \lambda \|\boldsymbol{w}\|_1$$

- 上式被称为：
- LASSO（Least Absolute Shrinkage & Selection Operator）
- L_1 范数和 L_2 范数正则化都有助于降低过拟合风险
- 但前者还会带来一个额外的好处：
- 它比后者更易于获得“稀疏（Sparse）解”
- 即它求得的 w 会有更少的非零分量

目录

- 基本概念
- 特征选择
- 特征提取
- 度量学习

特征提取

基本概念

- 特征提取（Feature Extraction）
- 与特征选择类似，特征提取的核心目的也是：
- 降低特征数量
- 得到一组“不相关”的特征
- 与特征选择不同，特征提取不再保留原始特征
- 而是产生新的特征，而这些特征是原始特征的某种组合

特征提取

基本概念

- ❖ 特征提取可以看作特征选择方法的一般化：
- ❖ 特征选择假设在原始数据中：
- ❖ 特征数目浩繁，但只有少数几个真正起作用
- ❖ 而特征提取则认为：
- ❖ 在所有特征可能的组合（例如：线性组合）中
- ❖ 只有少数几个真正起作用

特征提取

基本概念

- 常见的特征提取方法：
- 主成分分析（PCA）
- 核化（Kernelized）线性降维
- 非负矩阵分解（NMF）

特征提取

主成分分析

- 主成分分析（PCA）
- 最常用的一种降维方法
- 在介绍 PCA 之前，不妨先考虑这样一个问题：
- 对于正交特征空间中的样本点
- 如何用一个超平面（直线的高维推广）对所有样本进行恰当的表达？

特征提取

主成分分析

- 若存在这样的超平面，那么它大概应具有这样的性质：
- 最近重构性：样本点到这个超平面的距离都足够近
- 最大可分性：样本点在这个超平面上的投影能尽可能分开
- 有趣的是，基于最近重构性和最大可分性
- 能分别得到主成分分析的两种等价推导

特征提取

主成分分析

- 假定数据样本进行了中心化，即 $\sum_i x_i = 0$
- 再假定投影变换后得到的新坐标系为 $\{w_1, w_2, \dots, w_d\}$
- 其中 w_i 是标准正交基向量
- 若丢弃新坐标系中的部分坐标，即将维度降低到 $d' < d$
- 则样本点 x_i 在低维坐标系中的投影是：$z_i = (z_{i1}, z_{i2}, \dots, z_{id'})$
- 其中 $z_{ij} = w_j^T x_i$ 是 x_i 在低维坐标系下第 j 维的坐标
- 若基于 z_i 来重构 x_i，则会得到：$\hat{x}_i = \sum_{j=1}^{d'} z_{ij} w_j$

特征提取

主成分分析

- 考虑整个训练集
- 原样本点 x_i 与基于投影重构的样本点 $\hat{x}_i$ 之间的距离为：

$$\sum_{i=1}^{m}\left\|\sum_{j=1}^{d'} z_{ij}\boldsymbol{w}_j-\boldsymbol{x}_i\right\|_2^2=\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i-2\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i+\mathrm{const}$$

$$\propto -\operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\left(\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right)\mathbf{W}\right).$$

特征提取

主成分分析

- 根据最近重构性，上式应被最小化
- 考虑到 w_j 是标准正交基，$\sum_i x_i x_i^T$ 是协方差矩阵，有：

$$\begin{aligned} \min_{\mathbf{W}} \quad & -\operatorname{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}) \\ \text{s.t.} \quad & \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}. \end{aligned}$$

