模 式 识 别

第5章：特征选择与特征提取

主讲人：张治国
zhiguozhang@hit.edu.cn

本章内容

- ✓ 特征选择与特征提取的基本概念
- 类别可分性判据
  - 基于距离的可分性判据
  - 基于散布矩阵的可分性判据
- 特征选择
  - 分支定界法
  - 次优搜索算法
- ✓ 主成分分析
  - 算法、推导和相关问题
- ✓ 基于Fisher准则的可分性分析
  - 算法、推导和相关问题

特征的选择和提取

- 特征的选择和提取与应用直接相关，对分类和聚类都起至关重要的作用。

分类桃子和橙子

特征：{颜色、形状、外表、酸度、重量……}

↓ 特征选择和提取

特征矢量：{颜色、形状}

聚类不同的动物

羊，狗，猫，鲑鱼，鲸鱼，海鸥，麻雀，水母，青蛙，蜥蜴

特征：{繁殖方式、生活地点、幼仔喂养方式，大小、体重、食物……}

↓ 特征选择和提取

特征矢量：{繁殖方式、幼仔喂养方式}

羊，狗，猫，鲸鱼

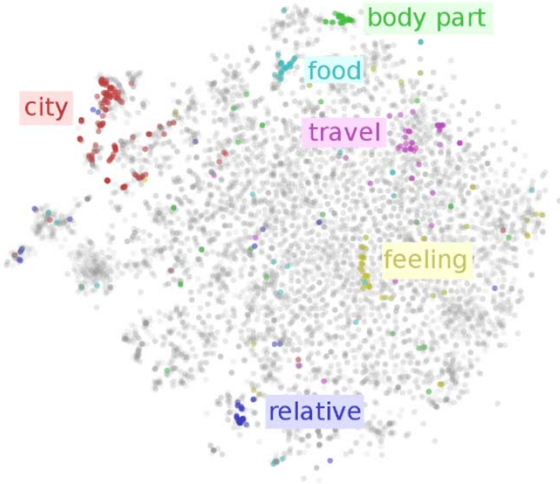
鲑鱼，海鸥，麻雀，水母，青蛙，蜥蜴

高维特征

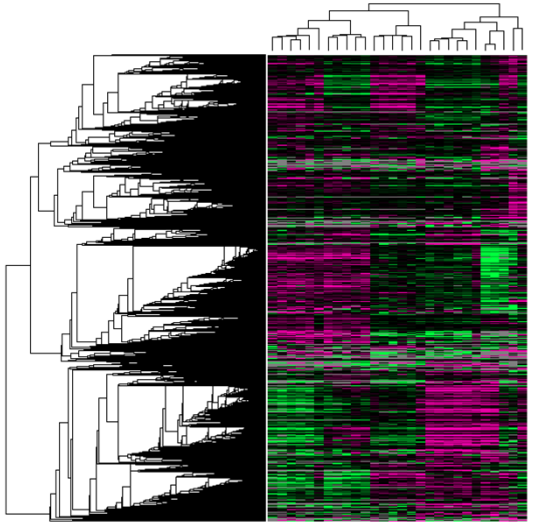
- 随着应用的扩展和深入，以及数据采集与分析技术发展，特征维度经历了飞速增长。

计算机视觉

自然语言处理

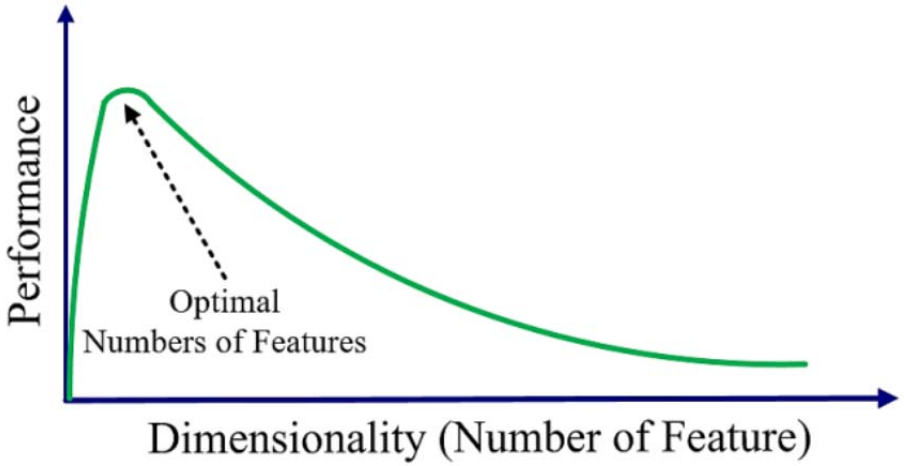


生物信息学



维数诅咒

- 更多的特征并不意味更好的分类结果。



- 过高的特征维度对分类器的设计学习带来困难：
  - 计算和存储复杂度增加，降低分类器效率；
  - 分类器过于复杂。

维数诅咒

- 分类器参数随着特征维数的增加而增多。
  - 线性分类器权值矢量维度随着特征维度线性增长；

$$g(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + w_0 = 0$$

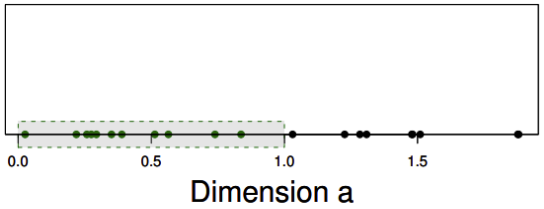
  - 更复杂分类器的参数随特征维度的增长速度更快。

- 分类器的学习是利用训练样本估计参数的过程。
  - 样本数越多，则参数估计越准确；
  - 样本数一定的条件下，参数数量越少则估计越准确；
  - 少量样本估计过多参数是不可靠的。

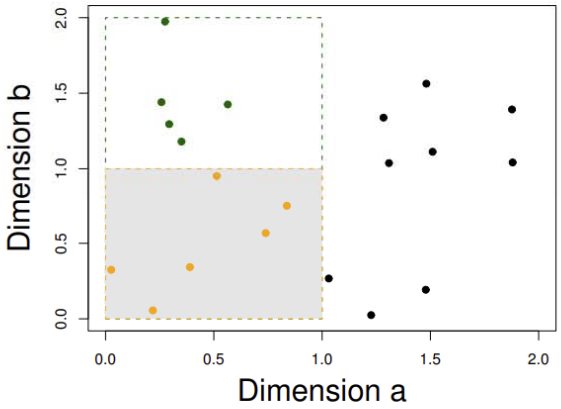
维数诅咒

- 维数诅咒（Curse of Dimensionality）：使用少量样本学习复杂分类器产生的种种问题。

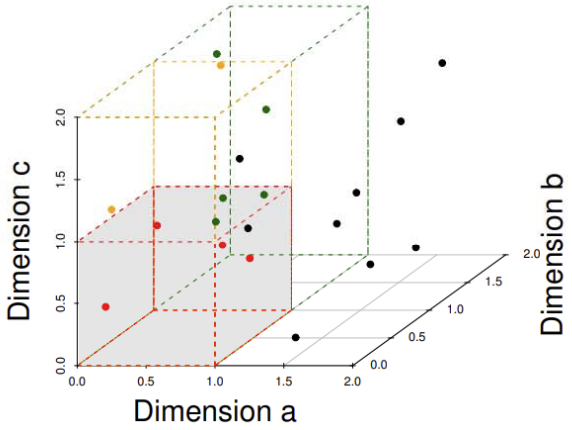
20 Objects/Samples in total



(a) 11 Objects in One Unit Bin



(b) 6 Objects in One Unit Bin

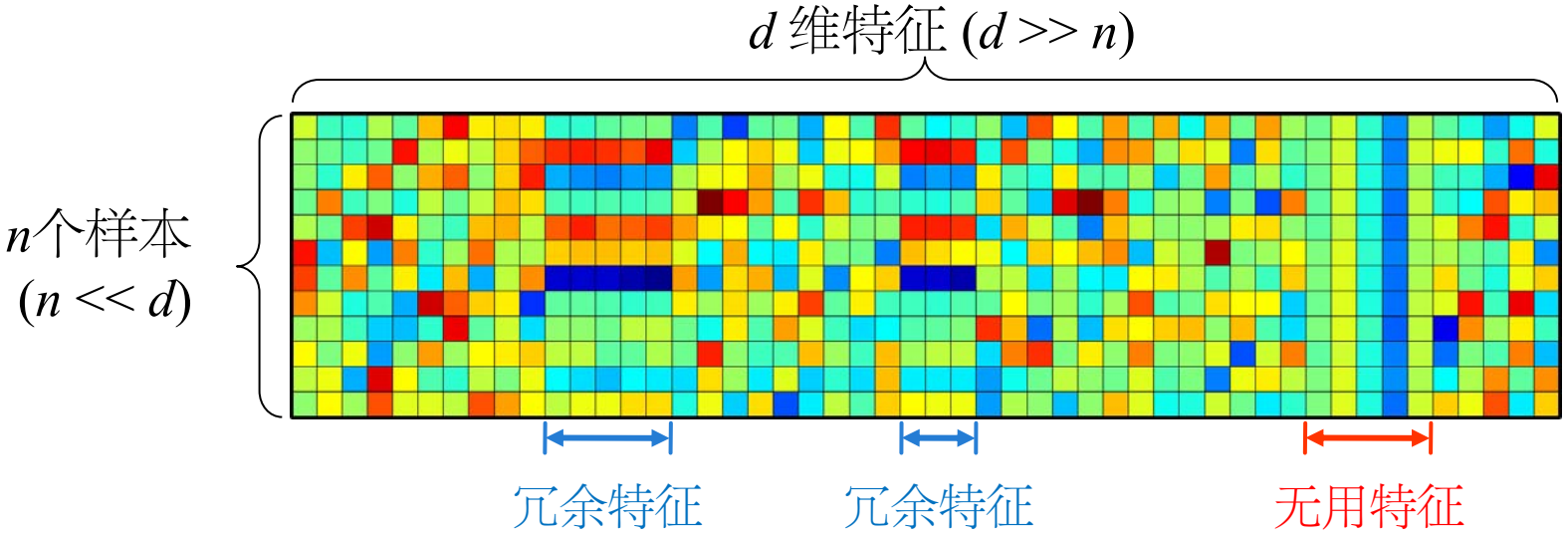


(c) 4 Objects in One Unit Bin

- 如果总样本数不变，随着特征维度增加，单位区域内的样本数急剧下降（稀疏化）。
- 随着维度的增加，理论上需要指数增长的样本数量覆盖到整个特征空间上时，才能保证有效估计参数。
- 稀疏化的样本难以正确估计模型参数，容易过拟合。

维数诅咒

- 高维特征面临的问题：
  - 很难找到真正的重要特征
  - 太多冗余特征
  - 无用“特征”导致过拟合



降维

- 降维（Dimension Reduction）：将数据/特征从高维空间转换到低维空间，但仍保留原数据/特征的有用信息。
- 降维的优势和潜在问题：
  - 降低计算的复杂度，降低存储器的占用，提高分类器的识别速度；
  - 降低分类器的复杂度，提高分类器的性能，提高泛化能力；
  - 有可能丢失可分性信息，降低分类准确率。
- 两类降维方法：特征选择和特征提取

特征选择

- 特征选择（Feature Selection）：从原始生成的 d 维特征 $\boldsymbol{x}=(x_1,x_2,\cdots,x_d)^T$ 中挑选 d' 个特征构成新的特征矢量 $\boldsymbol{x}'=(x_{i_1},x_{i_2},\cdots,x_{i_{d'}})^T$ 的过程，$d'<d$，$i_1,\cdots,i_{d'}\in\{1,\cdots,d\}$。
- 特征选择的目的是从原始特征中挑选出对分类最有价值的一组特征子集，抛弃掉与分类无关或贡献很小的特征；特征选择不生成新特征。
- 关键问题：如何快速准确找到最好的特征子集？

