机器学习

李成龙

安徽大学人工智能学院
“多模态认知计算”安徽省重点实验室
合肥综合性国家科学中心人工智能研究院

内容安排

- 什么是机器学习
- 机器如何学习
- 如何让机器学习的更好
- 为什么机器能学习

内容安排

- 机器如何学习
  - 有监督学习
    - 感知机
    - 支持向量机
    - 朴素贝叶斯分类
    - 决策树
    - 集成学习（Bagging算法与随机森林、Boosting算法）
    - 线性回归
    - 逻辑回归
    - Softmax回归
    - 神经网络与深度学习
  - 无监督学习
    - 聚类
    - 主成分分析

本节目录

本节目录

主成分分析

- **概述**
  - 主成分分析（principal component analysis, PCA）是一种常用的无监督学习方法
  - PCA方法简称为主分量分析法、主成分分析法或PCA方法，在多元统计分析或经济统计分析领域，亦通常将其称之为因子载荷分析或因子分析
  - 这一方法利用正交变换把由线性相关变量表示的观测数据转换为少数几个由线性无关变量表示的数据，线性无关的变量称为主成分
  - 主成分的个数通常小于原始变量的个数，所以主成分分析属于降维方法
  - 主成分分析主要用于发现数据中的基本结构，即数据中变量之间的关系

主成分分析

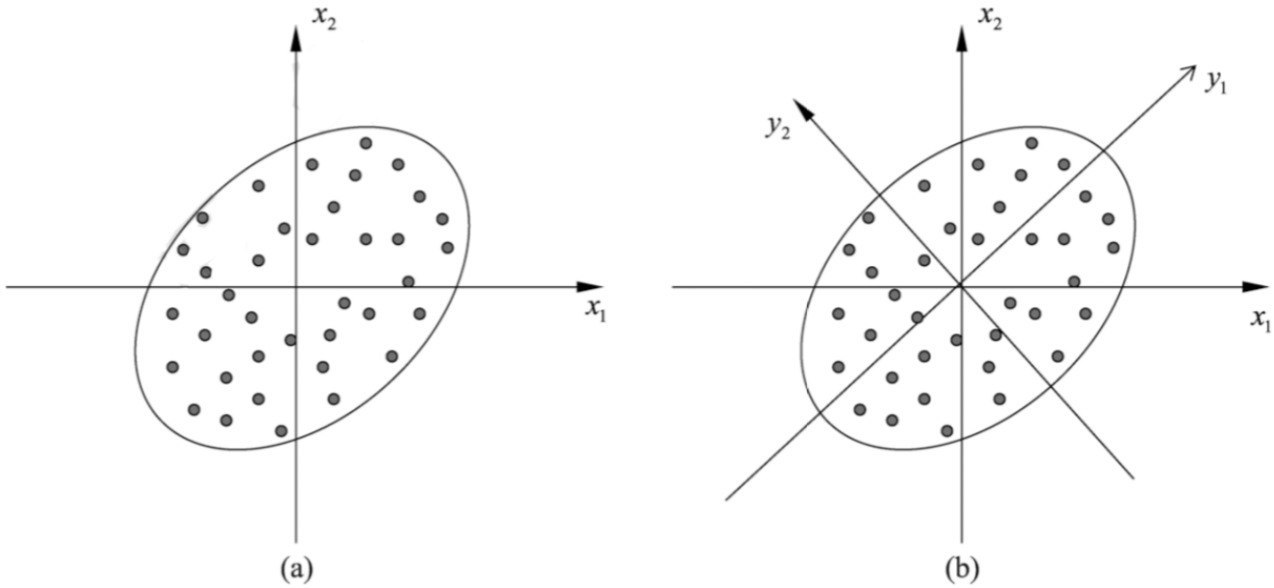
- **基本思想**
  - 主成分分析中，首先对给定数据进行规范化，使得数据每一变量的平均值为0，方差为1
  - 之后对数据进行正交变换，原来由线性相关变量表示的数据，通过正交变换变成由若干个线性无关的新变量表示的数据
  - 新变量是可能的正交变换中变量的方差的和（信息保存）最大的，方差表示在新变量上信息的大小
  - 可以用主分成近似地表示原始数据，发现数据的基本结构
  - 也可以把数据由少数主成分表示，对数据降维