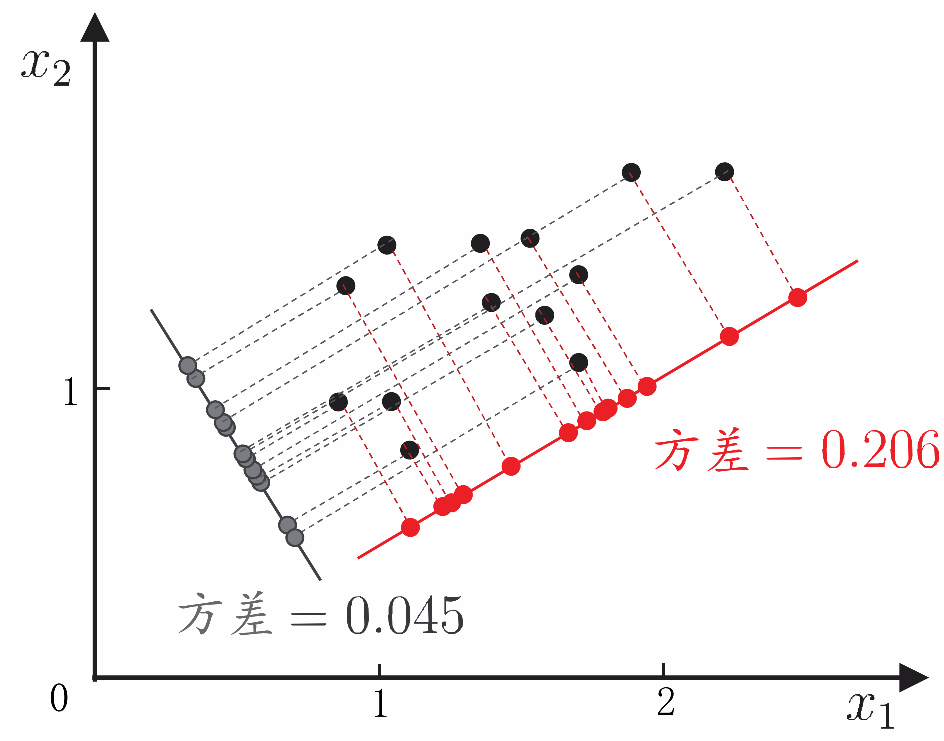
- 这就是主成分分析的优化目标

特征提取

主成分分析

- 从最大可分性出发，能得到主成分分析的另一种解释
- 我们知道，样本点 x_i 在新空间中超平面上的投影是 $W^T x_i$
- 若所有样本点的投影能尽可能分开
- 则应该使投影后样本点的方差最大化

$$\max_{\mathbf{W}} \quad \mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W})$$
$$\text{s.t.} \quad \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}.$$



特征提取

主成分分析

$$\max_{\mathbf{W}} \quad \mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}) \qquad \Longleftrightarrow \qquad \min_{\mathbf{W}} \quad -\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W})$$

$$\text{s.t.} \quad \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}. \qquad\qquad \text{s.t.} \quad \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}.$$

- ❖ 最近重构性和最大可分性的推导很明显在数学上是等价的
- ❖ 使用拉格朗日乘子法可得：

$$\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W} = \lambda\mathbf{W}.$$

特征提取

主成分分析

输入: 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
低维空间维数 d'.

过程:

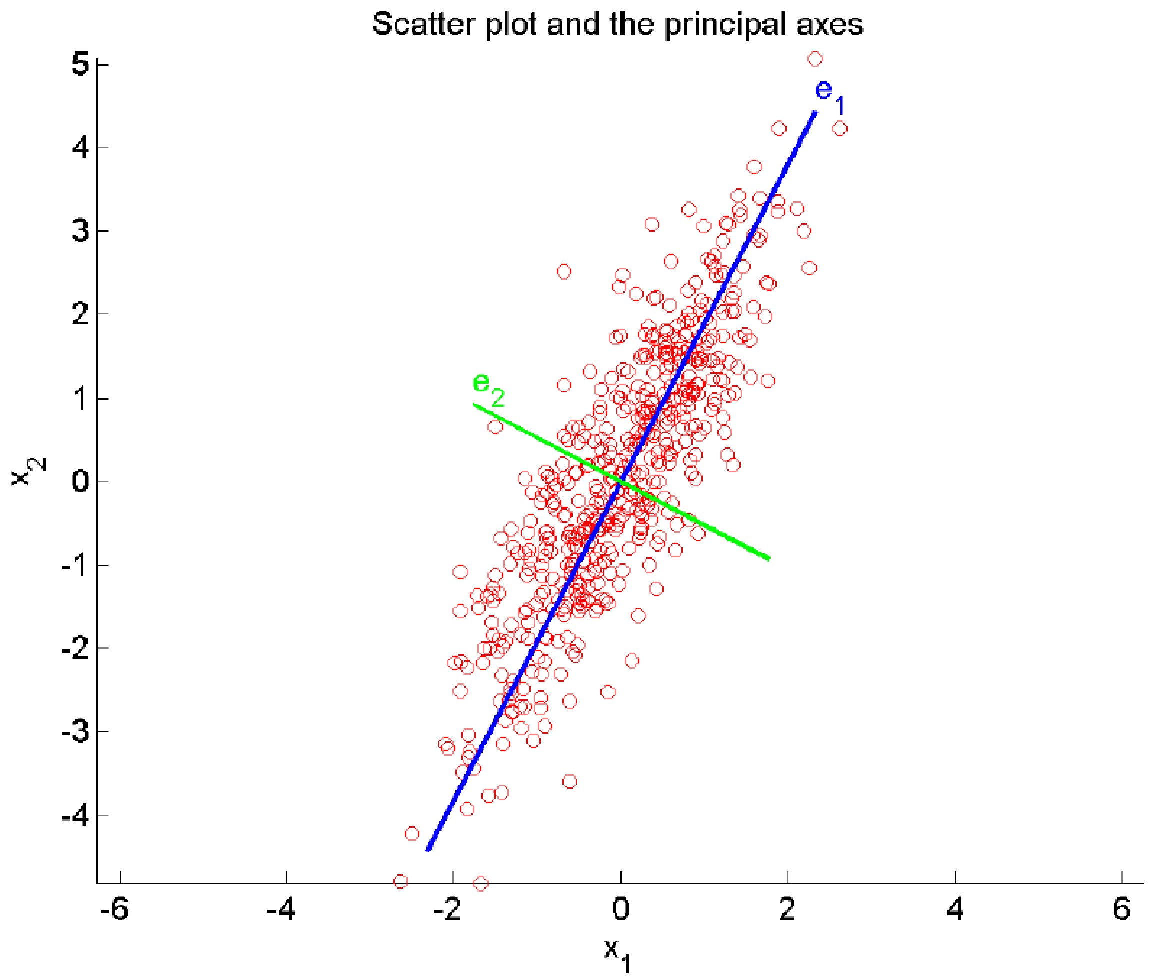
1: 对所有样本进行中心化: $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i$;
2: 计算样本的协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$;
3: 对协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 做特征值分解;
4: 取最大的 d' 个特征值所对应的特征向量 $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'}$.

输出: 投影矩阵 $\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'})$.