特征提取

- 特征提取（Feature Extraction）：将原始的 d 维特征 $\boldsymbol{x}=(x_1,x_2,\cdots,x_d)^T$ 经过某种变换 $y_i=f_i(\boldsymbol{x})$ 得到 d' 个特征 $\boldsymbol{y}=(y_1,y_2,\cdots,y_{d'})^T$，$d'<d$。
- 特征提取是根据原始特征的内在信息，通过某种函数变换重新生成一组新的特征；该组新特征可以极大限度地保留原有特征的信息量，并且去除了冗余信息。
- 关键问题：如何找到最合理有效的特征变换？

特征选择与特征提取

- 例：为区分桃子和橘子，计算图片RGB三个分量的颜色特征 (x_1, x_2, x_3)，以及高度宽度两个形状特征 (x_4, x_5)，生成5维特征矢量 $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5)^T$。
- 特征选择：区分桃子橘子的颜色分量主要是红色和绿色，因此可以从特征集合中去除蓝色。
- 特征提取：考虑到红绿分量相关性和高度宽度相关性，生成两个新特征：$y_1 = f_1(\boldsymbol{x}) = x_2 / x_1$ ，和 $y_2 = f_2(\boldsymbol{x}) = x_4 / x_5$ ，最终产生一个新的2维特征矢量 $\boldsymbol{y} = (y_1, y_2)^T$ 。

类别可分性判据

- 特征的价值体现在基于特征构建的分类器是否可有效区分不同类别的样本。
- 一般希望无需建立分类器（复杂且其性能受多因素影响）就可以在分类前依据样本集来度量特征对类别可分性的贡献。
- 因此，我们使用类别可分性判据作为度量特征分类价值的指标。

类别可分性判据

- 两种常用的类别可分性判据：
  - 基于距离的类别可分性判据
  - 基于散布矩阵的类别可分性判据

注，本节介绍的类别可分性判据与第4讲“聚类”介绍的“聚类准则”类似，区别在于：

- 聚类准则的样本是无监督的，目标是评价聚类后子集之间的区分程度；
- 类别可分性判据的样本有监督，评价的是样本集在不同特征子集上的区分程度。

类别可分性判据

- 如果同一类别的样本相似性大，不同类别的样本相似性小，那么分类越有利。因此，样本间距离可作为度量特征可分性的判据。
- 将 c 个类别的样本集表示为 $D_1,\cdots,D_c$，其中 $D_i=\{\boldsymbol{x}_1^{(i)},\cdots,\boldsymbol{x}_{n_i}^{(i)}\}$，样本上标表示所属类别，$n_i$ 为第 i 个类别的样本数，特征以集合的形式表示：$X=\{x_1,\cdots,x_d\}$。

注：以下距离判据均基于欧氏距离。

类别可分性判据

- 类内距离：度量特征集 X 上同类样本间相似程度

$$J_W(X)=\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n_i}\left\|\boldsymbol{x}_k^{(i)}-\boldsymbol{\mu}^{(i)}\right\|^2$$

其中 $\boldsymbol{\mu}^{(i)}=\frac{1}{n_i}\sum_{k=1}^{n_i}\boldsymbol{x}_k^{(i)}$ 是第 i 类的样本均值。

- 类间距离：度量不同类别样本之间的差异程度

$$J_B(X)=\sum_{i=1}^{c}\frac{n_i}{n}\left\|\boldsymbol{\mu}^{(i)}-\boldsymbol{\mu}\right\|^2$$

其中 $\boldsymbol{\mu}=\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n_i}\boldsymbol{x}_k^{(i)}$ 是所有类别样本均值。

类别可分性判据

- 类别可分性判据也可以用样本的散布矩阵计算。
- 第 i 类的类内散布矩阵：

$$S_W^{(i)} = \frac{1}{n_i}\sum_{k=1}^{n_i}(\boldsymbol{x}_k^{(i)} - \boldsymbol{\mu}^{(i)})(\boldsymbol{x}_k^{(i)} - \boldsymbol{\mu}^{(i)})^T$$

- 总的类内散布矩阵：

$$S_W = \sum_{i=1}^{c}\frac{n_i}{n}S_W^{(i)}$$

- 类间散布矩阵：

$$S_B = \sum_{i=1}^{c}\frac{n_i}{n}(\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu})^T$$

- 总体散布矩阵：

$$S_T = \frac{1}{n_i}\sum_{i=1}^{c}\sum_{k=1}^{n_i}(\boldsymbol{x}_k^{(i)} - \boldsymbol{\mu})(\boldsymbol{x}_k^{(i)} - \boldsymbol{\mu})^T = S_W + S_B$$

类别可分性判据

- 基于散布矩阵可以定义很多可分性判据：

$$J_1(X) = tr(S_W^{-1} S_B)$$

$$J_2(X) = \frac{tr(S_B)}{tr(S_W)}$$

$$J_3(X) = \frac{|S_B|}{|S_W|} = |S_W^{-1} S_B|$$

$$J_4(X) = \frac{|S_T|}{|S_W|}$$

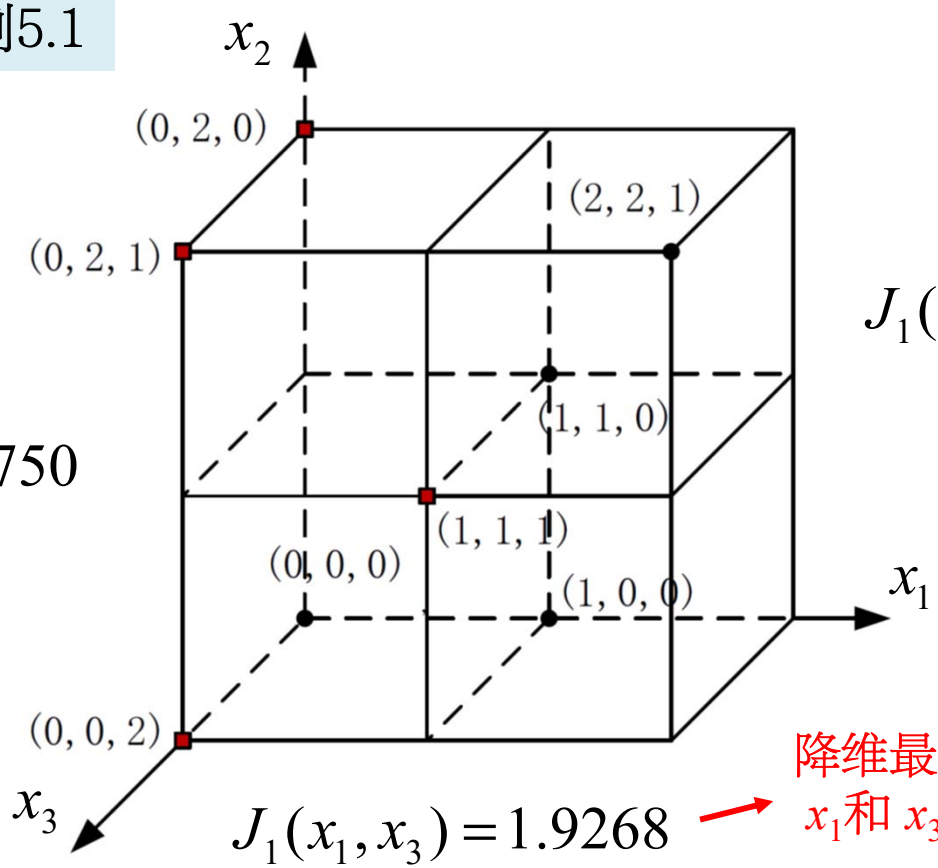
- 可分性判据越大越好（越可分）。
- 定义不同，原则类似：
  - ✓ 最小化类内距离
  - ✓ 最大化类间距离

其中，$tr(\cdot)$是矩阵的迹，$|\cdot|$是矩阵的行列式值。

类别可分性判据

- 例：下图中红点和黑点是两类样本，计算3维特征中任意2维的类别可分性判据 $J_1(X)=tr(S_W^{-1}S_B)$ 。

详见课本99页例5.1



$J_1(x_1,x_2)=0.8446$

$J_1(x_2,x_3)=0.3750$

$J_1(x_1,x_3)=1.9268$

降维最佳选择是投影到 x_1和 x_3 定义的平面

特征选择

- 特征选择：从原始特征集合 X 中挑选出一组最有利于分组的特征子集 X' 。

- 因为类别可分性可以评价特征对分类问题的有效性，所以，特征识别是对某种类别可分性依据的优化

$$X' = \arg\max_{\tilde{X} \subset X} J(\tilde{X})$$

其中原始特征集合 X 中包含 d 个特征，X' 中包含 $d' < d$ 个特征，$\tilde{X}$ 是任意包含 d' 个元素的 X 子集。

特征选择

- 特征选择的三个思路：过滤法、包装法、嵌入法
- 过滤法（Filter）：按照特征的可分性判据和与标签的相关性等指标挑选部分特征。
- 包装法（Wrapper）：将可分析判据或分类性能作为子集评价标准，比较多种可能子集的性能。
- 嵌入法（Embedded）：将特征选择与分类器训练融为一体，分类器训练过程中自动选择特征。

过滤法

- 过滤法：用可分性判据（或其他指标如特征和标签的相关性等）分别评价每个特征，根据判据值大小对特征排序

$$J(x_1) \geq J(x_2) \geq \cdots J(x_{d'}) \geq \cdots \geq J(x_d)$$

选择判据值最大的前 d' 个特征：$X' = \{x_1, x_2, \cdots x_{d'}\}$。

- 但是，只有特征相互独立时，本思路才可以保证解的最优性。如果特征间存在相关性，判据值最大的 d' 个特征组合在一起不保证最优可分性。

包装法

- 包装法：尝试多种 $\tilde{X} \in X$ 的特征组合，计算并比较每一种组合的判据值，选出最优组合。
- 包装法包括穷举法、分支定界法、次优搜索算法（顺序前进法、顺序后退法等）。
- 穷举法（尝试所有特征组合）可以选出最优的特征组合/子集，但运算量巨大，很难操作。