主成分分析

- **基本思想**
  - 数据集合中的样本由实数空间（正交坐标系） 中的点表示，空间的一个坐标轴表示一个变量，规范化处理后得到的数据分布在原点附近
  - 对原坐标系中的数据进行主成分分析等价于进行坐标系旋转变换，将数据投影到新坐标系的坐标轴上
  - 新坐标系的第一坐标轴、第二坐标轴等分别表示第一主成分、 第二主成分等
  - 数据在每一轴上的坐标值的平方表示相应变量的方差
  - 这个坐标系是在所有可能的新的坐标系中，坐标轴上的方差的和最大的

主成分分析

- **例题：**数据由线性相关的两个变量x_1和x_2表示
  - 主成分分析对数据进行旋转变换，并将数据在新坐标系表示
  - 主成分分析选择方差最大的方向（第一主成分）作为新坐标系的第一坐标轴，即y_1轴
  - 之后选择与第一坐标轴正交，且方差次之的方向第二主成分）作为新坐标系的第二坐标轴，即y_2轴



本节目录

主成分分析

- **模型结构**
  - 对于任意一组基向量$\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m\}$，可将数据集$D$中的任意给定的一个样本$X_i$表示为：

$$X_i = \sum_{j=1}^{m} \theta_{ij} \boldsymbol{w}_j$$

其中θ_{ij}为样本X_i的第j个分量在基向量$\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m\}$下的坐标

主成分分析

- **模型结构**
  - 样本数据在不同基向量下的坐标通常会有所不同
  - 例如，对于某个样本数据向量η，通常以标准正交基$(1,0)^T,(0,1)^T$为基向量将其表示为$X=(3,2)^T$，即有：

$$(3,2)^T=3(1,0)^T+2(0,1)^T$$

表示$(3,2)^T$是以$(1,0)^T$和$(0,1)^T$为基向量组成的二维空间中从原点到坐标点$(3,2)$的向量。亦可将其表示为如下形式：

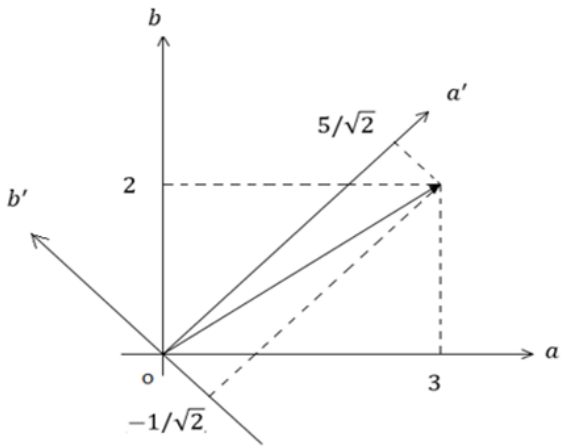
$$X=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}(3,2)^T$$

主成分分析

- **模型结构**
  - 事实上，如果使用标准正交基$(1/\sqrt{2},1/\sqrt{2})^T$和$(-1/\sqrt{2},1/\sqrt{2})^T$作为二维空间的坐标系，则样本数据向量$\eta$在该坐标系中表示形式为：

$$X' = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} (3,2)^T = (5/\sqrt{2}, -1/\sqrt{2})^T$$

  - 向量A的坐标分量取值在这两个坐标系中的变换过程如图所示，其中原坐标系的坐标轴为a轴和b轴，变换后的坐标轴分别为a'轴和b'轴



主成分分析

- **模型结构**
  - 以上分析可知，将样本数据向量η从某个坐标系映射到另一个坐标系，其实并未改变向量自身，而只改变了该向量的定量表示形式
  - 若希望将数据集D中样本数据由m降至k维（$k < m$），则可选择k个线性无关的m维向量$\boldsymbol{w_1, w_2, \ldots, w_k}$作为$k$维空间的一组基，将$D$中$m$维向量通过线性变换映射到以$\boldsymbol{w_1, w_2, \ldots, w_k}$为基向量的$k$维空间中，由此实现对数据集$D$中样本向量的降维

