模式识别
第七章 特征选择与提取

内容

* 基本概念
* 特征选择
* 特征提取

基本概念

* 特征选择和提取是模式识别中的一个关键问题
  * 前面讨论分类器设计的时候，一直假定已给出了特征向量维数确定的样本集，其中各样本的每一维都是该样本的一个特征；
  * 这些特征的选择是很重要的，它强烈地影响到分类器的设计及其性能；
  * 假若对不同的类别，这些特征的差别很大，具有良好区分能力的特征则比较容易设计出具有较好性能分类系统。

基本概念

* 特征选择和提取是构造模式识别系统时的一个重要部分
  * 在很多实际问题中，往往不容易找到那些最重要的特征，或受客观条件的限制，不能对它们进行有效的测量；
  * 因此在测量时，由于人们心理上的作用，只要条件许可总希望把特征取得多一些；
  * 另外，由于客观上的需要，为了突出某些有用信息，抑制无用信息，有意加上一些比值、指数或对数等组合计算特征；
  * 特征取多一些有没有问题?

基本概念

* 如果将数目很多的测量值不做分析，全部直接用作分类特征，不但耗时，而且会影响到分类的效果，产生“特征维数灾难”问题。
* 为了设计出效果好的分类器，通常需要对原始的测量值集合进行分析，经过选择或变换处理，组成有效的识别特征；
* 在保证一定分类精度的前提下，减少特征维数，即进行“降维”处理，使分类器实现快速、准确和高效的分类。

基本概念

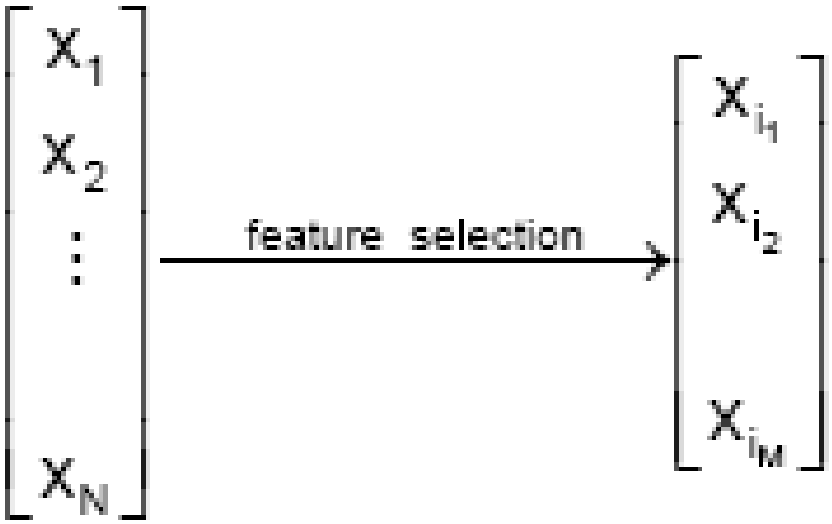
* 为达到上述目的，关键是所提供的识别特征应具有很好的可分性，使分类器容易判别。为此，需对特征进行选择：
  * 应去掉模棱两可、不易判别的特征；
  * 所提供的特征不要重复，即去掉那些相关性强且没有增加更多分类信息的特征。

基本概念

* 所谓特征选择，就是从n个度量值集合{x_1, x_2,…, x_n}中，按某一准则选取出供分类用的子集，作为降维（m维，m<n）的分类特征
* 所谓特征提取，就是使(x_1, x_2,…, x_n)通过某种变换，产生m个特征(y_1, y_2,…, y_m)(m<n)，作为新的分类特征（或称为二次特征）
* 其目的都是为了在尽可能保留识别信息的前提下，降低特征空间的维数，以达到有效的分类

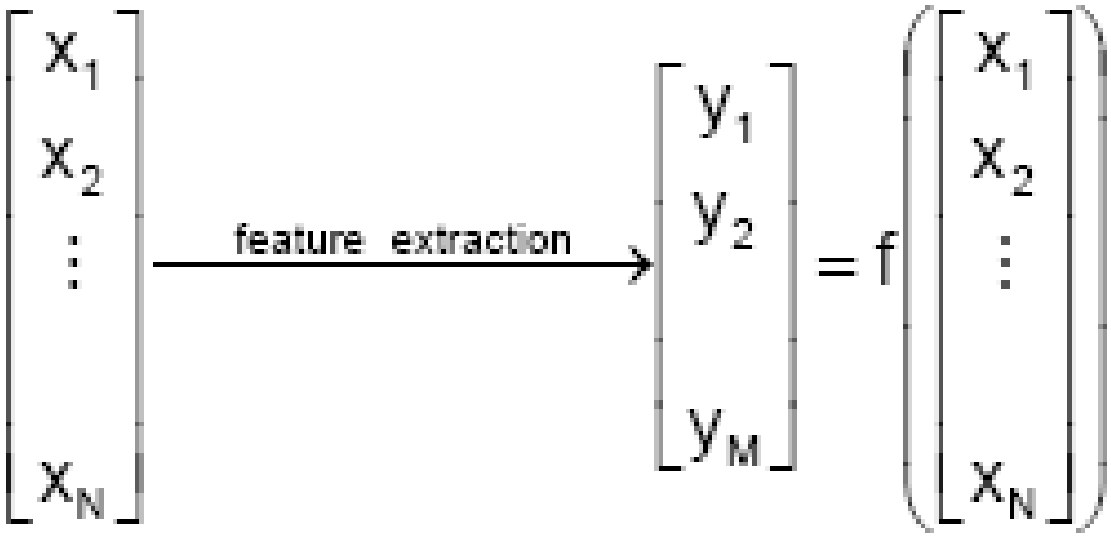
基本概念-数学表示

* 特征选择：在当前的特征集中选择一个子集，没有变换。



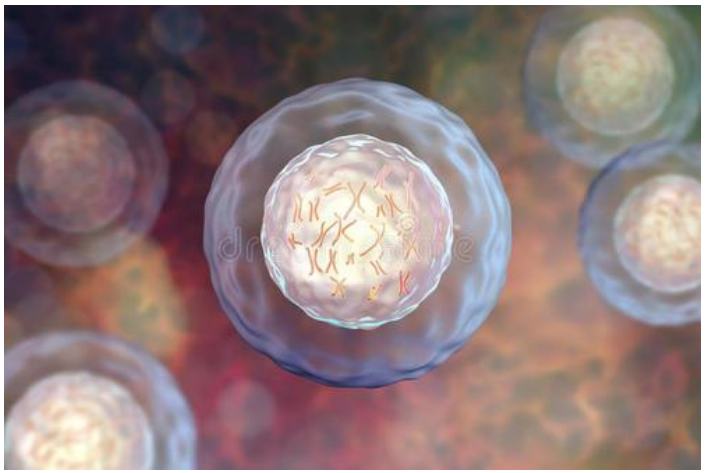
基本概念-数学表示

* 特征提取：将当前的特征集变换到一个更低维度的空间中。



基本概念-示例

* 以细胞自动识别为例：
* 通过图像输入得到一批包括正常细胞和异常细胞的图像，我们的任务是根据这些图像区分哪些细胞是正常的，哪些细胞是异常的



基本概念-示例

* 首先找出一组能代表细胞性质的特征，为此可计算
  * 细胞总面积
  * 总光密度
  * 胞核面积
  * 核浆比
  * 细胞形状
  * 核内纹理
  * ……