从 d 个特征中选出 d' 个特征共有 $C_d^{d'}$ 种组合：

- 从3个特征中选2个特征有3种可能的组合；
- 从100个特征中选10个特征有大于17万亿的组合。

分支定界法

- 分支定界法：可以减小穷举法计算复杂度的最优特征组合搜索算法，基于可分性判据的单调性：

$$X_1 \subset X_2 \Rightarrow J(X_1) \leq J(X_2)$$

即，如果从某特征集合中移除一个特征会减小判据值，如果增加一个特征会增大判据值。

- 注意：并非所有类别可分性判据具有单调性。
  - 满足单调性的判据：J_W，J_B，J_1，J_3
  - 不满足单调性的判据：J_2，J_4

见课件19页

- 分支定界法只有基于单调性判据时才可以保证搜索到最优特征组合。

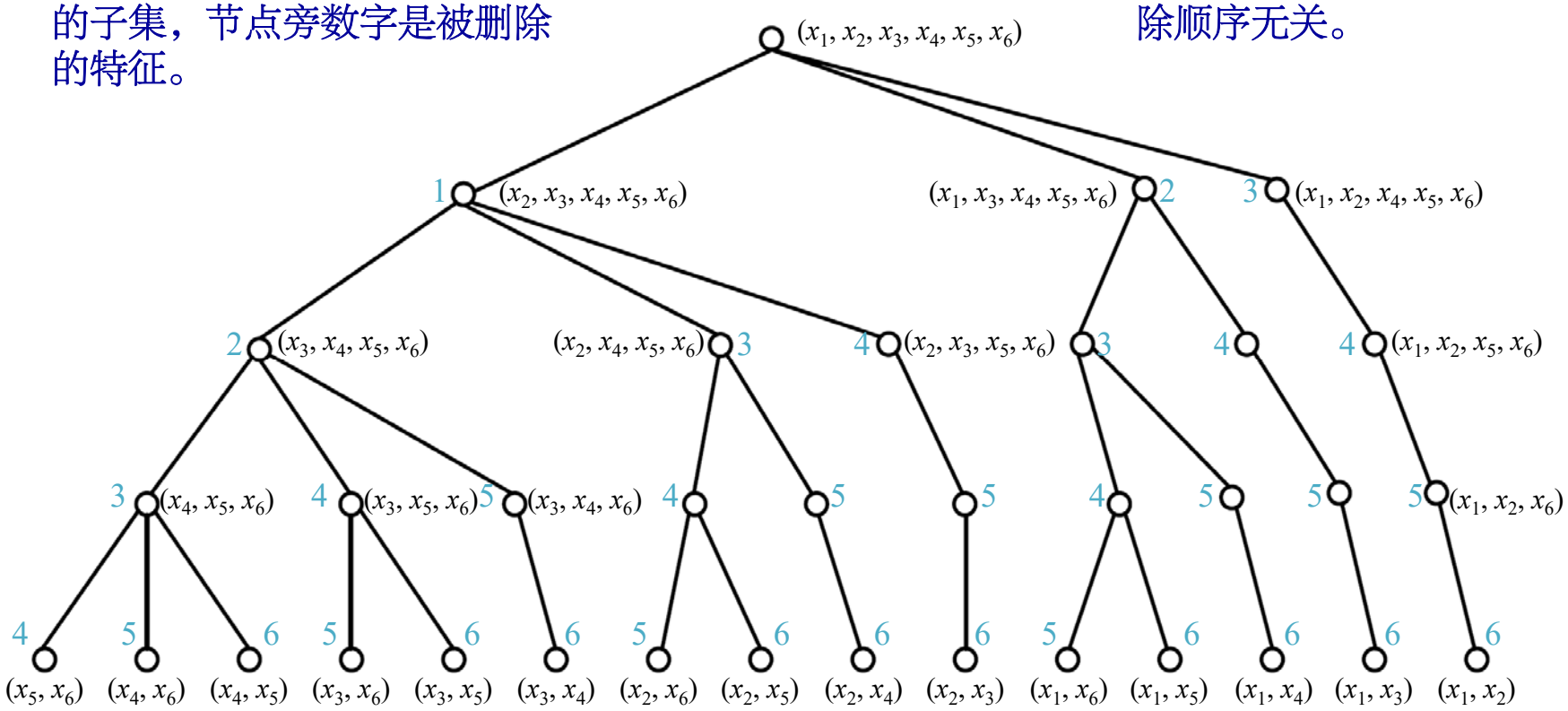
分支定界法

- 例：从原始6维特征中选出2个特征，构建搜索树

- 每个节点是一个特征组合。
- 每个节点的子节点对应（上方）父节点特征集中删除一个特征的子集，节点旁数字是被删除的特征。

- 最上方的根节点对应全部特征。

- 搜索树非对称，保证不会出现重复组合特征。
- 仅关注删除的特征，与删除顺序无关。



- 最下方的叶节点对应所有15种特征组合。

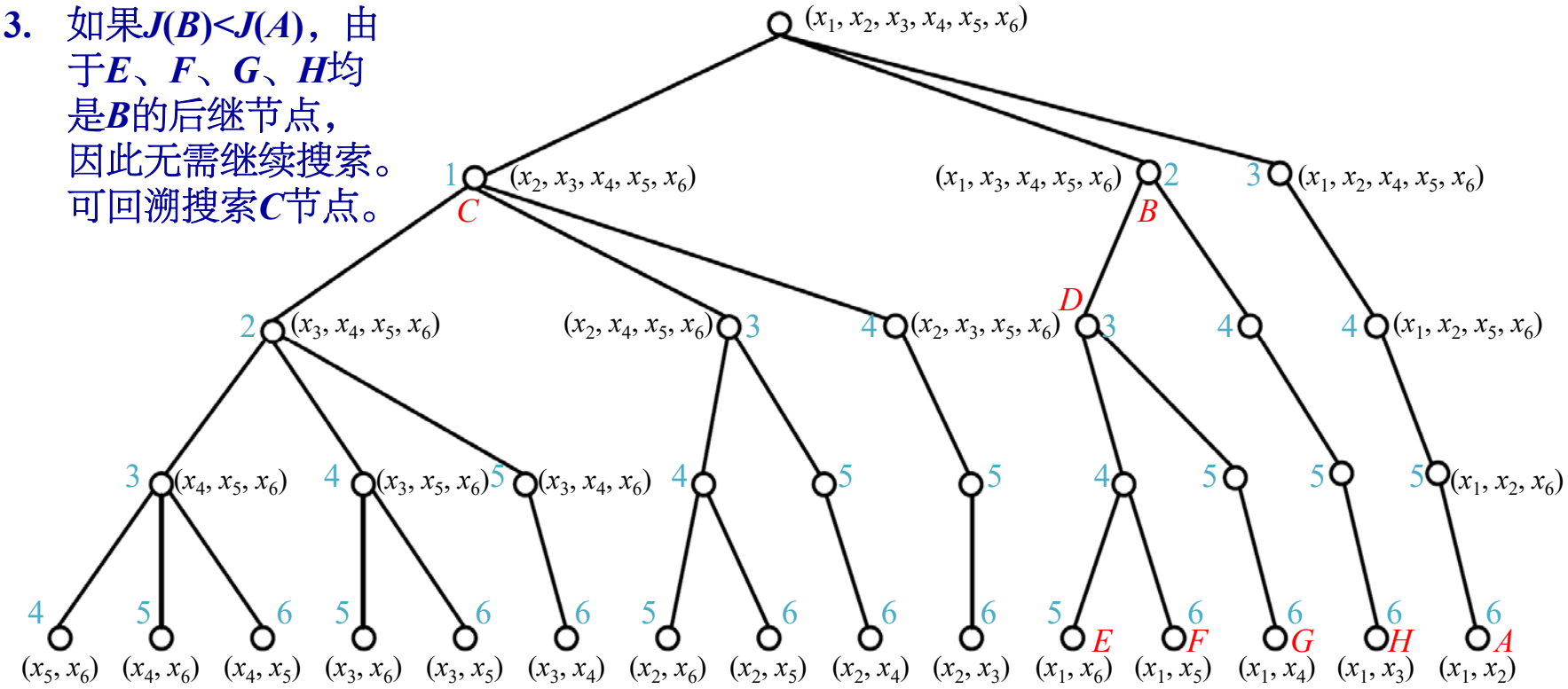
分支定界法

- 初始化：根据原始特征维数 d 和选择特征维数 d' 构建搜索树，设置界值 $B = 0$；
- 从右向左分支定界搜索：
  - 如果当前节点没有分支，则向下搜索，直到叶节点为止。计算叶节点代表的特征集合的可分性判据，如果大于界值 B，则将 B 替换为这个判据值，并记录这个特征集合，作为当前的最优选择；向上回溯，直到有节点存在未搜索过的分支为止，按照从右向左的顺序搜索其子节点；
  - 如果当前节点有分支，则计算当前节点代表特征集合的可分性判据，小于界值 B，则中止该节点向下的搜索；否则按照从右向左的顺序搜索其子节点。
- 输出：最优特征集合。

分支定界法

- 例：分支定界法按照从右向左深度优先进行搜索

1. 计算节点*A*特征组合的判据*J*(*A*)作为当前最优结果。
2. 回溯搜索到节点*B*，计算*J*(*B*)。
3. 如果*J*(*B*)<*J*(*A*)，由于*E*、*F*、*G*、*H*均是*B*的后继节点，因此无需继续搜索。可回溯搜索*C*节点。
4. 如果*J*(*B*)>*J*(*A*)，需要继续搜索节点*D*和*H*。
5. 从右向左重复搜索。



分支定界法的问题

- 分支定界法能否搜索到最优的特征组合依赖于所采用的类别可分性判据是否具有单调性。
- 分支定界法的计算复杂度是不确定的，与最优解所在位置有关。
  - 如果最优解分支在右端并且根节点的子节点判据值均小于最优解，则搜索效率高；
  - 如果每个分支的可分性判据都大于其左端分支的可分性判据，实际的计算复杂度会超过穷举法。

次优搜索算法

- 顺序前进法：从空集开始，每次向选择的特征集中加入一个特征，直到特征集中包含d'个特征为止，每次选择加入特征的原则是将其加入后使得可分性判据最大。
- 顺序前进法每一轮迭代只需要计算每一个未被选择的特征加入 X' 之后的判据值，因此选择出 d' 个特征需要计算判据值的次数为

$$\sum_{i=0}^{d'-1}(d-i)=\frac{d'(2d-d'+1)}{2}$$

次优搜索算法

- 顺序前进法

- 初始化：原始特征集合 X，设置选择特征集合$X' = \varnothing$；
- 循环直到 X' 中包含d'个特征为止：
  - 计算将任意未被选择的特征加入 X'后的可分性判据值：

$$J(X' \cup \{x_i\}),\ \forall x_i \in X - X'$$

  - 寻找最优特征：$x' = \arg\max_{x_i \in X - X'} J(X' \cup \{x_i\})$
  - 将最优特征加入选择特征集合：$X' = X' \cup \{x'\}$
- 输出：特征集合 X'。

次优搜索算法

- 顺序后退法：从整个特征集开始，每一轮从特征集中选择一个最差的特征删除，选择删除特征的原则是将其删除之后使得特征集的可分性判据值下降得最少。
- 顺序后退法每一轮迭代需要计算将 X' 中每个元素删除之后的判据值，直到 X' 中剩余 d' 个元素为止，需要迭代 $d-d'$ 次，需要计算判据值次数为

$$\sum_{i=0}^{d-d'-1}(d-i)=\frac{(d-d')(d+d'+1)}{2}$$

次优搜索算法

- 顺序后退法

  - 初始化：原始特征集合 X，设置选择特征集合 $X' = X$；
  - 循环直到 X' 中包含 d' 个特征为止：
    - 计算将任意一个 X' 中元素删除之后的可分性判据值：

$$J(X' - \{x_i\}),\ \forall x_i \in X'$$

    - 寻找最优的删除特征：$x' = \arg\max_{x_i \in X'} J(X' - \{x_i\})$
    - 将选择的特征移出特征集合：$X' = X' - \{x'\}$
  - 输出：特征集合 X'。

次优搜索算法

- 广义顺序前进/后退法：与顺序前进/后退法基本类似，但每次增加或删除 r 个特征。
- 如果共进行 k 轮迭代，判据值的计算次数为

$$\sum_{i=0}^{k-1} C_{d-ir}^{r} = \frac{1}{r!}\sum_{i=0}^{k-1}\frac{(d-ir)!}{(d-ir-r)!}$$

- 一般地，广义顺序前进/后退法的计算量大于顺序前进/后退法，但因为考虑了特征间的相关性，优化结果一般好于顺序前进/后退法。

次优搜索算法

- 增 l 减 r 法（l-r 法）：允许对特征选择过程进行回溯，先采用顺序前进法加入 l 个特征，再采用顺序后退法删除 r 个特征 $(l > r)$，循环直到 X' 中包含 d' 个特征为止。以上过程也可以相反执行：先删除 r 个特征，再加入 l 个特征$(l < r)$ 。
- 增 l 减 r 法解决了顺序前进/后退法无法回溯的问题：特征被加入/删除后无法被重新考虑是不合适的，因为选择时只考虑它与当前 X' 中特征的相关性，未考虑之后加入或删除某些特征的影响。