主成分分析

- **模型结构**
  - P令$m \times n$阶矩阵$\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$表示数据集$D$的初始数据，矩阵$\boldsymbol{X'} = (X_1', X_2', \ldots, X_n')$表示数据集$D$在以$\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_k}$为基向量的$k$维空间中样本数据，则有

$$\boldsymbol{X'} = (\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_k})^T \boldsymbol{X} = \begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \cdots \\ \boldsymbol{w}_k \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}^T$$

  - 记矩阵$\boldsymbol{W} = (\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_k})^T$，并称之为变换矩阵，则可将上式简写为：

$$\boldsymbol{X'} = \boldsymbol{WX}$$

本节目录

主成分分析

- **学习准则**
  - 任意选择的一组k个线性无关m维向量$\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k\}$作为基向量构成变换矩阵$\boldsymbol{W}$，将数据集$D$中样本数据降至$k$维。因此，如何选择一组适当的基向量$\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k\}$是实现对样本数据进行有效降维的关键技术
  - 对数据集D中样本数据的降维应尽可能保留原数据有效信息
  - 数据点分布的分散度可用方差度量，方差越大的属性，其包含的信息量就越大，故要求所选基向量$\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k\}$使得映射后数据方差尽可能地变大

本节目录

主成分分析

- **学习算法**
  - 为实现学习准则，需对数据集D中的样本数据进行规范化操作。假定数据集D中包含n个样本数据$X_1, X_2, \cdots, X_n$，每个样本数据X_i均为具有m个属性的m维向量。令第i个样本数据X_i的第j维分量取值为x_{ij}，可按下列公式将X_i的各个分量值x_{ij}转化为标准值z_{ij}

$$z_{ij} = \frac{x_{ij} - u_j}{s_j}, i = 1,2,\cdots,n, j = 1,2,\ldots,m$$

  - 其中u_j和s_j分别是样本数据集D中所有样本的第j维分量均值和标准差，即有

$$u_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\ ,\ \ s_j = \left(\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - u_i)^2\right)^{\frac{1}{2}}$$

主成分分析

- **学习算法**
  - 将规范化后数据组成新的数据矩阵$\boldsymbol{Z} = (Z_1, Z_2, \dots, Z_n)$并构造其协方差矩阵$\boldsymbol{C}$

$$\boldsymbol{C} = \frac{1}{m}\boldsymbol{Z}^T\boldsymbol{Z}$$

$$= \begin{pmatrix} \frac{1}{m}\sum_{j=1}^{m} {z_{1j}}^2 & \frac{1}{m}\sum_{j=1}^{m} z_{1j}\, z_{2j} & \cdots & \frac{1}{m}\sum_{j=1}^{m} z_{1j}\, z_{nj} \\ \frac{1}{m}\sum_{j=1}^{m} z_{1j}\, z_{2j} & \frac{1}{m}\sum_{j=1}^{m} {z_{2j}}^2 & \cdots & \frac{1}{m}\sum_{j=1}^{m} z_{2i}\, z_{nj} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{m}\sum_{j=1}^{m} z_{1j}\, z_{nj} & \frac{1}{m}\sum_{j=1}^{m} z_{2j}\, z_{nj} & \cdots & \frac{1}{m}\sum_{j=1}^{m} {z_{nj}}^2 \end{pmatrix}$$

主成分分析

- **学习算法**
  - 协方差矩阵$\boldsymbol{C}$中主对角线元素为样本属性数据的方差，非主对角线元素表示样本数据的两个属性之间的协方差
  - 现讨论构造适的当标准正交基$\{\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_k}\}$，使得标准化样本数据$\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)$变换到以$\{\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_k}\}$为坐标系的$k$维线性空间中能够得到最大的数据方差。首先构造第一个基向量$\boldsymbol{w_1}$，以标准化样本数据$(Z_1, Z_2, \ldots, Z_n)$关于第一个属性的方差作为目标函数进行最大值优化求解，对如下目标函数进行最大值优化求解

$$J(\boldsymbol{w_1}) = \frac{1}{n}\sum_{i=1}^{n}\left(Z_i^T \boldsymbol{w_1}\right)^2$$