基本概念-示例

* 这样产生出来的原始特征可能很多（几十甚至几百个），或者说原始特征空间维数很高，需要降低（或称压缩）维数以便分类
* 一种方式是从原始特征中挑选出一些最有代表性的特征，称之为特征选择
* 另一种方式是用映射（或称变换）的方法把原始特征变换为较少的特征，称之为特征提取

基本概念 – 问题

* 特征是否越多越好?
* 样本的精确表示是否是适合于分类的特征?

特征选择

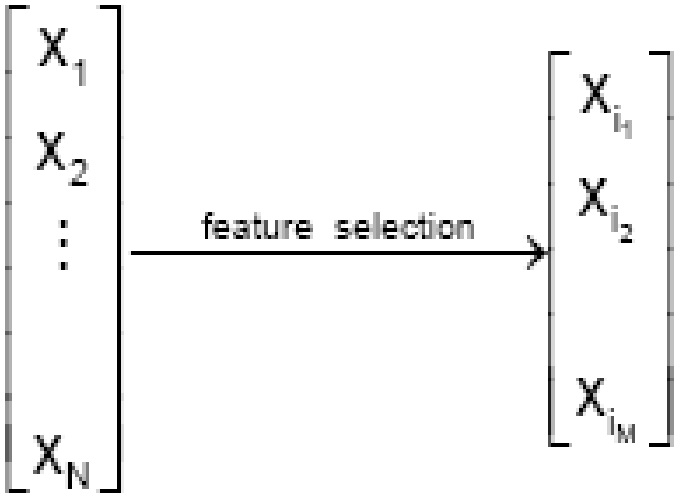
* 严格定义：
  * 给定一个特征集$X = \{x_i | i = 1, 2, \ldots, N\}$
  * 找到该特征集的子集

$$Y_M = \{x_{i_1}, x_{i_2}, \ldots, x_{i_M}\}, M < N$$

  * 使得所定义的目标函数$J(Y)$最优
  * 注：该目标函数在理想状态下，是正确分类的概率

特征选择

* 严格定义的数学表达：



$$\{x_{i_1}, x_{i_2}, \ldots, x_{i_M}\} = \underset{M, j_m}{\operatorname{argmax}}\left[J\{x_i \mid i = 1 \ldots N\}\right]$$

特征选择

* 一个问题：
* 为什么我们在有通用特征提取方法的情况下需要特征选择?

特征选择

* 特征选择在部分情况下是必须的
  * 特征获取代价昂贵
    * 传感器使用中需要对传感器进行挑选
  * 保持特征的物理特性
    * 特征变换时原特征的物理意义（单位长度、重量等）通常会消失
  * 特征可能不是量化的
    * 在机器学习问题中时有发生

特征选择

* 另外，更少的特征意味着模式识别系统的参数量更少
  * 增强通用性
  * 降低计算复杂度和时间

* 特征选择可以降低特征维数！

特征选择－常见方法

unsupervised

	Linear	Nonlinear
Selection	Correlation between inputs	Mutual information between inputs
Projection	Principal Component Analysis	Sammon's Mapping, Kohonen maps

supervised

	Linear	Nonlinear
Selection	Correlation between inputs and output	Mutual information between inputs and outputs, Greedy algorithms, Genetic algorithms
Projection	Linear Discriminant Analysis, Partial Least Squares	Projection pursuit

特征选择

* 为什么要降维?
* 理论上来说没用
  * 更多的信息意味着分类任务更容易完成
  * 模型可以忽略无关特征
* 理论上来说，理论和实际相同，但是实际中不是!

特征选择

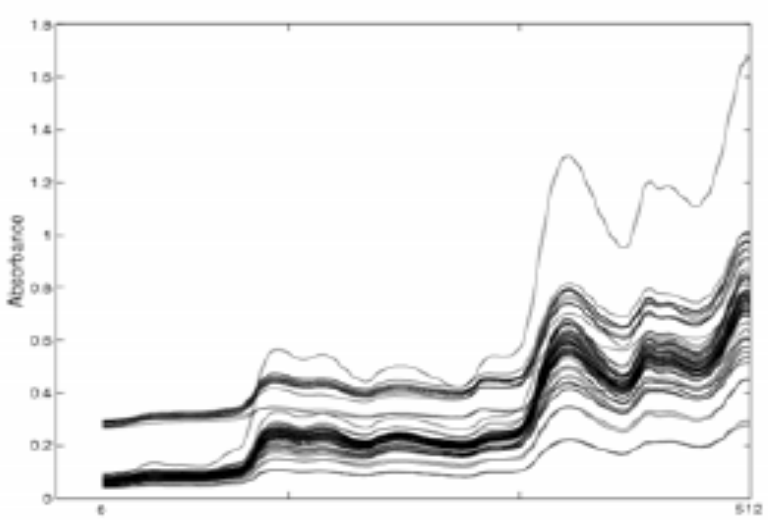
* 实际中，更多的特征输入意味着：
  * 更多的参数
  * 更大的输入空间
* 特征维度灾难和过拟合的风险增加！

特征选择 – 示例

Example : Spectrophotometry

To predict sugar concentration in an orange juce sample from light absorbtion spectra