图 10.5 PCA 算法

特征提取

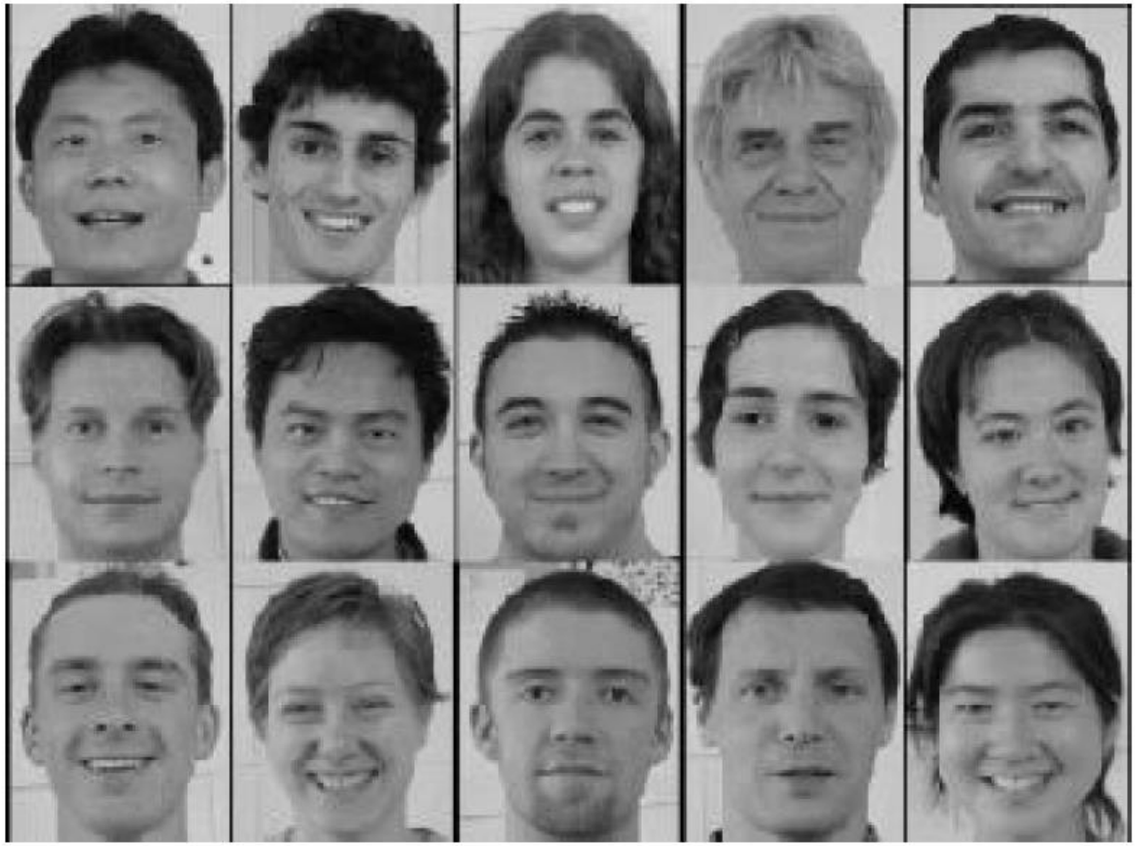
主成分分析

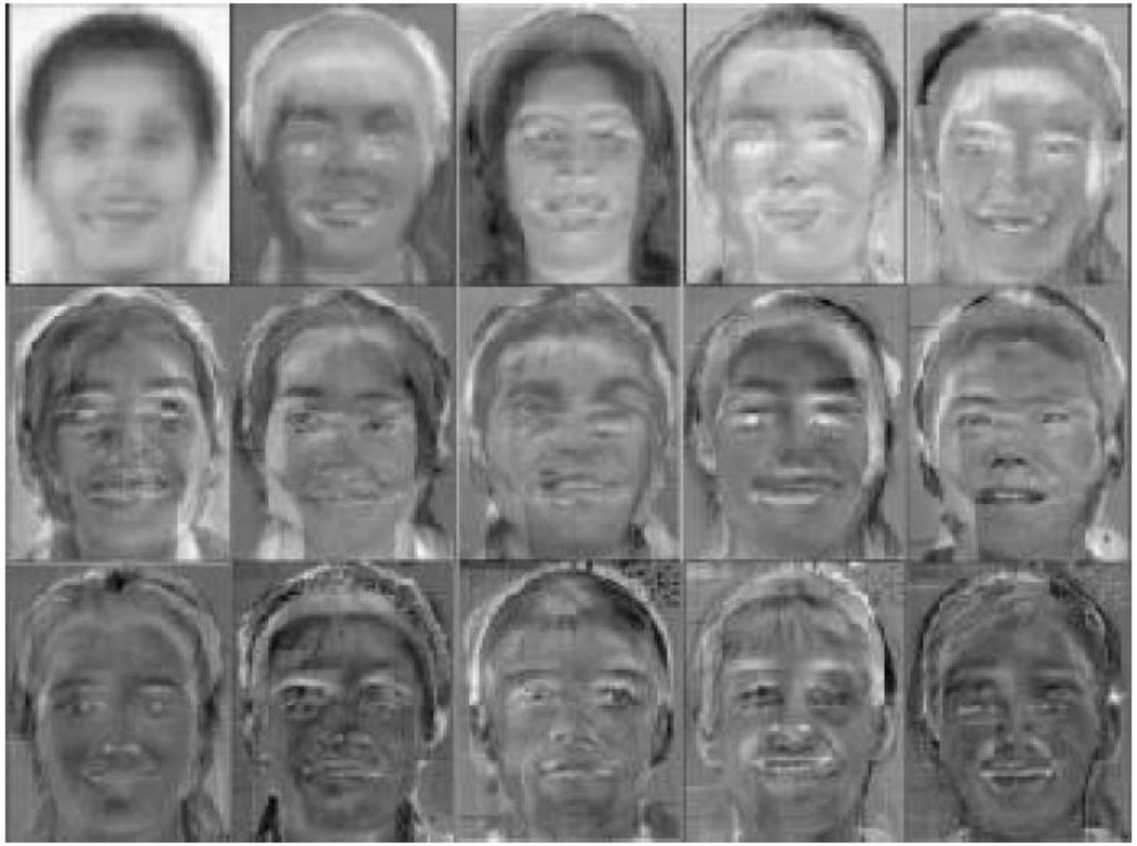


Scatter plot and the principal axes