主成分分析

- **学习算法**

$$J(\boldsymbol{w_1}) = \frac{1}{n}\sum_{i=1}^{n}\left(Z_i^T \boldsymbol{w_1}\right)^2$$

  - 由于$Z_i^T \boldsymbol{w_1}$是一个实数，其转置还是其自身，故可将上述目标函数转化为

$$J(\boldsymbol{w_1}) = \frac{1}{n}\sum_{i=1}^{n}\left(Z_i^T \boldsymbol{w_1}\right)^T\left(Z_i^T \boldsymbol{w_1}\right)$$

  - 即有$J(\boldsymbol{w_1}) = \frac{1}{n}\, \boldsymbol{w}_1^T(\sum_{i=1}^{n} Z_i Z_i^T)\boldsymbol{w}_1$
  - 由于$\sum_{i=1}^{n} Z_i Z_i^T = \boldsymbol{Z}\boldsymbol{Z}^T$
  - 故有$J(\boldsymbol{w_1}) = \frac{1}{n} w_1^T \boldsymbol{Z}\boldsymbol{Z}^T w_1$

主成分分析

- **学习算法**
  - 由于$\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k\}$为标准正交基，故需满足约束条件$w_1^T w_1 = 1$。将该约束条件与上述目标函数进行联立，可得到如下条件优化问题

$$\max_{w_1} \ \boldsymbol{w}_1^T Z Z^T \boldsymbol{w}_1; \ \text{s.t.} \ \boldsymbol{w}_1^T \boldsymbol{w}_1 = 1$$

  - 又因为协方差矩阵$\boldsymbol{C} = \frac{1}{n} Z Z^T$，故可由此构造如下拉格朗日函数

$$\max_{\boldsymbol{w}_1} \ \boldsymbol{w}_1^T \boldsymbol{C} \boldsymbol{w}_1 - \alpha(\boldsymbol{w}_1^T \boldsymbol{w}_1 - 1)$$

主成分分析

- **学习算法**
  - 令上述优化问题的目标函数对$\boldsymbol{w}_1$的偏导数为0，则有$2\boldsymbol{C}\boldsymbol{w}_1 - 2\alpha\boldsymbol{w}_1$=0，即有
$$\boldsymbol{C}\boldsymbol{w}_1 = \alpha\boldsymbol{w}_1$$
  - 因此，$\boldsymbol{w}_1$是协方差矩阵$\boldsymbol{C}$的一个特征向量，α是与该特性向量对应的特征根。又因
$$\boldsymbol{w}_1^T\boldsymbol{C}\boldsymbol{w}_1 = \alpha\boldsymbol{w}_1^T\boldsymbol{w}_1 = \alpha$$
  - 故要使得$\boldsymbol{w}_1^T ZZ^T\boldsymbol{w}_1$取得最大化，即使得$\boldsymbol{w}_1^T\boldsymbol{C}\boldsymbol{w}_1$取得最大化，$\boldsymbol{w}_1$即为协方差矩阵$\boldsymbol{C}$的最大特征根$\lambda_1$所对应的特征向量，可由此获得样本数据$\boldsymbol{X}$或$\boldsymbol{Z}$的第一个主成分
$$\boldsymbol{z}_1' = \boldsymbol{w}_1^T\boldsymbol{Z}$$

主成分分析

- **学习算法**
  - 如果第一主成分$\boldsymbol{z}_1'$不足以代表m维数据$\boldsymbol{X}$或$\boldsymbol{Z}$的信息，可以考虑计算样本数据的第二个主成分$\boldsymbol{z}_2' = \boldsymbol{w}_2^T\ \boldsymbol{Z}$。对于第二个主成分$\boldsymbol{z}_2'$的构造，可通过求解如下条件优化问

$$\max_{\boldsymbol{w}_2} \boldsymbol{w}_2^T \boldsymbol{C} \boldsymbol{w}_2 - \alpha(\boldsymbol{w}_2^T \boldsymbol{w}_2 - 1) - \beta(\boldsymbol{w}_2^T \boldsymbol{w}_1 - 0)$$