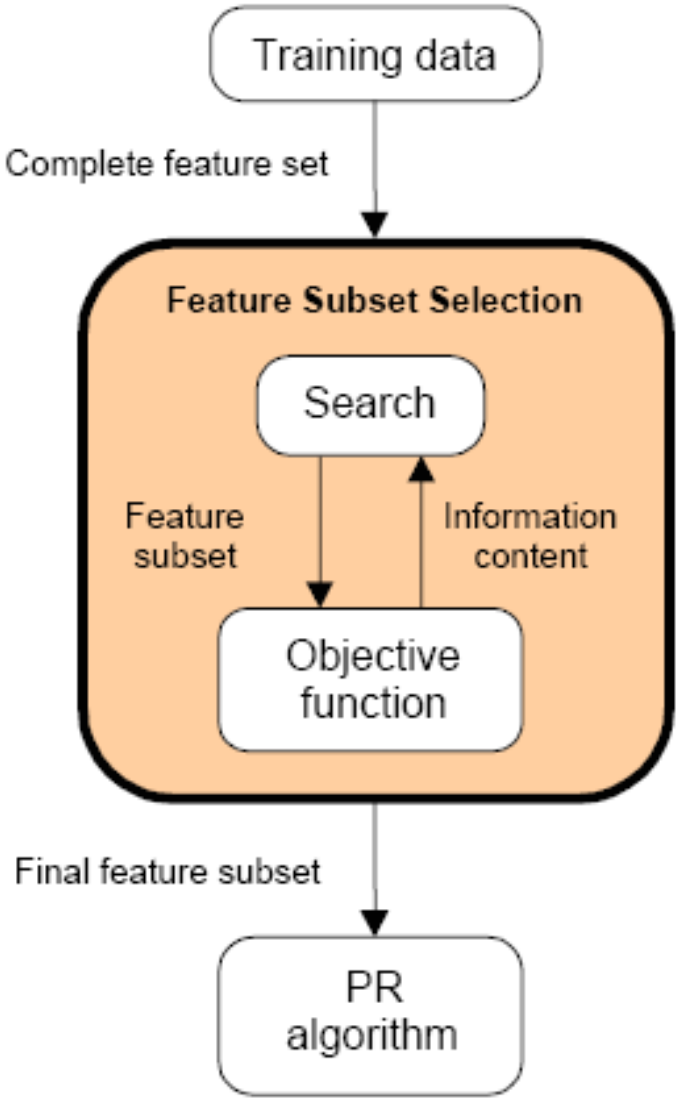
115 samples in dimension **512**



Even a linear model would lead to overfitting !

特征选择－方法

* 特征选择方法的关键
  * 关键1－子集相关性衡量－目标函数
  * 关键2－最优子集搜索



特征选择－方法

Search Strategy

- Exhaustive evaluation of feature subsets involves $\binom{N}{M}$ combinations for a fixed value of M, and 2^N combinations if M must be optimized as well
  - This number of combinations is unfeasible, even for moderate values of M and N, so a search procedure must be used in practice
  - For example, exhaustive evaluation of 10 out of 20 features involves 184,756 feature subsets; exhaustive evaluation of 10 out of 100 involves more than 10^{13} feature subsets [Devijver and Kittler, 1982]
- A search strategy is therefore needed to direct the FSS process as it explores the space of all possible combination of features

特征选择－方法

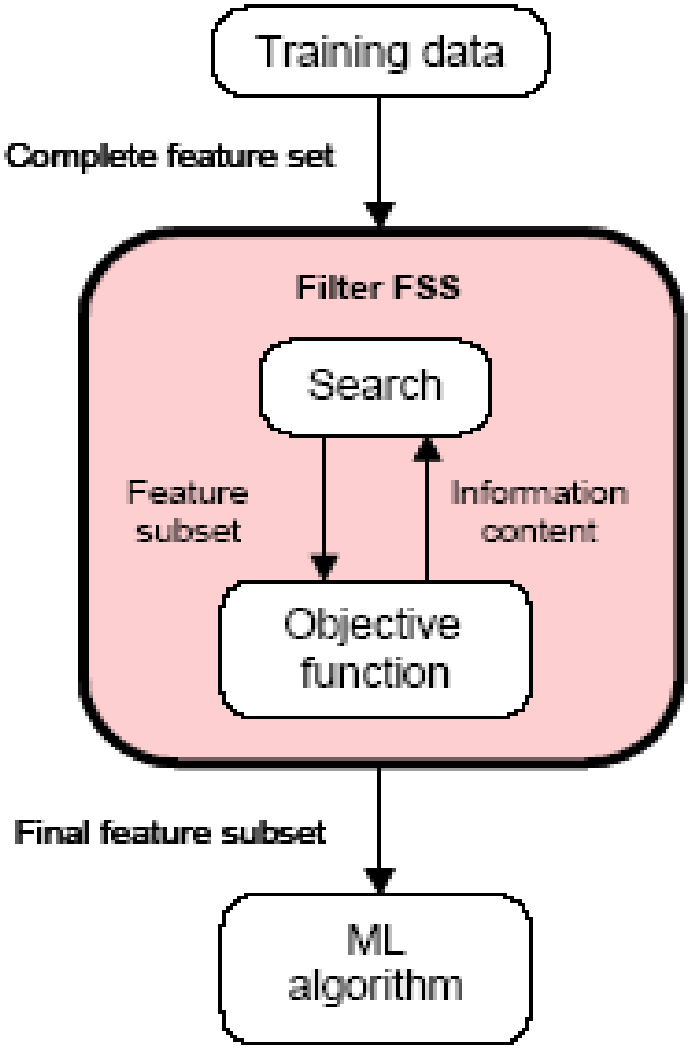
Objective Function

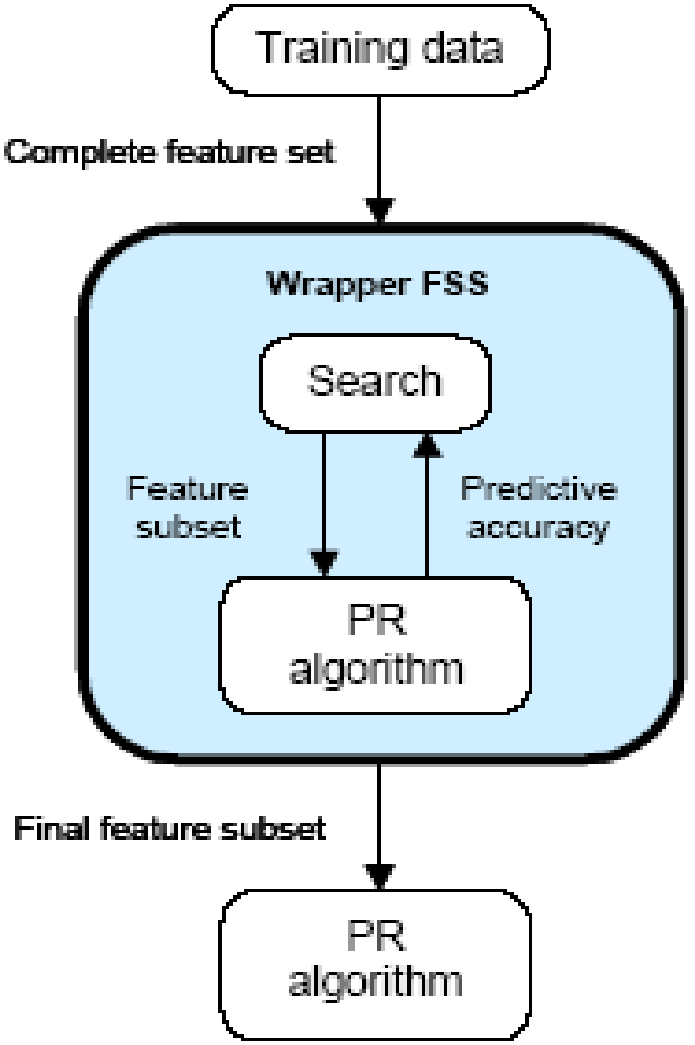
- The objective function evaluates candidate subsets and returns a measure of their "goodness", a feedback signal used by the search strategy to select new candidates