次优搜索算法

- 增 l 减 r 法（l-r 法）
  - 初始化：设置选择特征集合 $X' = \varnothing$；
  - 循环直到 X' 中包含 d' 个特征为止：
    - 调用顺序前进法 l 次，向 X' 中添加 l 个特征；
    - 调用顺序后退法 r 次，向 X' 中删除 r 个特征；
  - 输出：特征集合 X' 。

或

  - 初始化：设置选择特征集合 $X' = X$ ；
  - 循环直到 X' 中包含 d' 个特征为止：
    - 调用顺序后退法 r 次，向 X' 中删除 r 个特征；
    - 调用顺序前进法 l 次，向 X' 中添加 l 个特征；
  - 输出：特征集合 X' 。

递归式特征消除

- 递归式特征消除（Recursive Feature Elimination）： 基于一个为特征分配权重的外部估计量（例如线性回归模型里的回归系数），通过递归地考虑越来越小的特征集来选择特征。
  - 首先，对初始特征集训练估计器，通过特征权重等属性获得每个特征的重要性。
  - 然后从当前的特征集中删除最不重要的特征。
  - 在经过修剪的集合上递归地重复这个过程，直到最终达到所需的特征数量。

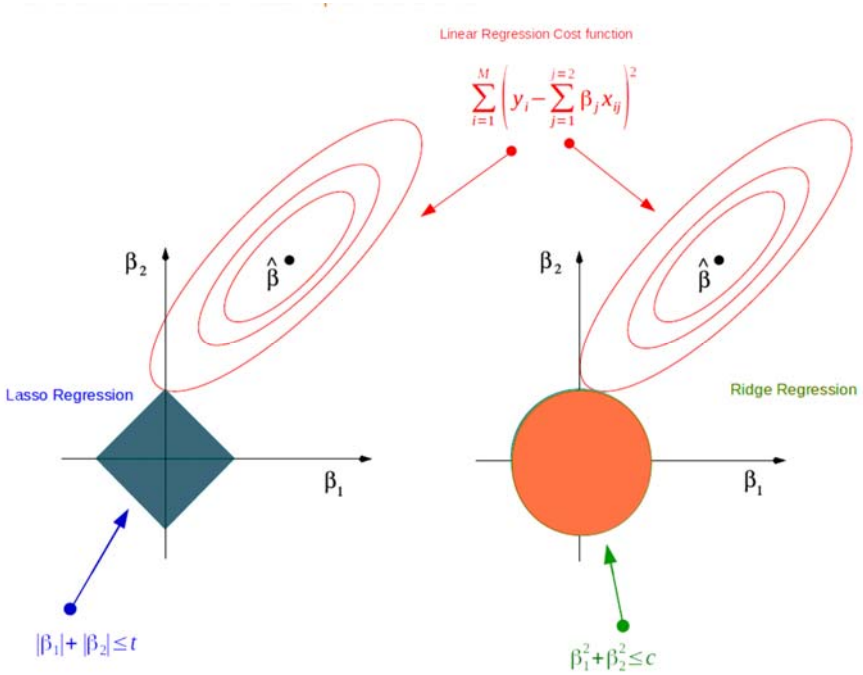
嵌入法

- 嵌入法：将所有特征包含在分类器中，特征选择与模型训练在同一个优化过程中完成。
- 一般可以通过正则化（regularization）算法，例如LASSO，自动实现特征选择。
- 正则化：对模型参数施加约束项（通常对参数添加稀疏性范数如 $\mathcal{L}_1$ 范数），使其中许多参数自动变为零（即对应特征对模型无预测能力），从而实现特征选择。

嵌入法

- LASSO：线性回归模型的优化目标设置为平方误差与特征的 $\mathcal{L}_1$ 范数（正则化）之和。

- $\mathcal{L}_1$ 范数有助于降低过拟合风险，而且可以获得稀疏（sparse）解，即求得的回归系数中只有少量非零值（d 个特征中仅有对应非零回归系数的特征才会出现在最终模型中）。



延伸阅读：https://www.cnblogs.com/wuliytTaotao/p/10837533.html

特征提取

- 降维的主要方法：特征提取和特征选择。
- 特征选择：从原始的 d 维特征 $\boldsymbol{x}=(x_1,x_2,\cdots,x_d)^T$ 中挑选 d' 个特征使得某种类别可分性判据最优。
- 特征提取：将原始特征经过某种变换 $y_i=f_i(\boldsymbol{x})$ 得到可极大限度保留原特征信息的少量新特征。
- 接下来介绍当 $f_i(\boldsymbol{x})$ 是线性函数时的两类经典线性特征提取方法：
  - 主成分分析
  - 基于Fisher准则的可分性分析

特征提取

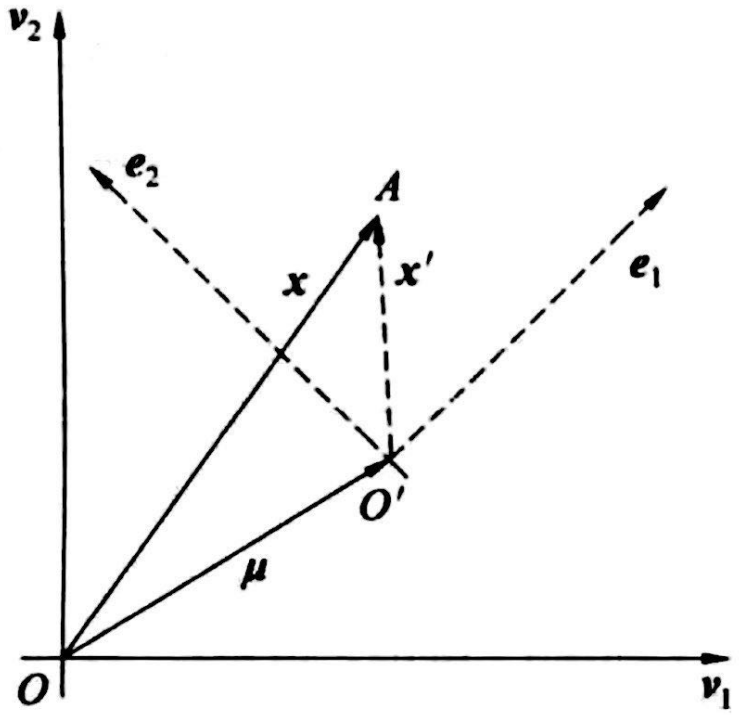
- 思考：区分鲈鱼和鲑鱼时有以下特征，重量、长度、宽度、周长、头部大小、鱼鳍大小等。这些特征是否相关？是否冗余？如何综合这些特征？
- 思考：人脸识别中，同一人的不同面部照片有多少重复信息和新颖信息？不同人的面部照片有什么相同信息和差别信息？

主成分分析

- 主成分分析（Principal Component Analysis，PCA）是一种最常用的线性成分分析方法。
- PCA的主要思想是寻找到数据的主轴方向，由主轴构成一个新的坐标系（维度比原维数低），然后数据由原坐标系向新的坐标系投影。
- PCA从尽量减少信息损失的角度实现降维。
- PCA的其它名称：离散K-L变换，Hotelling变换

主成分分析推导

- 样本集合中的每一个样本对应特征空间中的一个点，同一个点在不同坐标下对应不同矢量。



- 左图中，点 A
  - 在以 O 为原点$\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$为基矢量的坐标系下对应的矢量为 $\boldsymbol{x}$，
  - 在以 O' 为原点$\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$为基矢量的坐标系下对应的矢量为 $\boldsymbol{x}'$。

- 如果新坐标系原点O' 在原坐标系下的矢量是 $\boldsymbol{\mu}$，则有

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{x}'$$

主成分分析推导

- 分别将矢量 $\boldsymbol{x}$ 和 $\boldsymbol{x}'$ 写成两个坐标系下的坐标形式 $\boldsymbol{x}=(x_1,\cdots,x_d)^T$，$\boldsymbol{x}'=(a_1,\cdots,a_d)^T$，则有

直角坐标系下，矢量可以表示为基矢量的线性组合。见附录A.3。

$$\boldsymbol{x}=\boldsymbol{\mu}+\boldsymbol{x}'=\boldsymbol{\mu}+\sum_{i=1}^{d}a_i\boldsymbol{e}_i$$

- 新坐标系下，矢量 $\boldsymbol{x}'$ 的元素可以由原坐标系下的矢量 $\boldsymbol{x}$ 以及 $\boldsymbol{\mu}$ 和基矢量（正交）计算得到：

$$a_i=\boldsymbol{e}_i^T(\boldsymbol{x}-\boldsymbol{\mu}),\quad i=1,\cdots,d$$

主成分分析推导

- $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{x}' = \boldsymbol{\mu} + \sum_{i=1}^{d} a_i \boldsymbol{e}_i$ 意味着由新坐标下的矢量 $\boldsymbol{x}'$ 恢复/重建原矢量 $\boldsymbol{x}$ 是没有误差的。
- 可以仅选择保留新坐标系下的 $d' < d$ 个特征，然后用保留的 d' 个特征恢复原坐标系下的 d 维特征矢量：

$$\hat{\boldsymbol{x}} = \boldsymbol{\mu} + \sum_{i=1}^{d'} a_i \boldsymbol{e}_i$$

- 但是，显然 $\hat{\boldsymbol{x}}$ 是对 $\boldsymbol{x}$ 的近似，用 $\hat{\boldsymbol{x}}$ 代替 $\boldsymbol{x}$ 的时候会出现误差。误差大小和新坐标的位置、基矢量方向和保留的特征有关。

主成分分析推导

- 主成分分析中，新坐标系原点选在样本集的均值矢量 $\boldsymbol{\mu}$ 上，然后寻找一组最优基矢量 $\{\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d\}$，使得在保留前 d' 个特征的条件下，恢复样本集的均方误差最小，即求解优化问题

$$\min_{\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d} J(\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d)=\frac{1}{n}\sum_{k=1}^{n}\| \boldsymbol{x}_k-\hat{\boldsymbol{x}}_k \|^2$$

其中 $\hat{\boldsymbol{x}}_k$ 是将第 k 个的样本 $\boldsymbol{x}_k$ 由原坐标系变换到新坐标系下后只使用前 d' 个特征恢复得到的矢量。

主成分分析推导

- 如用 a_{ki} 表示第 k 个样本在新坐标系下的第 i 维的特征，可得

$$\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k = \sum_{i=d'+1}^{d} a_{ki}\boldsymbol{e}_i$$

- 优化问题变为：

$$\min_{\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d} J(\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d) = \frac{1}{n}\sum_{k=1}^{n} \| \sum_{i=d'+1}^{d} a_{ki}\boldsymbol{e}_i \|^2$$

$$= \frac{1}{n}\sum_{k=1}^{n} (\sum_{i=d'+1}^{d} a_{ki}\boldsymbol{e}_i)^T (\sum_{i=d'+1}^{d} a_{ki}\boldsymbol{e}_i)$$

$$= \frac{1}{n}\sum_{k=1}^{n}\sum_{i=d'+1}^{d} a_{ki}^2$$

利用基矢量间的正交性

主成分分析推导

- 因为标量 $a_i = \boldsymbol{e}_i^T(\boldsymbol{x}-\boldsymbol{\mu}) = [\boldsymbol{e}_i^T(\boldsymbol{x}-\boldsymbol{\mu})]^T$ ，优化函数变为：

$$\min_{\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d} J(\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d) = \frac{1}{n}\sum_{k=1}^{n}\sum_{i=d'+1}^{d}[\boldsymbol{e}_i^T(\boldsymbol{x}_k-\boldsymbol{\mu})][\boldsymbol{e}_i^T(\boldsymbol{x}_k-\boldsymbol{\mu})]^T$$

$$= \sum_{i=d'+1}^{d}\boldsymbol{e}_i^T\left[\frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{x}_k-\boldsymbol{\mu})(\boldsymbol{x}_k-\boldsymbol{\mu})^T\right]\boldsymbol{e}_i$$

$$= \sum_{i=d'+1}^{d}\boldsymbol{e}_i^T\boldsymbol{\Sigma}\boldsymbol{e}_i$$

其中 $\boldsymbol{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{x}_k-\boldsymbol{\mu})(\boldsymbol{x}_k-\boldsymbol{\mu})^T$ 是样本集协方差矩阵的估计。