  - 令上述优化问题的目标函数对$\boldsymbol{w}_2$的偏导数为0，则有

$$2\boldsymbol{C}\boldsymbol{w}_2 - 2\alpha\boldsymbol{w}_2 - \beta\boldsymbol{w}_1 = 0$$

主成分分析

- **学习算法**

$$2\boldsymbol{C}\boldsymbol{w}_2 - 2\alpha\boldsymbol{w}_2 - \beta\boldsymbol{w}_1 = 0$$

  - 用$\boldsymbol{w}_1^T$左乘上式，得：$2\boldsymbol{w}_1^T\boldsymbol{C}\boldsymbol{w}_2 - 2\alpha\boldsymbol{w}_1^T\boldsymbol{w}_2 - \beta\boldsymbol{w}_1^T\boldsymbol{w}_1$=0
  - 由于$\boldsymbol{w}_1^T\boldsymbol{w}_2 = 0$且$\boldsymbol{w}_1^T\boldsymbol{C}\boldsymbol{w}_2$作为标量等于其转置$\boldsymbol{w}_2^T\boldsymbol{C}\boldsymbol{w}_1$，注意到$\boldsymbol{w}_1$为协方差矩阵$\boldsymbol{C}$中以$\lambda_1$为特征根的特征向量，故有

$$\boldsymbol{w}_1^T\boldsymbol{C}\boldsymbol{w}_2 = \boldsymbol{w}_2^T\boldsymbol{C}\boldsymbol{w}_1 = \lambda_1\boldsymbol{w}_2^T\boldsymbol{w}_1 = 0$$

  - 由此可得$\beta\boldsymbol{w}_1^T\boldsymbol{w}_1 = 0$，即有$\beta = 0$，从而可将$2\boldsymbol{C}\boldsymbol{w}_2 - 2\alpha\boldsymbol{w}_2 - \beta\boldsymbol{w}_1$=0简化为$\boldsymbol{C}\boldsymbol{w}_2 = \alpha\boldsymbol{w}_2$
  - 因此，协方差矩阵$\boldsymbol{C}$中除λ_1之外的最大特征值$\lambda_2 = \alpha$所对应的特征向量即为所求的第二个正交基向量$\boldsymbol{w}_2$，由此可得第二主成分$\boldsymbol{z}_2' = \boldsymbol{w}_2^T\boldsymbol{Z}$
  - 可同理依此求出第三、第四、……、第k个基向量和相应的主成分

主成分分析

- **算法流程**
  - 为方便求解，可对协方差矩阵$\boldsymbol{C}$做对角化处理，对协方差矩阵$\boldsymbol{C}$实施对角化的具体计算公式为

$$\boldsymbol{\lambda = PCP^T}$$

  - 其中$\boldsymbol{\lambda}$为$\boldsymbol{C}$的全部特征根组成的对角矩阵，$\boldsymbol{P}$是$\boldsymbol{C}$的全部特性向量组成的正交矩阵
  - 可选择矩阵$\boldsymbol{P}$的前k行组成主分量分析的变换矩阵$\boldsymbol{W}$，由此可得到PCA方法的基本步骤如下：

主成分分析

- **算法流程**
  - 对数据集D中样本数据按如下公式进行标准化 $\boldsymbol{z_{ij}=\frac{x_{ij}-u_j}{s_j}, i=1,2,\cdots,n, j=1,2,\ldots,m}$，并组成新的数据矩阵$\boldsymbol{Z}$
  - 根据数据矩阵$\boldsymbol{Z}$计算协方差矩阵$\boldsymbol{C=\frac{1}{m}Z^TZ}$
  - 求出协方差矩阵$\boldsymbol{C}$全部特征根并将这些特征根按照从大到小次序排列，选择前k个特征值所对应特征向量按行排列构成变换矩阵$\boldsymbol{W}$
  - 使用变换矩阵$\boldsymbol{W}$对原数据进行降维$\boldsymbol{X'=WX}$，或对标准化数据进行降维$\boldsymbol{Z'=WZ}$