Objective functions are divided in two groups

- **Filters**: The objective function evaluates feature subsets by their information content, typically interclass distance, statistical dependence or information-theoretic measures
- **Wrappers**: The objective function is a pattern classifier, which evaluates feature subsets by their predictive accuracy (recognition rate on test data) by statistical resampling or cross-validation

特征选择－方法


Training data
Complete feature set
Filter FSS
Search
Feature subset
Information content
Objective function
Final feature subset
ML algorithm


Training data
Complete feature set
Wrapper FSS
Search
Feature subset
Predictive accuracy
PR algorithm
Final feature subset
PR algorithm

Filter types

- **Distance or separability measures**
  - These methods use distance metrics to measure class separability, such as
    - Distance between classes: Euclidean, Mahalanobis, etc.
    - Determinant of $S_W^{-1}S_B$ (LDA eigenvalues)
- **Correlation and information-theoretic measures**
  - These methods are based on the rationale that good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other
  - Linear relation measures
    - Linear relationship between variables can be measured using the correlation coefficient $$J(Y_M) = \frac{\sum_{i=1}^{M} \rho_{ic}}{\sum_{i=1}^{M}\sum_{j=i+1}^{M} \rho_{ij}}$$
      - Where ρ_{ic} is the correlation coefficient between feature 'i' and the class label and ρ_{ij} is the correlation coefficient between features 'i' and 'j'
  - Non-Linear relation measures
    - Correlation is only capable of measuring linear dependence. A more powerful measure is the mutual information $I(Y_k;C)$
    $$J(Y_M) = I(Y_M;C) = H(C) - H(C \mid Y_M) = \sum_{c=1}^{C} \int_{Y_M} P(Y_M, \omega_c) \lg \frac{P(Y_M, \omega_c)}{P(Y_M)\,P(\omega_c)} dx$$
    - The mutual information between the feature vector and the class label $I(Y_M;C)$ measures the amount by which the uncertainty in the class $H(C)$ is decreased by knowledge of the feature vector $H(C|Y_M)$, where $H(\cdot)$ is the entropy function
      - Note that mutual information requires the computation of the multivariate densities $P(Y_M)$ and $P(Y_M,\omega_C)$, which is an ill-posed problem for high-dimensional spaces. In practice [Battiti, 1994], mutual information is replaced by a heuristic like
      $$J(Y_M) = \sum_{m=1}^{M} I(x_{l_m};C) - \beta \sum_{m=1}^{M} \sum_{n=m+1}^{M} I(x_{l_m};x_{l_n})$$

Filters vs. Wrappers

Filters

- Advantages
  - **Fast execution**: Filters generally involve a non-iterative computation on the dataset, which can execute much faster than a classifier training session
  - **Generality**: Since filters evaluate the intrinsic properties of the data, rather than their interactions with a particular classifier, their results exhibit more generality: the solution will be "good" for a larger family of classifiers
- Disadvantages
  - **Tendency to select large subsets**: Since the filter objective functions are generally monotonic, the filter tends to select the full feature set as the optimal solution. This forces the user to select an arbitrary cutoff on the number of features to be selected

Wrappers

- Advantages
  - **Accuracy**: wrappers generally achieve better recognition rates than filters since they are tuned to the specific interactions between the classifier and the dataset
  - **Ability to generalize**: wrappers have a mechanism to avoid overfitting, since they typically use cross-validation measures of predictive accuracy
- Disadvantages
  - **Slow execution**: since the wrapper must train a classifier for each feature subset (or several classifiers if cross-validation is used), the method can become unfeasible for computationally intensive methods
  - **Lack of generality**: the solution lacks generality since it is tied to the bias of the classifier used in the evaluation function. The "optimal" feature subset will be specific to the classifier under consideration

特征选择 – Filter方法示例

* 示例：
* 基于类别可分性判据的方法

特征选择 – Filter方法示例

* 距离和散布矩阵
* [类内散布矩阵]
  * 对属于同一类的模式样本，类内散布矩阵表示各样本点围绕其均值周围的散布情况，这里即为该分布的协方差矩阵。
* [类间距离和类间散布矩阵]
* [多类模式集散布矩阵]
* 以上各类散布矩阵反映了各类模式在模式空间的分布情况，但它们与分类的错误率没有直接联系。

特征选择 – Filter方法示例

* 设有n个可用作分类的测量值，为了在不降低（或尽量不降低）分类精度的前提下，减小特征空间的维数以减少计算量，需从中直接选出m个作为分类的特征。
* 问题：在n个测量值中选出哪一些作为分类特征，使其具有最小的分类错误?