5
4
3
2
1
0
-1
-2
-3
-4
-6
-4
-2
0
2
4
6
x_1
x_2
e_1
e_2

特征提取

主成分分析



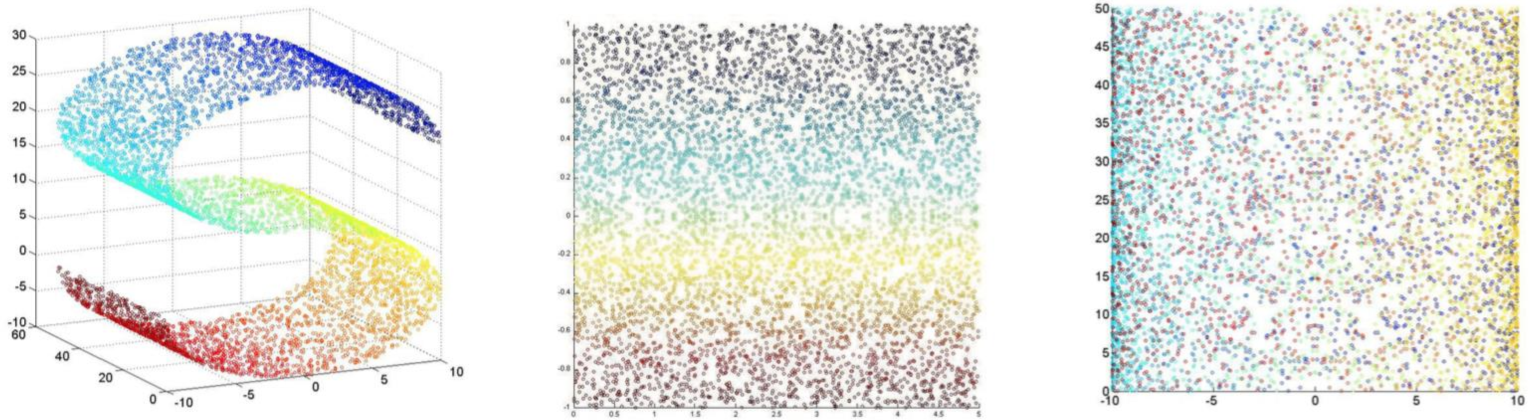


“本征脸（Eigenface）”