本节目录

主成分分析

- **方差贡献率**
  - 若$k = m$，则转换后数据保留了原数据的全部信息；若$k = 0$，则相当于完全不展示原数据的信息
  - 在确定k的具体取值时，通常会考虑不同k值可保留方差的百分比并称这种分量方差占总方差的百分比为该分量对总方差的贡献率，简称为方差贡献率
  - 令$\lambda_1, \lambda_2, \ldots, \lambda_n$表示协方差矩阵$\boldsymbol{C}$的全部特征值且按由大到小顺序排列，$\boldsymbol{w}_i$为特征值$\lambda_i$所对应的特征向量，若保留变换后样本数据前$k$个分量，则得到相应累计方差贡献率$\Omega$为

$$\Omega = \sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{n} \lambda_i$$

  - 通常选择k以保留99%或97%的累计方差贡献率，即选取满足$\Omega \geq 0.99$或$\Omega \geq 0.97$的最小k值

主成分分析

- **例题：**现有我国大陆30个省、直辖市、自治区的经济发展状况数据集如表4-12所示，包括8项经济指标：国民生产总值（$\boldsymbol{a}_1$）；居民消费水平（$\boldsymbol{a}_2$）；固定资产投资（$\boldsymbol{a}_3$）；职工平均工资（$\boldsymbol{a}_4$）；货物周转量（$\boldsymbol{a}_5$）；居民消费指数（$\boldsymbol{a}_6$）；商品零售价格指数（$\boldsymbol{a}_7$）；工业总产值（$\boldsymbol{a}_8$），试用PCA方法将这8项经济指标融合成3项综合指标

主成分分析

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
北京	1394.89	2505	519.01	8144	373.9	117.3	112.6	843.43
天津	920.11	2720	345.46	6501	342.8	115.2	110.6	582.51
河北	2849.52	1258	704.87	4839	2033.3	115.2	115.8	1234.85
山西	1092.48	1250	290.9	4721	717.3	116.9	115.6	697.25
内蒙古	832.88	1387	250.23	4134	781.7	117.5	116.8	419.39
辽宁	2793.37	2397	387.99	4911	1371.1	116.1	114	1840.55
吉林	1129.2	1872	320.45	4430	497.4	115.2	114.2	762.47
黑龙江	2014.53	2334	435.73	4145	824.8	116.1	114.3	1240.37
上海	2462.57	5343	996.48	9279	207.4	118.7	113	1642.95
江苏	5155.25	1926	1434.95	5934	1025.5	115.8	114.3	2026.64
浙江	3524.79	2249	1006.39	6619	754.4	116.6	113.5	916.59
安徽	2003.58	1254	474	4609	908.3	114.8	112.7	824.14
福建	2160.52	2320	553.97	5857	609.3	115.2	114.4	433.67
江西	1205.1	1182	282.84	4211	411.7	116.9	115.9	571.84
山东	5002.34	1527	1229.55	5145	1196.6	117.6	114.2	2207.69
河南	3002.74	1034	670.35	4344	1574.4	116.5	114.9	1367.92
湖北	2391.42	1527	571.68	4685	849	120	116.6	1220.72
湖南	2195.7	1408	422.61	4797	1011.8	119	115.5	843.83
广东	5381.72	2699	1639.83	8250	656.5	114	111.6	1396.35
广西	1606.15	1314	382.59	5150	556	118.4	116.4	554.97
海南	364.17	1814	198.35	5340	232.1	113.5	111.3	64.33
四川	3534	1261	822.54	4645	902.3	118.5	117	1431.81
贵州	630.07	942	150.84	4475	301.1	121.4	117.2	324.72
云南	1206.68	1261	334	5149	310.4	121.3	118.1	716.65
西藏	55.98	1110	17.87	7382	4.2	117.3	114.9	5.57
陕西	1000.03	1208	300.27	4396	500.9	119	117	600.98
甘肃	553.35	1007	114.81	5493	507	119.8	116.5	468.79
青海	165.31	1445	47.76	5753	61.6	118	116.3	105.8
宁夏	169.75	1355	61.98	5079	121.8	117.1	115.3	114.4
新疆	834.57	1469	376.95	5348	339	119.7	116.7	428.76