特征选择 – Filter方法示例

* 从n个测量值中选出m个特征，一共有C_n^m种可能的选法。
* 一种“穷举”办法：对每种选法都用训练样本试分类一下，测出其正确分类率，然后做出性能最好的选择，此时需要试探的特征子集的种类达到$C_n^m = \frac{n!}{m!(n-m)!}$种，非常耗时。
* 需寻找一种简便的可分性准则，间接判断每一种子集的优劣。
  * 对于独立特征的选择准则
  * 一般特征的散布矩阵准则

特征选择 – Filter方法示例

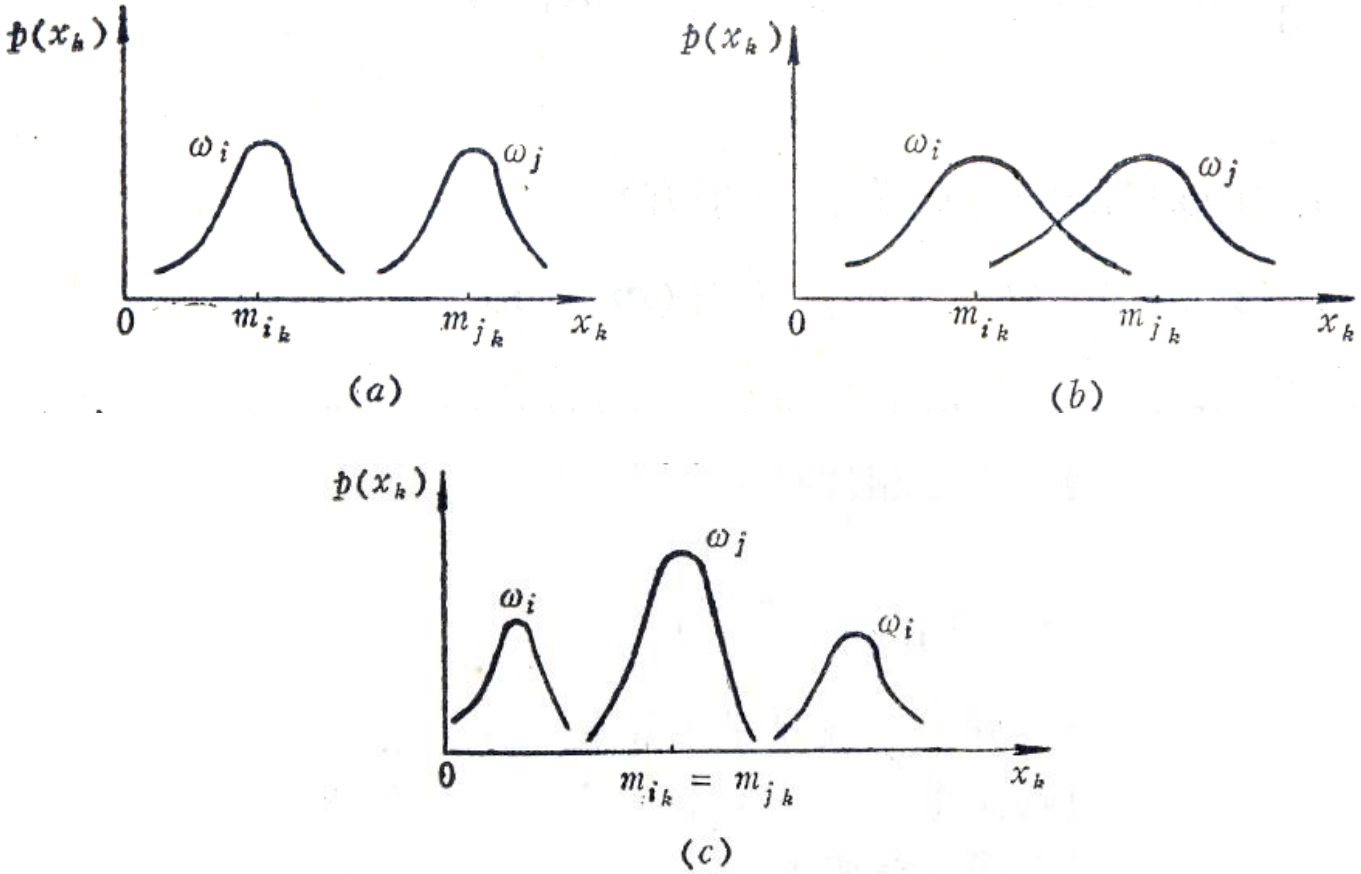
* 对于独立特征的选择准则
  * 类别可分性准则应具有这样的特点，即不同类别模式特征的均值向量之间的距离应最大，而属于同一类的模式特征，其方差之和应最小。
  * 假设各原始特征测量值是统计独立的，此时，只需对训练样本的n个测量值独立地进行分析，从中选出m个最好的作为分类特征即可

* [例：对于ω_i和ω_j两类训练样本的特征选择]

特征选择 – Filter方法示例

* 基于距离测度的可分性准则，其适用范围与模式特征的概率分布有关。
* 三种不同模式分布的情况
  * (a)中特征x_k的分布有很好的可分性，通过它足以分离ω_i和ω_j两种类别；
  * (b)中的特征分布有很大的重叠，单靠x_k达不到较好的分类，需要增加其它特征；
  * (c)中的ω_i类特征x_k的分布有两个最大值，虽然它与ω_j的分布没有重叠，但计算G_k约等于0，此时再利用G_k作为可分性准则已不合适。

特征选择 – Filter方法示例



特征选择 – Filter方法示例

* 因此，假若类概率密度函数不是或不近似正态分布，均值和方差就不足以用来估计类别的可分性，此时该准则函数不完全适用。

特征选择 – Filter方法示例

* 一般特征的散布矩阵准则
  * [类内、类间和总体的散布矩阵S_w、S_b和S_t]
  * S_w的行列式值越小且S_b的行列式值越大，可分性越好。
  * [散布矩阵准则J_1和J_2形式]
  * 使J_1或J_2最大的子集可作为所选择的分类特征
* 注：这里计算的散布矩阵不受模式分布形式的限制，但需要有足够数量的模式样本才能获得有效的结果。

特征选择 – Filter方法示例

* 参考文献
  * Jain A.,Zongker D., “Feature selection: Evaluation, Application, and small sample performance”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.19(2),pp153-158, 1997
  * Zhang H., Sun G,“Feature Selection Using Tabu Search Method”, Pattern Recognition, Vol.35, pp701-711, 2002