特征提取

核化线性降维

- 线性阵维方法假设从高维空间到低维空间的函数映射是线性的
- 然而，在现实任务中，可能需要非线性映射才能找到恰当的低维嵌入



(a) 三维空间中的观察　(b) 本真二维结构　(c) PCA 降维结果

图 10.6　三维空间中观察到的 3000 个样本点，是从本真二维空间中矩形区域采样后以 S 形曲面嵌入，此情形下线性降维会丢失低维结构．图中数据点的染色显示出低维空间的结构．

特征提取

核化线性降维

- 非线性降维的一种常用方法
- 是基于核技巧对线性降维方法进行“核化”（Kernelized）
- 假定我们将在高维特征空间中把数据投影到由 W 确定的超平面上
- 即 PCA 欲求解：

$$\left(\sum_{i=1}^{m} z_i z_i^{\mathrm{T}}\right) \mathbf{W} = \lambda \mathbf{W}.$$

特征提取

核化线性降维

- 假定 z_i 是由原始属性空间中的样本点 x_i 通过映射 ϕ 产生，即：

$$z_i = \phi(x_i)$$

- 若 ϕ 能被显式表达出来，则通过它将样本映射至高维特征空间
- 再在特征空间中实施 PCA 即可

特征提取

核化线性降维

- 但一般情形下，我们不清楚 ϕ 的具体形式，于是引入核函数：

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^{\mathrm{T}} \phi(\boldsymbol{x}_j).$$

- 最终即可求出原始 PCA 的解
- 这个方法叫做：核主成分分析（Kernelized PCA，KPCA）

特征提取

非负矩阵分解

- 非负矩阵分解（Non-negative Matrix Factorization，NMF）
- 顾名思义，就是将非负的大矩阵分解成两个非负的小矩阵，使得：

$$V \approx WH$$

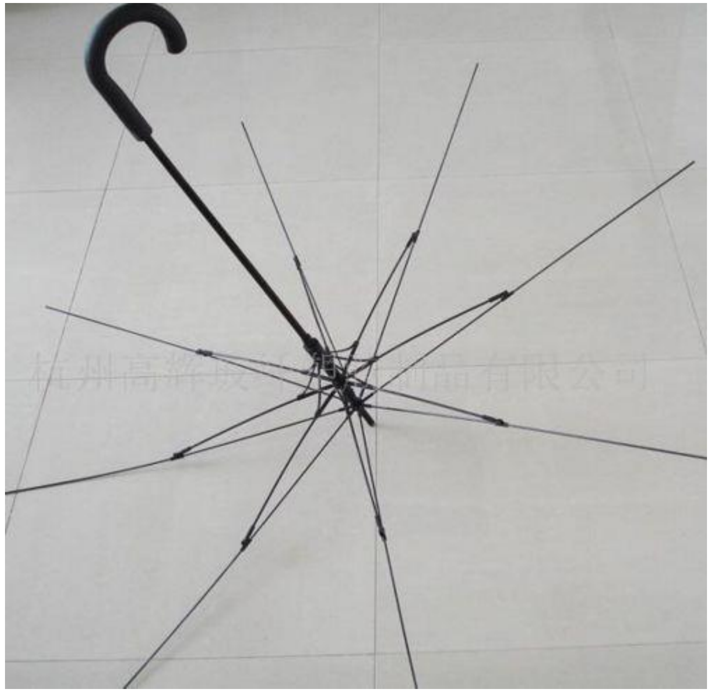
- V 是 $n * m$ 维的矩阵
- W 是 $n * k$ 维的矩阵，即 k 个基向量
- H 是 $k * m$ 的矩阵，每一列为 V 投影到 W 上得到的向量