主成分分析

- 首先将表中数据进行标准化处理，并依据标准化处理后数据建立协方差矩阵，下表为所求的协方差矩阵

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
a_1	1.0000	0.2668	0.9506	0.1899	0.6172	-0.2726	-0.2636	0.8737
a_2	0.2668	1.0000	0.4261	0.7178	-0.1510	-0.2351	-0.5927	0.3631
a_3	0.9506	0.4261	1.0000	0.3989	0.4306	-0.2805	-0.3591	0.7919
a_4	0.1899	0.7178	0.3989	1.0000	-0.3562	-0.1342	-0.5384	0.1033
a_5	0.6172	-0.1510	0.4306	-0.3562	1.0000	-0.2532	0.0217	0.6586
a_6	-0.2726	-0.2351	-0.2805	-0.1342	-0.2532	1.0000	0.7628	0.1252
a_7	-0.2636	-0.5927	-0.3591	-0.5384	0.0217	0.7628	1.0000	-0.1921
a_8	0.8737	0.3631	0.7919	0.1033	0.6586	-0.1252	-0.1921	1.0000

主成分分析

- 然后，对协方差矩阵进行对角化处理，下表为所求特征值并按从大到小的次序排列。下表中最后两列分别是各特征值的方差贡献率及其累计值

编号	特征值	方差贡献率	累计方差贡献率
1	3.754	46.925%	46.925%
2	2.197	27.4625%	74.3875%
3	1.215	15.1875%	89.575%
4	0.403	5.0375%	94.6125%
5	0.213	2.6625%	97.275%
6	0.138	1.725%	99%
7	0.065	0.8125%	99.8125%
8	0.015	0.1875%	100%

主成分分析

- 依题意，选择较大的三个特征值所对应特征向量作为基向量进行降维，所选三个特征值的方差百分比累计值为89.575%。根据特征值计算得到三个基向量的分量取值如下表所示

	w_1	w_2	w_3
w_{i1}	0.45679	0.25851	0.1099
w_{i2}	0.31301	-0.40379	0.24587
w_{i3}	0.47056	0.10839	0.19243
w_{i4}	0.23996	-0.48777	0.33405
w_{i5}	0.2509	0.49801	-0.24933
w_{i6}	-0.26244	0.16988	0.7227
w_{i7}	-0.31966	0.40102	0.39716
w_{i8}	0.42468	0.28769	0.19147

主成分分析

- 根据以上分析计算，故可将8维的原始数据 $\boldsymbol{X}=(\boldsymbol{a}_1,\boldsymbol{a}_2,\boldsymbol{a}_3,\boldsymbol{a}_4,\boldsymbol{a}_5,\boldsymbol{a}_6,\boldsymbol{a}_7,\boldsymbol{a}_8)^T$ 降维为3维的综合指标数据 $\boldsymbol{Z}'=(\boldsymbol{z}_1',\boldsymbol{z}_2',\boldsymbol{z}_3')^T$，其中

$$\boldsymbol{z}_i'=\boldsymbol{w}_i{}^T\boldsymbol{X},\qquad i=1,2,3$$

- 即有：

$$\begin{aligned}\boldsymbol{z}_1' &= 0.4568\boldsymbol{a}_1+0.3130\boldsymbol{a}_2+0.4706\boldsymbol{a}_3+0.2400\boldsymbol{a}_4+0.2509\boldsymbol{a}_5\\ &\quad -0.2624\boldsymbol{a}_6-0.3197\boldsymbol{a}_7+0.4247\boldsymbol{a}_8\\ \boldsymbol{z}_2' &= 0.2585\boldsymbol{a}_1-0.4038\boldsymbol{a}_2+0.1084\boldsymbol{a}_3-0.4878\boldsymbol{a}_4+0.4980\boldsymbol{a}_5\\ &\quad +0.1699\boldsymbol{a}_6+0.4010\boldsymbol{a}_7+0.2877\boldsymbol{a}_8\\ \boldsymbol{z}_3' &= 0.1099\boldsymbol{a}_1+0.2459\boldsymbol{a}_2+0.1924\boldsymbol{a}_3+0.3340\boldsymbol{a}_4-0.2493\boldsymbol{a}_5\\ &\quad +0.7227\boldsymbol{a}_6+0.3972\boldsymbol{a}_7+0.1915\boldsymbol{a}_8\end{aligned}$$