主成分分析推导

- 当 $\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d=0$ 时，$\min\limits_{\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d} J(\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d)=\sum\limits_{i=d'+1}^{d}\boldsymbol{e}_i^T\boldsymbol{\Sigma}\boldsymbol{e}_i$ 有最小值，但零矢量无意义；需要约束 $\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d$ 的长度。
- 主成分分析优化下列约束问题：

$$\min_{\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d} J(\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d)=\sum_{i=d'+1}^{d}\boldsymbol{e}_i^T\boldsymbol{\Sigma}\boldsymbol{e}_i$$

$$\text{约束：}\ \|\boldsymbol{e}_i\|^2=1,\ i=1,\cdots,d$$

主成分分析推导

- 有约束的优化问题可以通过拉格朗日法（见附录B.6）求解下列包含拉格朗日系数 λ_i 的函数

$$L(\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d,\lambda_1,\cdots,\lambda_d)=\sum_{i=d'+1}^{d}(\boldsymbol{e}_i^T\boldsymbol{\Sigma}\boldsymbol{e}_i-\lambda_i(\boldsymbol{e}_i^T\boldsymbol{e}_i-1))$$

- 对每一个基函数 $\boldsymbol{e}_i$ 求导可得：

$$\frac{\partial L(\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d,\lambda_1,\cdots,\lambda_d)}{\partial \boldsymbol{e}_i}=2\boldsymbol{\Sigma}\boldsymbol{e}_i-2\lambda_i\boldsymbol{e}_i=0$$

- 因此优化问题的解需要满足：$\boldsymbol{\Sigma}\boldsymbol{e}_i=\lambda_i\boldsymbol{e}_i$

主成分分析推导

- 使得 $\boldsymbol{\Sigma e}_i = \lambda_i \boldsymbol{e}_i$ 成立的 λ_i 和 $\boldsymbol{e}_i$ 分别为 $\boldsymbol{\Sigma}$ 的特征值和对应的特征向量。
- 因此，如果希望将一个样本集 D 的维度在新坐标下降低，可以将新坐标系原点放在样本集 D 均值位置，以集合 D 的协方差矩阵的特征矢量作为基矢量，可以保证只用保留的 d' 维特征恢复原矢量的时候均方误差最小。
- 保留哪些特征使得均方误差最小？

主成分分析推导

- 因为 $J(\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d)=\sum_{i=d'+1}^{d}\boldsymbol{e}_i^T\boldsymbol{\Sigma}\boldsymbol{e}_i=\sum_{i=d'+1}^{d}\lambda_i\boldsymbol{e}_i^T\boldsymbol{e}_i=\sum_{i=d'+1}^{d}\lambda_i$ ，所以要最小化 $J(\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d)$ ，只需要选择最小的 $d-d'$ 个特征值。
- 因此，在新坐标系下保留下来的是 $\boldsymbol{\Sigma}$ 最大的前 d' 个特征值，并将其对应的特征矢量作为新坐标的基矢量。

主成分分析算法

- 输入样本集合 D，估计均值矢量 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$；
- 计算协方差矩阵 $\boldsymbol{\Sigma}$ 的特征值和特征矢量，按照特征值由大到小排序；
- 选择前 d' 个特征矢量作为列矢量构成矩阵：

$$\boldsymbol{E} = (\boldsymbol{e}_1 \ \ \boldsymbol{e}_2 \ \ \cdots \ \ \boldsymbol{e}_{d'})；$$

- d 维特征矢量 $\boldsymbol{x}$ 可以转换为 d' 维特征矢量 $\boldsymbol{x}'$：

$$\boldsymbol{x}' = \boldsymbol{E}^T(\boldsymbol{x} - \boldsymbol{\mu})。$$

- 【常用】由降维后的矢量 $\boldsymbol{x}'$ 恢复原矢量 $\boldsymbol{x}$：

$$\hat{\boldsymbol{x}} = \boldsymbol{E}\boldsymbol{x}' + \boldsymbol{\mu}$$

主成分分析举例

- 例：用PCA将下列样本的特征从2维降为1维：

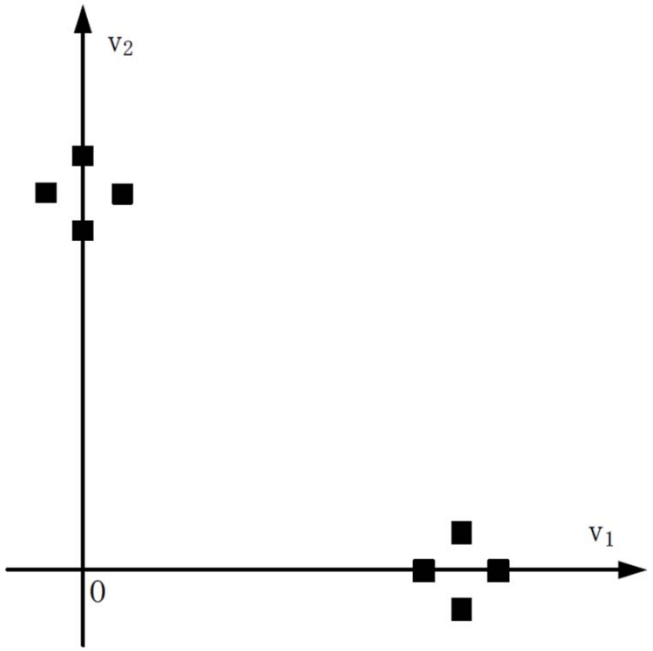
$\boldsymbol{x}_1=[10, 1]^T, \boldsymbol{x}_2=[9, 0]^T, \boldsymbol{x}_3=[10, -1]^T, \boldsymbol{x}_4=[11, 0]^T,$

$\boldsymbol{x}_5=[0, 9]^T, \boldsymbol{x}_6=[1, 10]^T, \boldsymbol{x}_7=[0, 11]^T, \boldsymbol{x}_8=[-1, 10]^T$

均值 $\boldsymbol{\mu}=[5, 5]^T$

协方差矩阵

$$\boldsymbol{\Sigma}=\begin{bmatrix} 25.5 & -25 \\ -25 & 25.5 \end{bmatrix}$$



主成分分析举例

- 例：用PCA将下列样本的特征从2维降为1维：

$\boldsymbol{x}_1=[10, 1]^T, \boldsymbol{x}_2=[9, 0]^T, \boldsymbol{x}_3=[10, -1]^T, \boldsymbol{x}_4=[11, 0]^T,$

$\boldsymbol{x}_5=[0, 9]^T, \boldsymbol{x}_6=[1, 10]^T, \boldsymbol{x}_7=[0, 11]^T, \boldsymbol{x}_8=[-1, 10]^T$

求解特征值和特征向量

$$|\boldsymbol{\Sigma}-\lambda\boldsymbol{I}|=\begin{vmatrix}25.5-\lambda & -25\\ -25 & 25.5-\lambda\end{vmatrix}=0$$

$$\Updownarrow$$

$$\lambda_1=50.5, \lambda_2=0.5; \boldsymbol{e}_1=[\sqrt{2}/2,-\sqrt{2}/2], \boldsymbol{e}_2=[\sqrt{2}/2,\sqrt{2}/2]$$

主分量

主成分分析举例

- 例：用PCA将下列样本的特征从2维降为1维：

$\boldsymbol{x}_1=[10, 1]^T, \boldsymbol{x}_2=[9, 0]^T, \boldsymbol{x}_3=[10, -1]^T, \boldsymbol{x}_4=[11, 0]^T,$

$\boldsymbol{x}_5=[0, 9]^T, \boldsymbol{x}_6=[1, 10]^T, \boldsymbol{x}_7=[0, 11]^T, \boldsymbol{x}_8=[-1, 10]^T$

使用主分量恢复原矢量

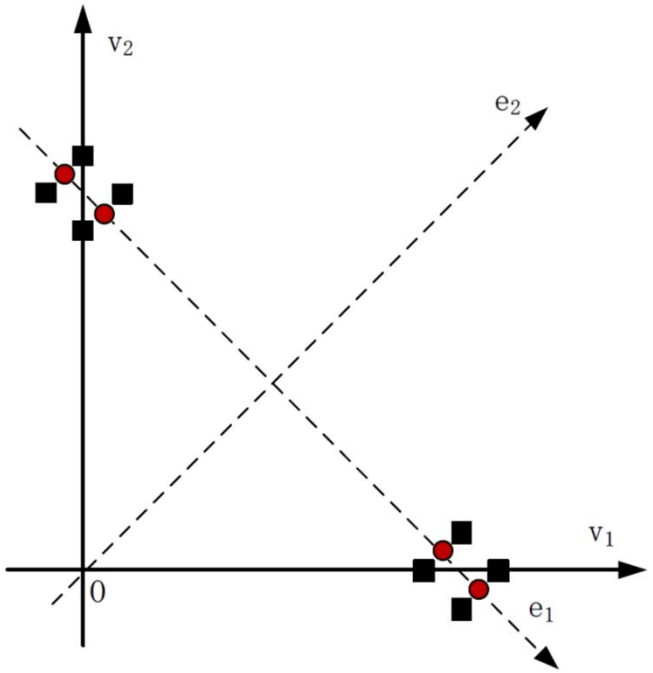
$$\boldsymbol{x}_i' = \boldsymbol{e}_1^T(\boldsymbol{x}_i - \boldsymbol{\mu})$$

$$\boldsymbol{x}_1' = 9\sqrt{2}/2, \boldsymbol{x}_2' = 9\sqrt{2}/2,$$

$$\boldsymbol{x}_3' = 11\sqrt{2}/2, \boldsymbol{x}_4' = 11\sqrt{2}/2,$$

$$\boldsymbol{x}_5' = -9\sqrt{2}/2, \boldsymbol{x}_6' = -9\sqrt{2}/2,$$

$$\boldsymbol{x}_7' = -11\sqrt{2}/2, \boldsymbol{x}_8' = -11\sqrt{2}/2$$



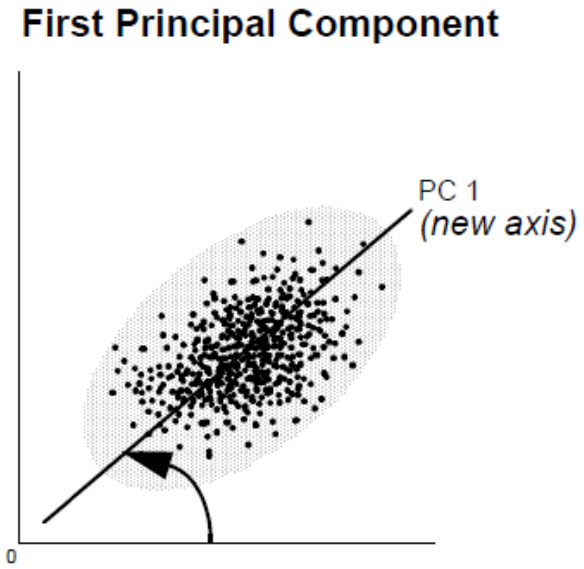
主成分分析讨论

- 特征矢量正交：协方差矩阵 Σ 是实对称阵，其特征值为实数，其特征矢量正交。即，特征矢量构成一组正交基，主成分分析得到的新坐标系是一个直角坐标系。

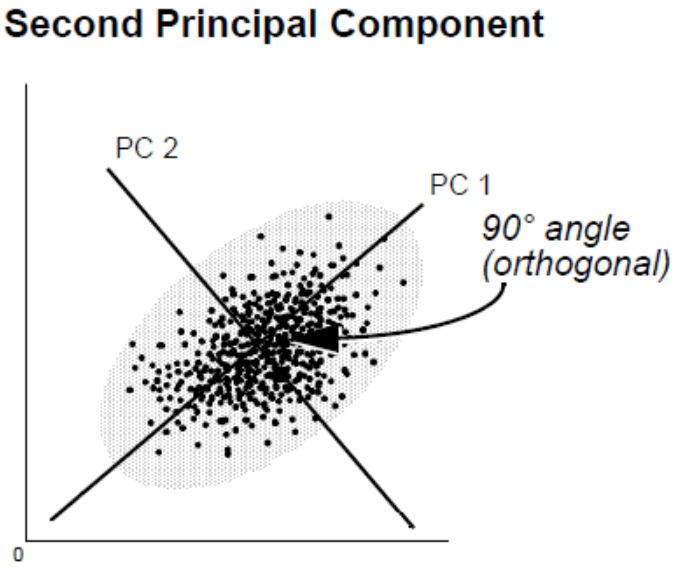
- 变换后特征不相关：将数据向新的坐标轴/正交基矢量投影之后，特征间是不相关的（证明见课本109页）。即，主成分分析可以有效消除特征间相关性。