特征提取

非负矩阵分解

- PCA 的原数据是有正有负的
- PCA 的基向量是指向四面八方的，相互正交着

- NMF 的原数据只分布在非负子空间
- 它的基向量则在这个非负子空间靠近边缘的区域，像一组长短不一、间隔不一的伞骨



特征提取

非负矩阵分解

- NMF 如何求解？
- 有点复杂，可以参考：
- https://en.wikipedia.org/wiki/Non-negative_matrix_factorization
- http://oa.ee.tsinghua.edu.cn/~zhangyujin/Download-Paper/C151=Dzxb-08.pdf

目录

- 基本概念
- 特征选择
- 特征提取
- 度量学习

度量学习

基本概念

- 对高维数据进行降维的主要目的是希望找到一个合适的低维空间
- 在此空间中进行学习能比原始空间性能更好
- 事实上，每个空间对应了在样本特征上定义的一个距离度量
- 而寻找合适的空间，实质上就是在寻找一个合适的距离度量
- 那么，为何不直接尝试“学习”出一个合适的距离度量呢？
- 这就是度量学习（Metric Learning）的基本动机

度量学习

基本概念

- 欲对距离度量进行学习，必须有一个便于学习的距离度最表达形式
- 我们在之前的课程中给出了很多种距离度量的表达式
- 但它们都是“固定的”、没有可调节的参数
- 因此不能通过对数据样本的学习来加以改善
- 为此，我们先来做一个推广

度量学习

基本概念

- 对两个 d 维样本 x_i 和 x_j，它们之间的平方欧氏距离可写为：

$$\mathrm{dist}^2_{\mathrm{ed}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2 = dist^2_{ij,1} + dist^2_{ij,2} + \cdots + dist^2_{ij,d},$$

- 其中 $dist_{ijk}$ 表示 x_i 与 x_j 在第 k 维上的距离
- 若假定不同特征的重要性不同，则可引入特征权重 w，得到：

$$\begin{aligned}\mathrm{dist}^2_{\mathrm{wed}}(\boldsymbol{x}_i, \boldsymbol{x}_j) &= ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2 = w_1 \cdot dist^2_{ij,1} + w_2 \cdot dist^2_{ij,2} + \cdots + w_d \cdot dist^2_{ij,d} \\ &= (\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \mathbf{W} (\boldsymbol{x}_i - \boldsymbol{x}_j),\end{aligned}$$

度量学习

基本概念

- 上式中的 W 可通过学习确定，但我们还能再往前走一步
- W 的非对角元素均为零，这意味着坐标轴是正交的，即特征之间无关
- 但现实问题中往往不是这样
- 例如考虑西瓜的“重量”和“体积”这两个特征
- 它们显然是正相关的，其对应的坐标轴不再正交

度量学习

基本概念

- 为此，将上式中的 W 替换为一个普通的半正定对称矩阵 M
- 于是就得到了“马氏距离（Mahalanobis Distance）”

$$\mathrm{dist}^2_{\mathrm{mah}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \mathbf{W} (\boldsymbol{x}_i - \boldsymbol{x}_j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2_{\mathbf{M}},$$

- 其中 M 亦称“度量矩阵”，而度量学习则是对 M 进行学习
- 求解 M 的具体过程和优化目标请参考教科书第 10.6 节

总结

- 「西瓜书」周志华《机器学习》，清华大学出版社
- https://item.jd.com/12762673.html

- 「南瓜书」谢文睿、秦州《机器学习公式详解》，人民邮电出版社
- https://github.com/datawhalechina/pumpkin-book/

总结

- Scikit-Learn
- https://scikit-learn.org

- TensorFlow
- https://www.tensorflow.org


HAPPY
Mother's
DAY

Thank You!

丁烨，计算机科学与技术学院
dingye@dgut.edu.cn