主成分分析

- 带入各省市标准化后的数据，各省份经济数据的主成分如下表

	北京	天津	河北	山西	内蒙古	辽宁	吉林
z_1'	-0.8266	0.6564	1.3585	-0.9888	-1.6211	1.6632	-0.3868
z_2'	2.2582	-2.6378	2.3513	0.3905	0.7253	0.9719	-0.4226
z_3'	0.5399	-1.1725	-1.3128	-0.5717	-0.3819	-0.6231	-1.2106
	河南	湖北	湖南	广东	广西	海南	四川
z_1'	1.023	-0.2825	-0.4101	4.6123	-1.1412	-0.5639	0.5699
z_2'	2.1457	1.4488	1.063	-1.2982	0.3656	-2.2874	1.9764
z_3'	-0.9401	1.1445	0.2546	0.0959	0.3815	-2.4087	0.852
	上海	江苏	浙江	安徽	福建	江西	山东
z_1'	3.1951	3.5689	1.883	0.4451	0.4181	-1.3898	3.0006
z_2'	-3.2802	1.2629	-0.4864	0.1197	-0.9188	0.3006	2.0659
z_3'	2.8822	0.3835	0.2257	-1.862	-0.6569	-0.5293	0.5468
	云南	西藏	陕西	甘肃	青海	宁夏	新疆
z_1'	-2.0197	-2.0175	-1.7772	-2.1163	-2.3478	-2.1619	-1.7232
z_2'	0.7238	-2.0169	0.7078	0.1682	-1.074	-0.9936	-0.0422
z_3'	1.8898	0.0156	0.459	0.6939	0.2626	-0.4881	1.019

主成分分析

$$
\begin{aligned}
z_1' &= 0.4568a_1 + 0.3130a_2 + 0.4706a_3 + 0.2400a_4 + 0.2509a_5 \\
&\quad - 0.2624a_6 - 0.3197a_7 + 0.4247a_8 \\
z_2' &= 0.2585a_1 - 0.4038a_2 + 0.1084a_3 - 0.4878a_4 + 0.4980a_5 \\
&\quad + 0.1699a_6 + 0.4010a_7 + 0.2877a_8 \\
z_3' &= 0.1099a_1 + 0.2459a_2 + 0.1924a_3 + 0.3340a_4 - 0.2493a_5 \\
&\quad + 0.7227a_6 + 0.3972a_7 + 0.1915a_8
\end{aligned}
$$

- 经PCA方法降维后得到的三个数据向量 z_1', z_2', z_3' 均由原始数据的8个指标国民生产总值（a_1）；居民消费水平（a_2）；固定资产投资（a_3）；职工平均工资（a_4）；货物周转量（a_5）；居民消费指数（a_6）；商品零售价格指数（a_7）；工业总产值（a_8）通过线性组合得到，z_1' 的前3个指标 a_1, a_2, a_3 的组合系统均较大，这三个指标分量对 z_1' 的构成起主要作用，故可将 z_1' 看成是由国民生产总值、固定资产投资和居民消费水平所刻画的反映经济发展状况的综合指标
- 可将 z_2' 看成是由国民生产总值，固定资产投资和居民消费水平所刻画的反映经济发展状况的综合指标，将 z_3' 单独看成是居民消费指数指标

总结

- **基本思想**
- **模型结构**
  - 正交变换
- **学习准则**
  - 方差最大化
- **学习算法**
  - 特征值/奇异值分解
- **主成分选择**
  - 方差贡献率

思考题

- PCA中，为什么先对数据进行规范化

练习题

- 现有10组正样本（见positive.mat）和10组负样本（见negative.mat），属性维度为15。试用PCA把数据降维到二维平面上进行分析，分别用不同颜色标识不同的类别