主成分分析讨论

- 特征值即信息量：特征值的模表示数据在相应正交基上的投影长度。 特征值越大，说明矩阵在对应特征向量上的方差越大，样本点越分散，越容易区分，信息量也就越多。



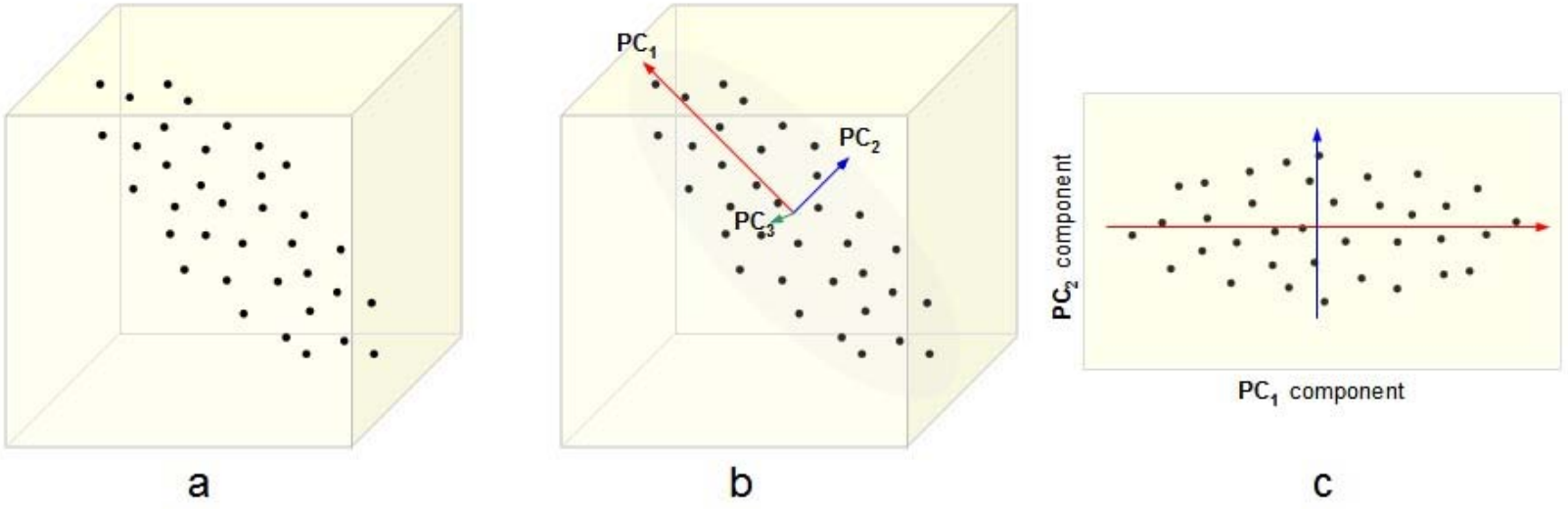
第一主成分，数据方差最大的投影方向



第二主成分和第一主成分正交，数据方差次大的投影方向

主成分分析讨论

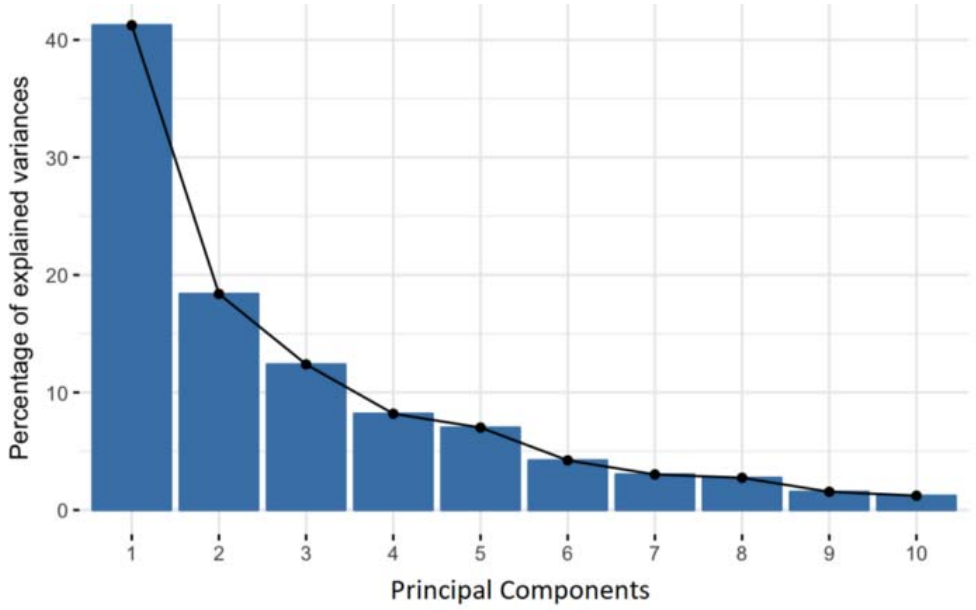
- 冗余特征： 特征值描述了变换后各维的重要性，特征值为0（或接近0）的特征为冗余特征。



a 给定的数据集为三维变量
b 数据的三个正交主成分，按照方差大小排序
c 数据集被投影到前两个主成分中，丢弃第三个成分

主成分分析讨论

- 主成分分析时应该保留多少特征?
- 主成分分析用 d' 个新特征表示原始的 d 维特征引起的误差是 $J(\boldsymbol{e}_1,\cdots,\boldsymbol{e}_d)=\sum_{i=d'+1}^{d}\lambda_i$，舍弃的特征值越小则误差越小。



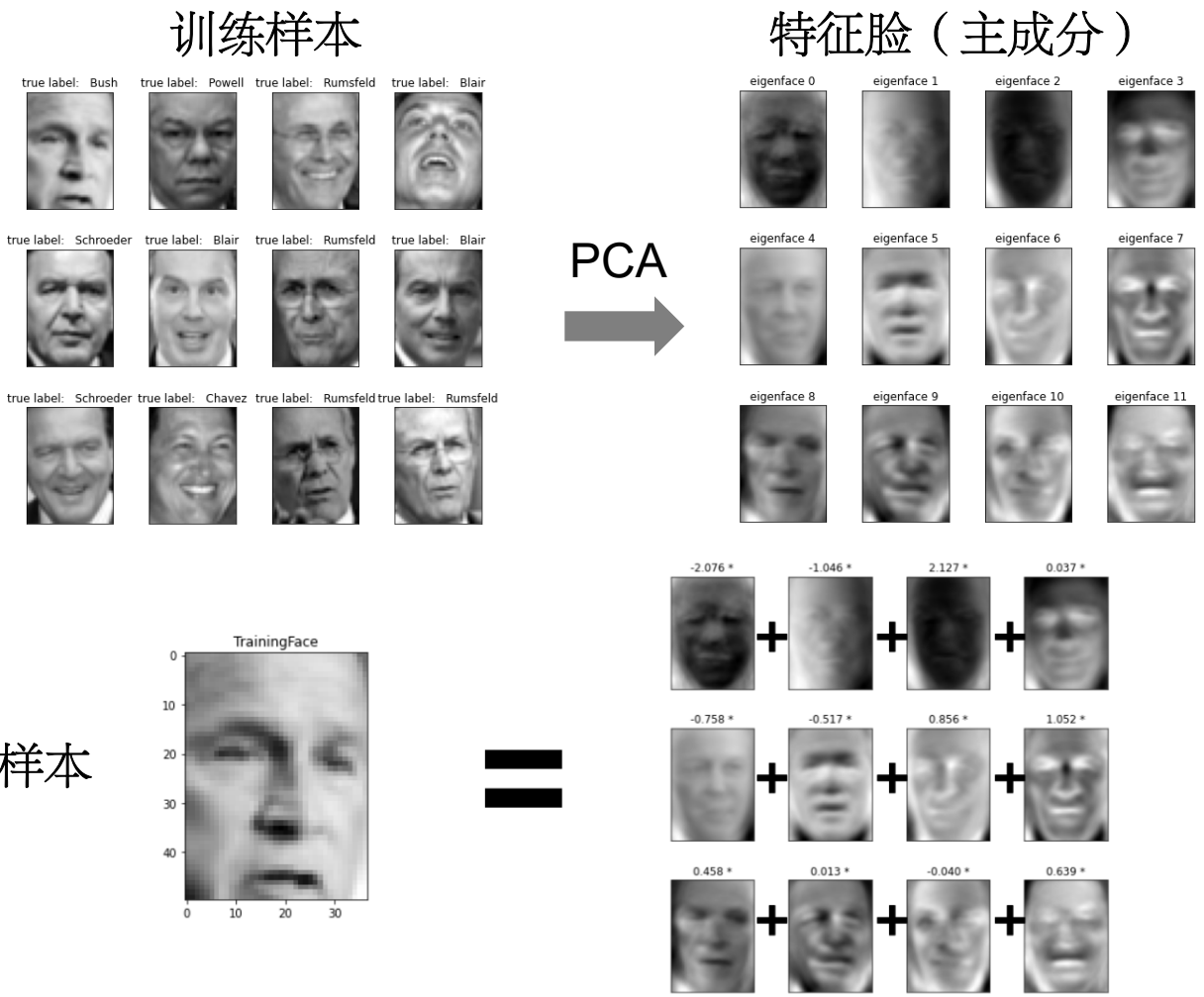
主成分分析讨论

- 常见选择 d' 的方法是在从大到小排序特征值之后计算累加值，以累加值与所有特征值总和的比例超过95%（或其他比例）为原则选择 d'。这意味着希望降维的误差不超过5%。

$$d' = \arg\min_{1 \le k \le d} \left[\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \ge 95\% \right]$$