特征提取

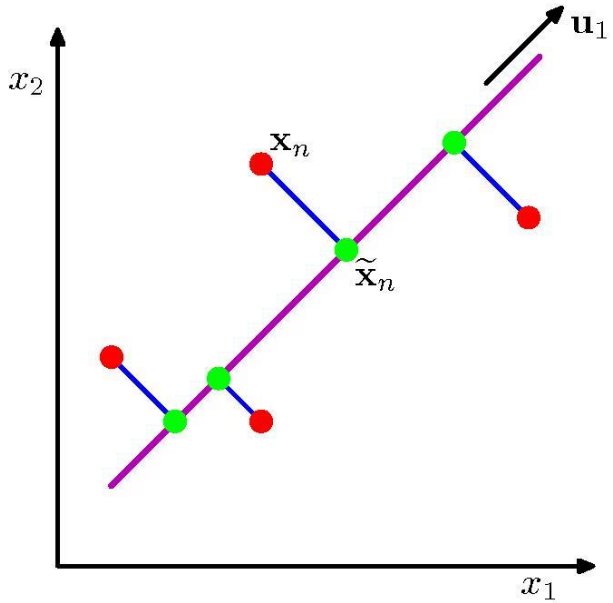
* 基于变换/映射的方法

特征提取-主成分分析概述

* K. Pearson(1901年论文) 针对非随机变量
* H. Hotelling(1933年论文) 推广到随机向量

特征提取-主成分分析概述

* 主成分分析(Principal Component Analysis, PCA)，将原有众多具有一定相关性的指标重新组合成一组少量互相无关的综合指标。



使得降维后样本的方差尽可能大

使得降维后数据的均方误差尽可能小

特征提取-主成分分析算法

* 最大方差思想
* 使用较少的数据维度保留住较多的原数据特性
* 将D维数据集 $\{\mathbf{x}_n\}, n = 1, 2, ..., N$ 降为M维，M<D

特征提取-主成分分析算法

* 首先考虑M=1，定义这个空间的投影方向为D维向量$\mathbf{u}_1$
* 出于方便且不失一般性，令$\mathbf{u}_1^T\mathbf{u}_1 = 1$
* 每个数据点$\mathbf{x}_n$在新空间中表示为标量$\mathbf{u}_1^T\mathbf{x}_n$
* 样本均值在新空间中表示为$\mathbf{u}_1^T\bar{\mathbf{x}}$，其$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$
* 投影后样本方差表示为$\frac{1}{N}\sum_{n=1}^{N}\{\mathbf{u}_1^T\mathbf{x}_n - \mathbf{u}_1^T\bar{\mathbf{x}}\}^2 = \mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$
* 其中原样本方差

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$$

特征提取-主成分分析算法

* 目标是最大化 $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1, \quad s.t. \quad \mathbf{u}_1^T\mathbf{u}_1 = 1$
* 利用拉格朗日乘子法 $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^T\mathbf{u}_1)$
* 对$\mathbf{u}_1$求导置零得到 $\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1$

u_1是S的特征向量

* 进一步得到$\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 = \lambda_1$

u_1是S最大特征值对应的特征向量时
方差取到极大值，称u_1为第一主成分

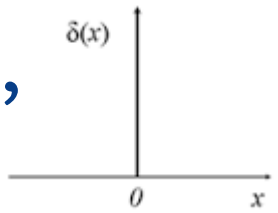
特征提取-主成分分析算法

* 考虑更一般性的情况(M>1), 新空间中数据方差最大的最佳投影方向由协方差矩阵S的M个特征向量$\mathbf{u}_1, \ldots, \mathbf{u}_M$定义, 其分别对应M个最大的特征值
* 首先获得方差最大的1维，生成该维的补空间；
* 继续在补空间中获得方差最大的1维，生成新的补空间；
* 依次循环下去得到M维的空间。

特征提取-主成分分析算法

* 最小均方误差思想
* 使原数据与降维后的数据(在原空间中的重建)的误差最小
* 定义一组正交的D维基向量 $\{\mathbf{u}_i\}, i = 1, ..., D$，满足

$$\mathbf{u}_i^T \mathbf{u}_j = \delta_{ij}$$



* 由于基是完全的，每个数据点可以表示为基向量的线性组合

$$\mathbf{x}_n = \sum_{i=1}^{D} \alpha_{ni} \mathbf{u}_i$$

特征提取-主成分分析算法

* 相当于进行了坐标变换

$$\{\mathbf{x}_{n1},\ldots,\mathbf{x}_{nD}\} \xrightarrow{\{\mathbf{u}_i\}} \{\alpha_{n1},\ldots,\alpha_{nD}\}$$

$$\alpha_{nj} = \mathbf{x}_n^T \mathbf{u}_j$$

* 那么

$$\mathbf{x}_n = \sum_{i=1}^{D} (\mathbf{x}_n^T \mathbf{u}_i)\mathbf{u}_i$$

特征提取-主成分分析算法

* 在M维变量(M<D)生成的空间中对其进行表示

$$\tilde{\mathbf{x}}_n = \sum_{i=1}^{M} z_{ni}\mathbf{u}_i + \sum_{i=M+1}^{D} b_i\mathbf{u}_i$$

独特的　　共享的

* 目标最小化失真度

$$J = \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \tilde{\mathbf{x}}_n\|^2$$

* 导数置零得到

$$z_{nj} = \mathbf{x}_n^T\mathbf{u}_j, j = 1, ..., M$$

$$b_j = \bar{\mathbf{x}}^T\mathbf{u}_j, j = M+1, ..., D$$

特征提取-主成分分析算法

* 有 $$\mathbf{x}_n - \tilde{\mathbf{x}}_n = \sum_{i=M+1}^{D} \{(\mathbf{x}_n - \bar{\mathbf{x}})^T \mathbf{u}_i\}\mathbf{u}_i$$

$$J = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=M+1}^{D}(\mathbf{x}_n^T\mathbf{u}_i - \bar{\mathbf{x}}^T\mathbf{u}_i)^2 = \sum_{i=M+1}^{D}\mathbf{u}_i^T\mathbf{S}\mathbf{u}_i$$

* 拉格朗日乘子法

$$\tilde{J} = \sum_{i=M+1}^{D}\mathbf{u}_i^T\mathbf{S}\mathbf{u}_i + \sum_{i=M+1}^{D}\lambda_i(1 - \mathbf{u}_i^T\mathbf{u}_i)$$