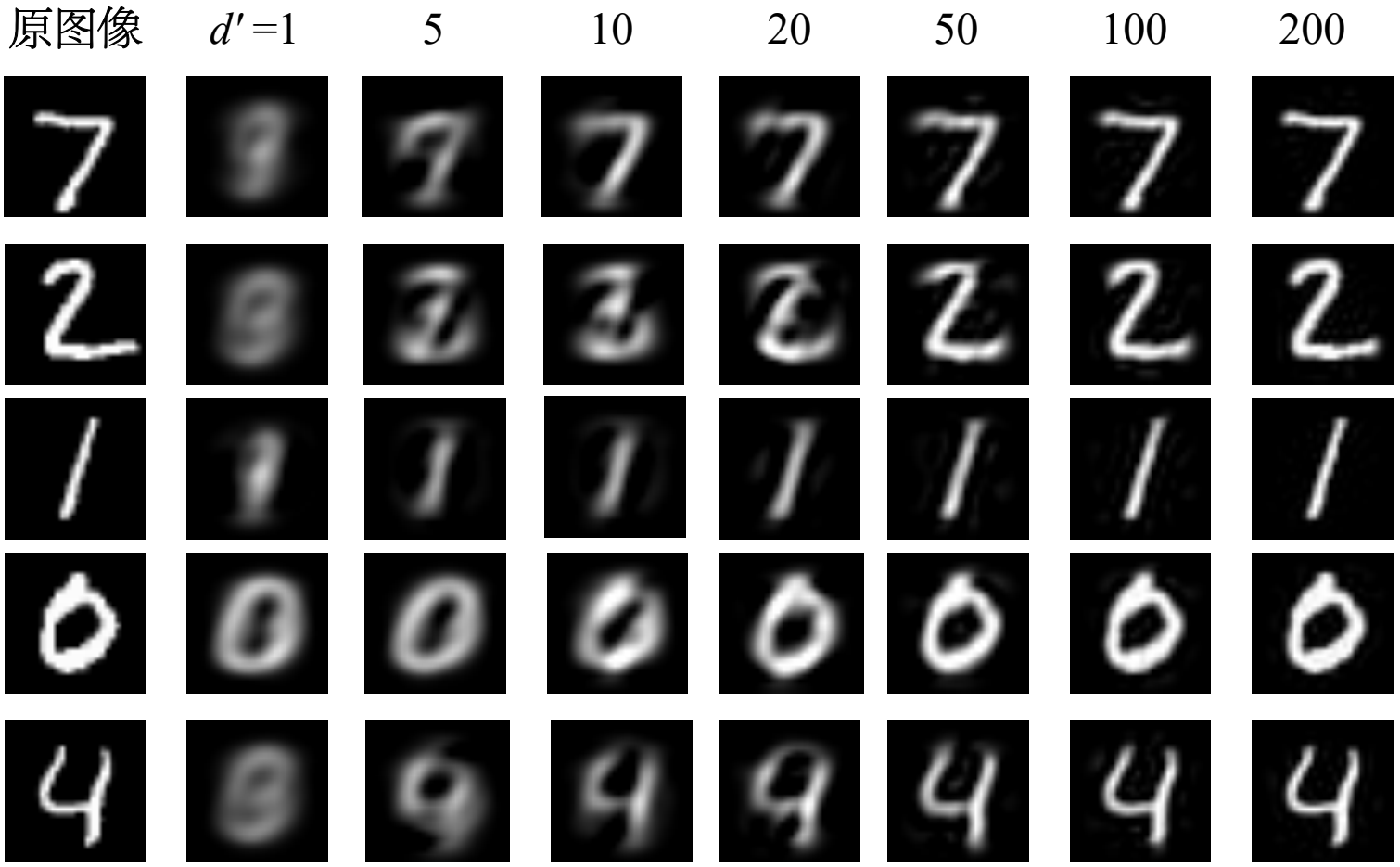
主成分分析应用

- 特征脸EigenFaces



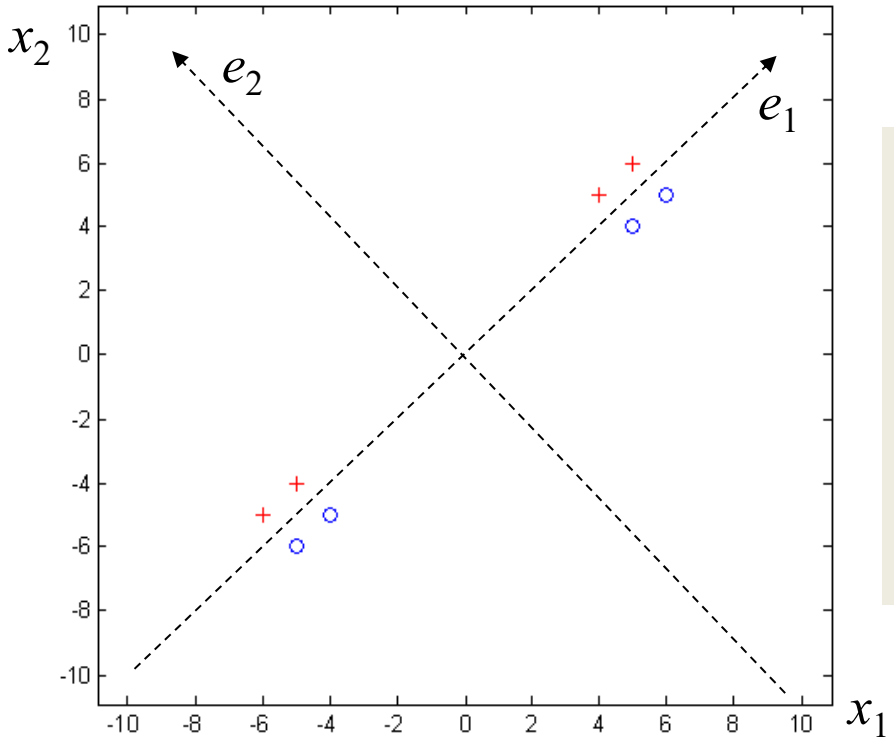
主成分分析应用

- 图像重建（基于不同个数的PCA成分）



基于Fisher准则的可分性分析

- PCA降维时没有考虑样本的类别属性，属于无监督学习。



PCA保留的投影方向未必包含类别可分信息，忽略的投影方向有可能包含了重要的可分性信息。

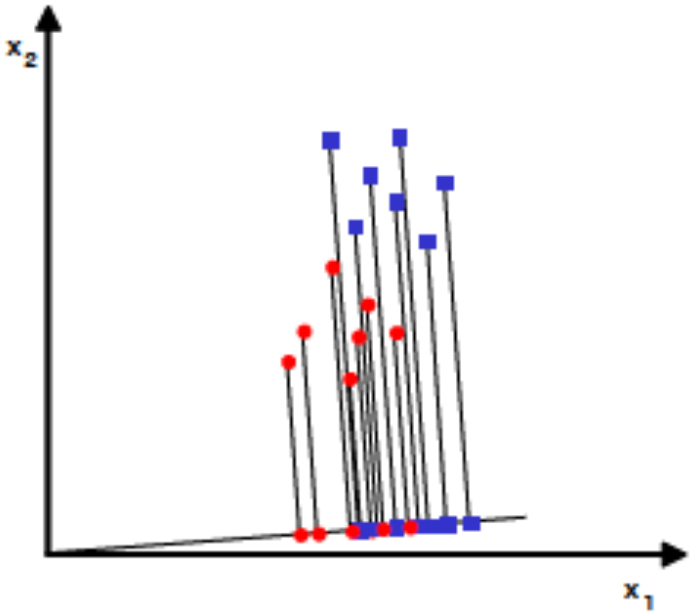
基于Fisher准则的可分性分析

- 基于Fisher准则的可分性分析（Fisher Discriminant Analysis，FDA）是在可分性最大意义下的最优线性映射，充分保留了样本的类别可分性信息。

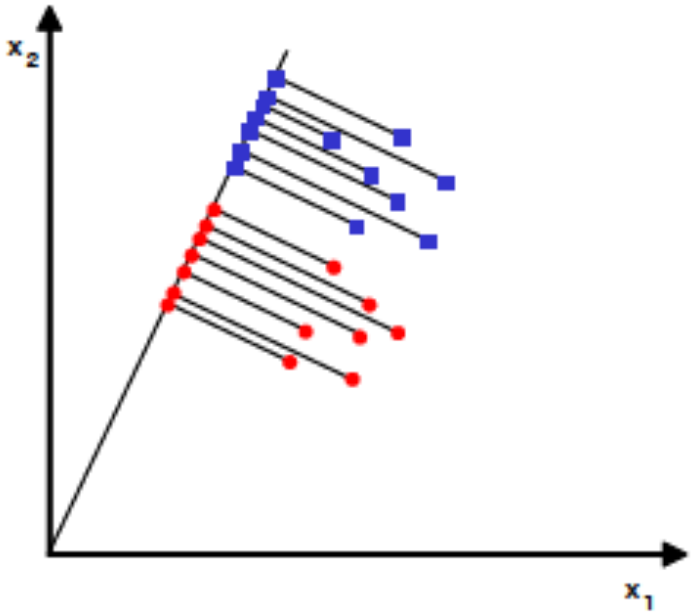
- 基于Fisher准则的可分性分析也被称为线性判别分析（Linear Discriminant Analysis，LDA）。

基于Fisher准则的可分性分析

- 如何选择对分类最优的投影方向？



差的投影
不同类别混在一起



好的投影
不同类别完好分开

Fisher可分性分析推导

- 假设两类问题的样本集为：$D_1 = \{\boldsymbol{x}_1^{(1)}, \cdots, \boldsymbol{x}_{n_1}^{(1)}\}$ 和 $D_2 = \{\boldsymbol{x}_1^{(2)}, \cdots, \boldsymbol{x}_{n_2}^{(2)}\}$，投影直线的单位矢量为 $\boldsymbol{w}$，则 d 维空间矢量 $\boldsymbol{x}$ 在这条直线的投影为标量：

$$y = \boldsymbol{w}^T \boldsymbol{x}$$

- 两类原始矢量样本集在投影后变为两个标量集：$D_1 \to Y_1 = \{y_1^{(1)}, \cdots, y_{n_1}^{(1)}\}$，$D_2 \to Y_2 = \{y_1^{(2)}, \cdots, y_{n_2}^{(2)}\}$。
- 按类别可分性判据，不同类样本越分散，同类样本越集中，则类别间的可分性高。

Fisher可分性分析推导

- 用两类的样本均值之差 $(\tilde{\mu}_1 - \tilde{\mu}_2)^2$ 度量两类样本间的分散程度；用两类样本各自的方差之和 $\tilde{s}_1^2 + \tilde{s}_2^2$ 度量样本类内的离散程度，可定义Fisher准则：

$$J(\boldsymbol{w}) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

- Fisher准则越大，则类别可分性越强。因此，最大化 $J(\boldsymbol{w})$ 就可以求出最优的$\boldsymbol{w}$。
- 下面计算 $J(\boldsymbol{w})$ 关于 $\boldsymbol{w}$ 的显式表达式。

Fisher可分性分析推导

- 投影后类别均值

$$\tilde{\mu}_i = \frac{1}{n_i}\sum_{y\in Y_i} y = \frac{1}{n_i}\sum_{x\in D_i} \boldsymbol{w}^T \boldsymbol{x} = \boldsymbol{w}^T \boldsymbol{\mu}_i,\ \ i = 1,2$$

- 投影后均值之差的平方可以表示为

$$\begin{aligned}(\tilde{\mu}_1 - \tilde{\mu}_2)^2 &= (\boldsymbol{w}^T \boldsymbol{\mu}_1 - \boldsymbol{w}^T \boldsymbol{\mu}_2)^2 \\ &= \boldsymbol{w}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{w} \\ &= \boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}\end{aligned}$$

其中 $\boldsymbol{S}_b$ 是类间散布矩阵。

Fisher可分性分析推导

- 类似地，计算类内样本方差：

$$\tilde{s}_i^2 = \sum_{y \in Y_i} (y - \tilde{\mu}_i)^2 = \sum_{\boldsymbol{x} \in D_i} (\boldsymbol{w}^T \boldsymbol{x} - \boldsymbol{w}^T \boldsymbol{\mu}_i)^2$$

$$= \sum_{\boldsymbol{x} \in D_i} \boldsymbol{w}^T (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{w}$$

$$= \boldsymbol{w}^T \boldsymbol{S}_i \boldsymbol{w}$$

其中 $\boldsymbol{S}_i$ 是第 i 类的类内散布矩阵。

- 总的方差为：

$$\tilde{s}_1^2 + \tilde{s}_2^2 = \boldsymbol{w}^T \boldsymbol{S}_1 \boldsymbol{w} + \boldsymbol{w}^T \boldsymbol{S}_2 \boldsymbol{w} = \boldsymbol{w}^T (\boldsymbol{S}_1 + \boldsymbol{S}_2) \boldsymbol{w} = \boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}$$

Fisher可分性分析推导

- 因此，Fisher准则函数变为：

$$J(\boldsymbol{w}) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}}$$

- 上式称为Rayleigh商优化问题。该问题存在无穷解，因为同方向不同长度的 $\boldsymbol{w}$ 有一样的准则值。
- 因此可适当调整 $\boldsymbol{w}$ 使得Fisher准则的分母 $\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}$ 等于常数 C，可得一个有约束的优化问题。

Fisher可分性分析推导

- 有约束的优化问题：

$$\max_{\boldsymbol{w}} J(\boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}$$

约束： $\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} = C$

- 构造拉格朗日函数转化为无约束优化问题：

$$L(\boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w} - \lambda(\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} - C)$$

- 上式对 $\boldsymbol{w}$ 求导

$$\frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}} = 2\boldsymbol{S}_b \boldsymbol{w} - 2\lambda \boldsymbol{S}_w \boldsymbol{w} = 0$$

- 设上式等于零，得到： $\boldsymbol{S}_b \boldsymbol{w} = \lambda \boldsymbol{S}_w \boldsymbol{w}$

Fisher可分性分析推导

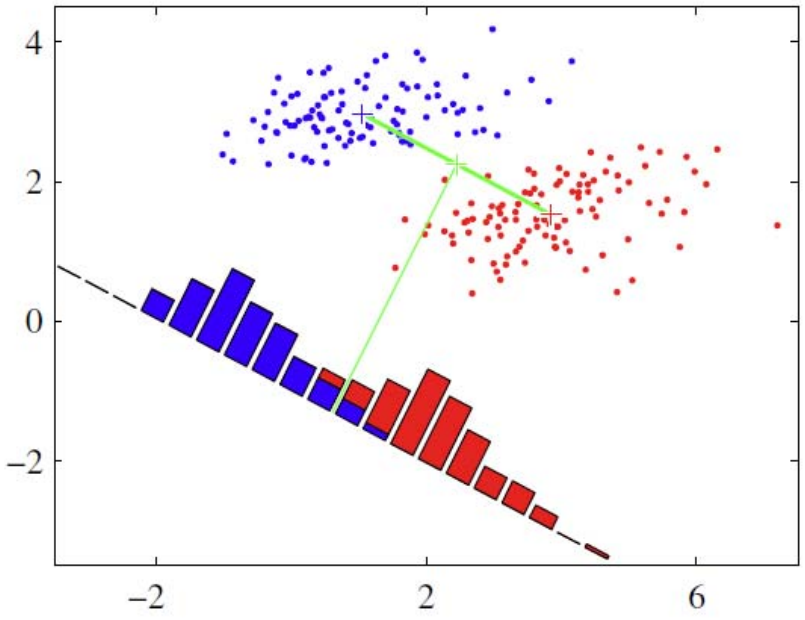
- 如果$\boldsymbol{S}_w$可逆，则有 $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b\boldsymbol{w}=\lambda\boldsymbol{w}$ 。
- 那么，λ 和 $\boldsymbol{w}$ 分别是矩阵$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ 的特征值和对应的特征向量。
- 将求解得到的一个特征矢量 $\boldsymbol{w}_i$ 带入 $J(\boldsymbol{w})$，得

$$J(\boldsymbol{w})=\frac{\boldsymbol{w}_i^T\boldsymbol{S}_b\boldsymbol{w}_i}{\boldsymbol{w}_i^T\boldsymbol{S}_w\boldsymbol{w}_i}=\frac{\lambda_i\boldsymbol{w}_i^T\boldsymbol{S}_w\boldsymbol{w}_i}{\boldsymbol{w}_i^T\boldsymbol{S}_w\boldsymbol{w}_i}=\lambda_i$$

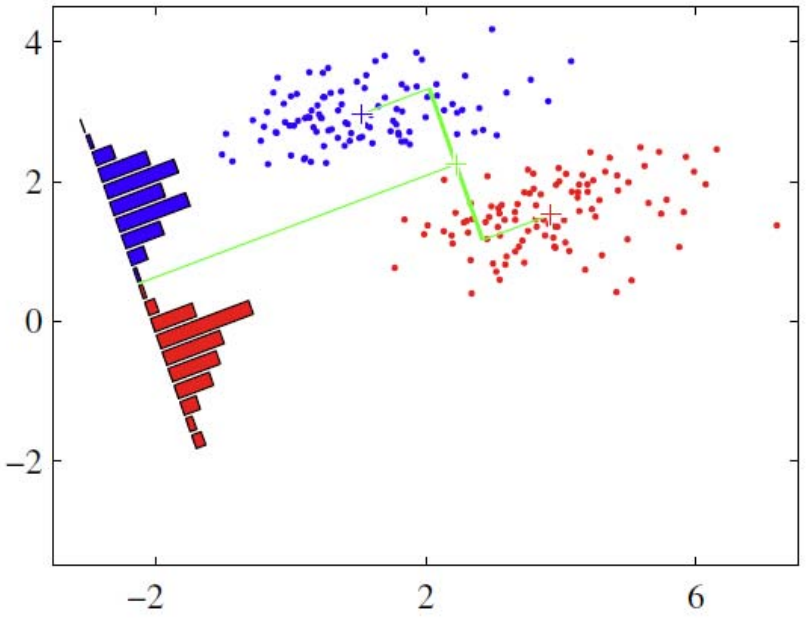
- 可见 $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ 最大特征值对应的特征矢量是使得Fisher准则取得最大值的方向矢量。

Fisher可分性分析推导

- 两个类别样本向一条直线上投影，当直线方向 $\boldsymbol{w}$ 为矩阵 $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ 最大特征值对应的特征矢量时，可以使得投影有最大可分性。



非FDA投影



FDA投影

Fisher可分性分析推导

- 推广到 c 个类别的样本向 d' 个方向上的投影，即将 d 维特征降到 d' 后，使得降维后的样本具有最大的可分型。
- 可定义矩阵：

$$\boldsymbol{S}_w = \sum_{i=1}^{c} \boldsymbol{S}_i, \qquad \boldsymbol{S}_b = \sum_{i=1}^{c} n_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$$

其中 $\boldsymbol{\mu}$ 是所有样本的均值矢量。

- 可证明使得Fisher准则最大的 d' 个投影矢量是对应于矩阵 $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ 最大的 d' 个特征值的特征矢量。

Fisher可分性分析算法

- 计算矩阵$\boldsymbol{S}_w$和$\boldsymbol{S}_b$；
- 计算矩阵$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$的特征值和特征矢量，特征值由大到小排序；
- 选择前d'个特征矢量作为列矢量构成矩阵：

$$\boldsymbol{E} = (\boldsymbol{e}_1 \ \ \boldsymbol{e}_2 \ \ \cdots \ \ \boldsymbol{e}_{d'});$$

- d维特征矢量$\boldsymbol{x}$可以转换为d'维特征矢量$\boldsymbol{x}'$：

$$\boldsymbol{x}' = \boldsymbol{E}^T\boldsymbol{x}\ 。$$

Fisher可分性分析讨论

- 经FDA变换后，新的坐标系不正交；$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ 不是对称矩阵，特征矢量不具正交性，变换后特征之间仍具有一定相关性。

- 当样本数足够多时，才能够保证类内散度矩阵 $\boldsymbol{S}_w$ 为非奇异矩阵（存在逆矩阵）；样本数少时，$\boldsymbol{S}_w$ 可能是奇异矩阵（此时可用奇异值分解求解）。

Fisher可分性分析讨论

- FDA需要确定一个参数 d'，即降维之后的特征矢量维数。
- 矩阵 $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ 特征值的大小描述了向相应特征矢量方向投影的可分性。
- 可证明，对于 c 个类别的样本集来说，矩阵 $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ 至多只存在 $c-1$ 个特征值大于等于0，其他的 $d-c+1$ 个特征值均为0。
- 因此FDA后新坐标维数最多为 $c-1$。

Fisher可分性分析讨论

- FDA可以作为线性分类器学习方法：
  1. 找到使两类样本可分性最强的投影方向 $\boldsymbol{w}$；
  2. 将所有的样本变换为1维特征 $\boldsymbol{w}^T\boldsymbol{x}$；
  3. 设定一个合适的阈值 $-w_0$ 进行分类

$$\boldsymbol{w}^T\boldsymbol{x}\begin{cases}\geq -w_0, & \boldsymbol{x}\in\omega_1\\ < -w_0, & \boldsymbol{x}\in\omega_2\end{cases}$$

可以由贝叶斯决策确定

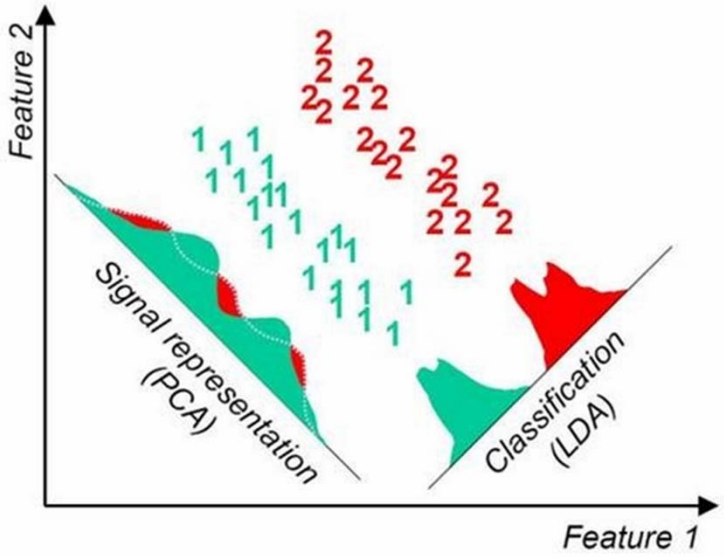
定义 $g(\boldsymbol{x})=\boldsymbol{w}^T\boldsymbol{x}+w_0$ ，则有

$$g(\boldsymbol{x})=\boldsymbol{w}^T\boldsymbol{x}+w_0\begin{cases}\geq 0, & \boldsymbol{x}\in\omega_1\\ < 0, & \boldsymbol{x}\in\omega_2\end{cases}$$

线性判别函数

PCA和FDA对比

	PCA	FDA
有无监督	无	有
特征值分解的矩阵	协方差矩阵 $\boldsymbol{\Sigma}$	$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$
第一投影方向	最大方差的方向	类别差异性最大的方向
最大维度	特征原始维度 d	$c-1$，c为类别数
坐标系是否正交	是	否
变化后特征是否相关	否	是



- ✓ PCA选择所有（不分类别）样本点投影具有最大方差的方向；
- ✓ FDA选择不同类别样本点投影差异性最大的方向。

其他特征提取方法

- 其他成分分析方法：
  - 独立成分分析（Independent Component Analysis，ICA）
  - 多维尺度变换（Multidimensional Scaling，MDS）
  - 典型相关分析（Canonical Correlation Analysis，CCA）
  - 偏最小二乘（Partial Least Squares，PLS）

- 非线性变化方法：
  - 流形学习（Manifold Learning）
  - 线性方法结合“核方法”（Kernel Methods）

本章小结

- 介绍了特征选择与特征提取的在模式识别中的必要性和两者的区别
- 介绍了多种基于距离和基于散布矩阵的类别可分性判据
- 介绍了特征选择的三大类方法：滤波法、包装法（包括分支定界法、次优搜索算法和递归式特征消除）、嵌入法

本章小结

- 介绍了两类主要的特征提取方法：主成分分析和Fisher可分性分析
- 介绍了主成分分析的算法、推导、应用、参数选择和注意事项等
- 介绍了Fisher可分性分析的算法、推导、应用、参数选择和注意事项等