* 求导得到

$$\mathbf{S}\mathbf{u}_i = \lambda_i\mathbf{u}_i$$

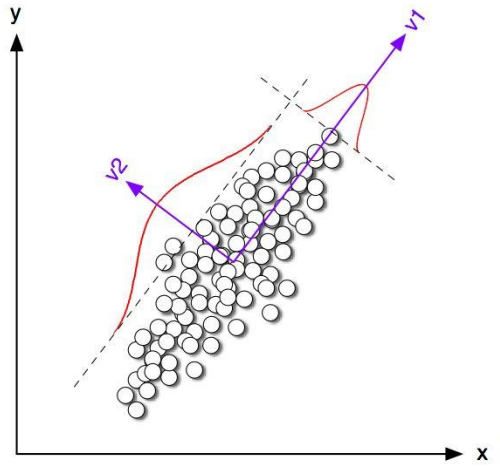
*J*最小时取*D-M*个最小的特征值

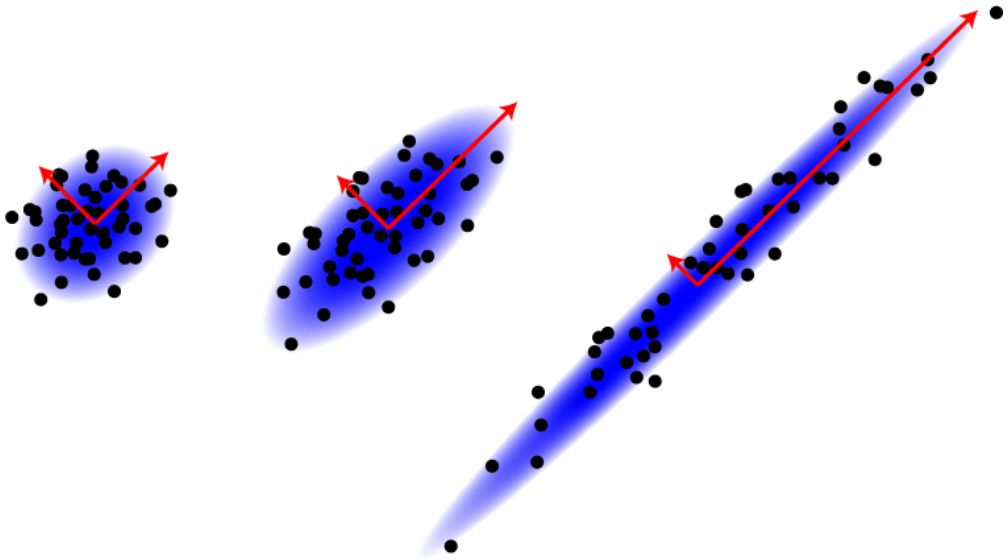
* 对应失真度为 $J = \sum_{i=M+1}^{D}\lambda_i$ 主子空间对应*M*个最大特征值

特征提取-主成分分析算法

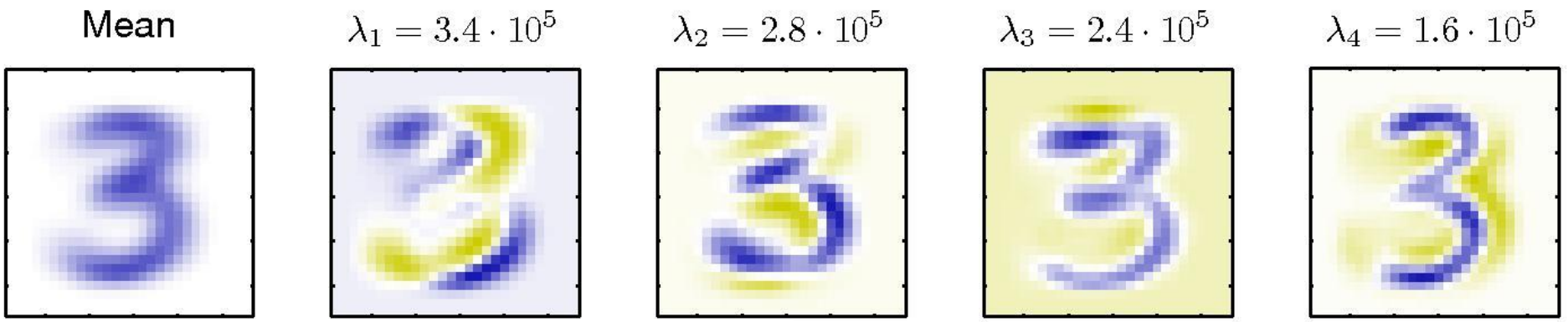
* 主成分分析算法计算步骤
* ①计算给定样本$\{\mathbf{x}_n\}, n = 1, 2, ..., N$的均值$\bar{\mathbf{x}}$和协方差矩阵S;
* ②计算S的特征向量与特征值;
* ③将特征值从大到小排列，前M个特征值$\lambda_1, ..., \lambda_M$所对应的特征向量$\mathbf{u}_1, ..., \mathbf{u}_M$构成投影矩阵。

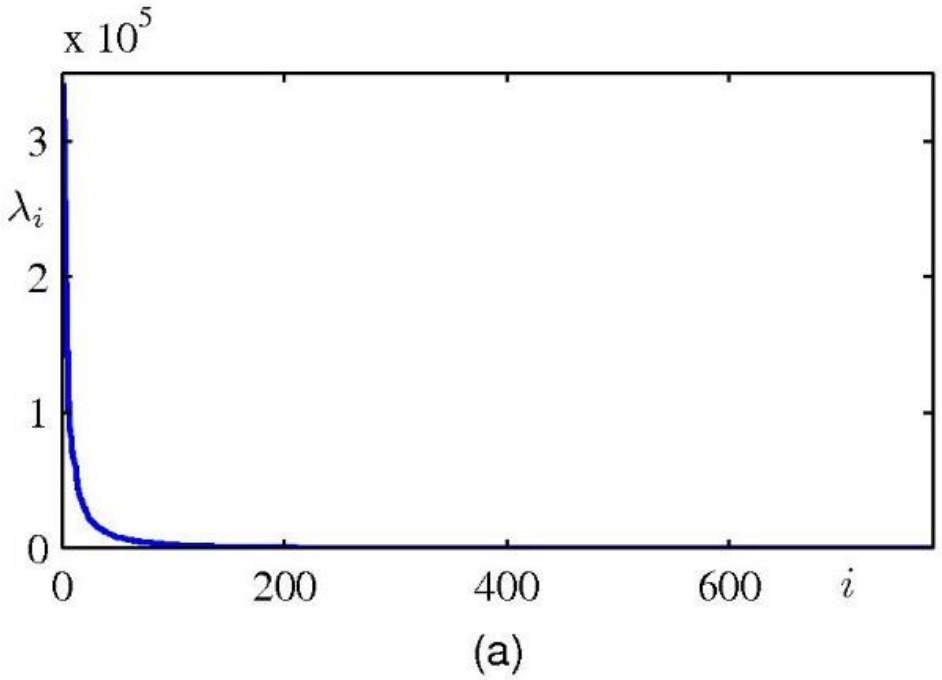
特征提取-主成分分析算法





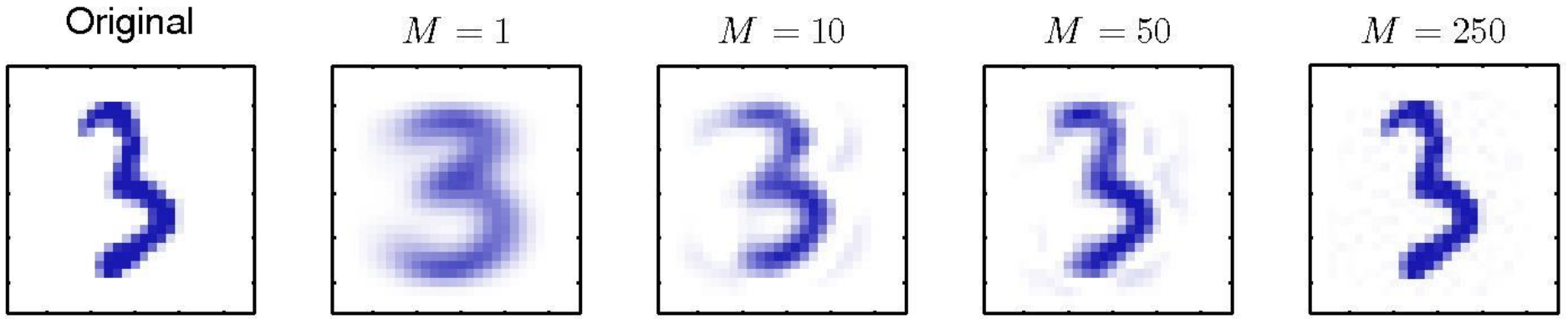
特征提取-主成分分析的应用

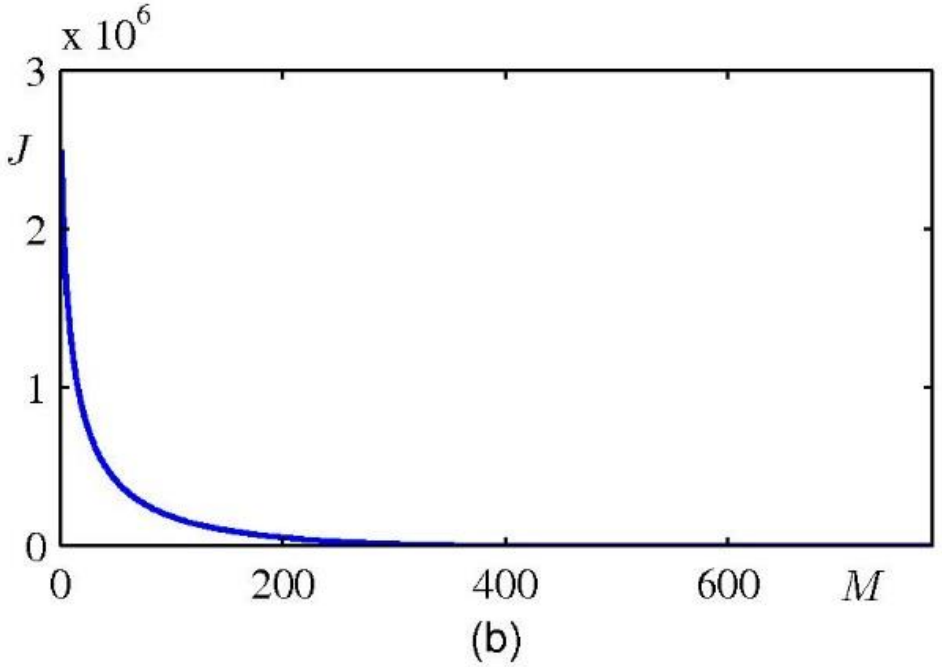




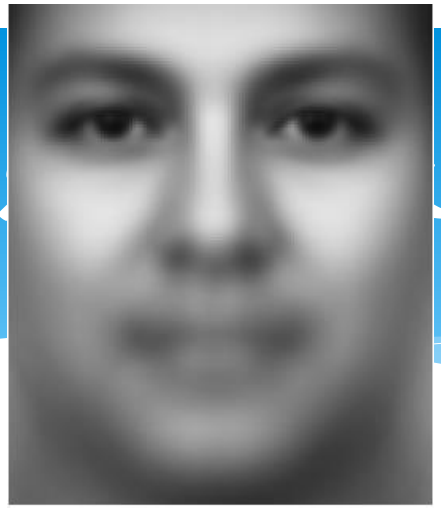
特征值分布谱
特征值由大到小排列

特征提取-主成分分析的应用





失真度分布谱
随M取值由小到大排列



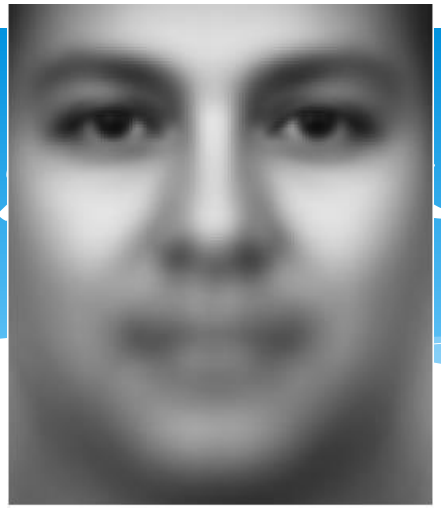
-主成分分析的应用

特征脸(Eigenfaces)#1~#8







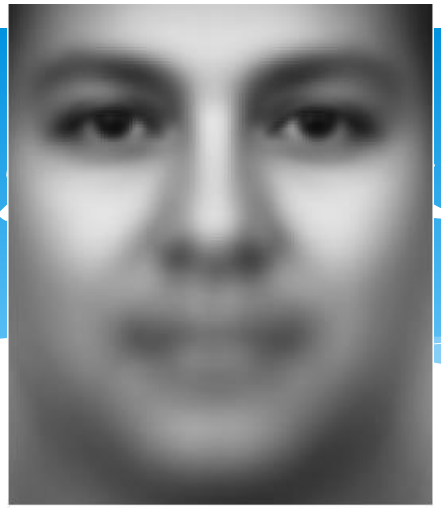
-主成分分析的应用

特征脸(Eigenfaces)#101~#108







-主成分分析的应用

特征脸(Eigenfaces)#501~#508








思考讨论

* 例如人脸识别问题，傅利叶谱是否可以直接作为特征?
* 什么样的特征适合于识别？鲁棒?
* 其他的特征提取或选择方法?

自学内容

* 基于PCA的人脸识别方法

第二次大作业

* 模式识别方法在人脸识别上的应用
* 作业要求
  * 提取不同的人脸特征来进行人脸识别，提取特征不限
  * 人脸识别方法不限
  * 识别率大小不做要求，越高越好

谢谢