机器学习及其MATLAB实现—从基础到实践 第13课

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

课程目录

主成分分析概述

- 主成分分析（Principle Component Analysis, PCA）
- 用少数的若干新变量（原变量的线性组合）替代原变量，新变量要尽可能多地反映原变量的数据信息，同时，新变量之间相互正交，可以消除原变量中相互重叠的信息。
- 设样本的标准化输入变量矩阵为：

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots\cdots & x_{1k} \\ x_{21} & x_{22} & \cdots\cdots & x_{2k} \\ & & \cdots\cdots & \\ x_{n1} & x_{n2} & \cdots\cdots & x_{nk} \end{bmatrix}$$

- 要求构造一个变量 P_1 满足：$P_1 = Xt_1 \ , \ \|t_1\| = 1$
- 同时，使得变量 P_1 能携带标准化输入变量矩阵 $X_{n\times k}$ 的信息。

主成分分析概述

• 从概率统计观点可知，变量的方差越大，该变量包含的信息越多。因此，上述问题可以转化为要求变量P_1 的方差最大。P_1 的方差为

$$Var(P_1)=\frac{1}{n}\|P_1\|^2=\frac{1}{n}t_1^{'}X^{'}Xt_1=t_1^{'}Vt_1 \qquad V=\frac{1}{n}X^{'}X$$

• 构造拉格朗日函数

$$L=t_1^{'}Vt_1-\lambda_1(t_1^{'}t_1-1)$$

其中，λ_1 为拉格朗日系数。分别计算 L 对 λ_1和 $\boldsymbol{t_1}$的偏导数，并令其为零，则有：

$$\begin{cases}\dfrac{\partial L}{\partial t_1}=2Vt_1-2\lambda_1t_1=0\\ \dfrac{\partial L}{\partial \lambda_1}=-(t_1^{'}t_1-1)=0\end{cases} \qquad Vt_1=\lambda_1t_1$$

由此可知，$\boldsymbol{t_1}$ 是 V 的一个标准化特征向量，λ_1 为其对应的特征值。此时，$Var(P_1)=t_1^{'}Vt_1=t_1^{'}\lambda_1t_1=\lambda_1t_1^{'}t_1=\lambda_1$

主成分分析概述

- 也就是说，所要求的 t_1 是矩阵 V 的最大特征值 λ_1 所对应的标准化特征向量。此时所对应的构造变量 $P_1 = Xt_1$ 称为第一主成分。
- 以此类推，可以求出 X 的第 m 个主成分 $P_m = Xt_m$
- 由上面的分析可知，第一主成分携带的信息最多，第二主成分次之，……。
- 前 m 个主成分携带的信息总和为

$$\sum_{i=1}^{m} Var(P_i) = \sum_{i=1}^{m} \lambda_i$$

主成分回归分析 = 主成分分析 + 多元线性回归分析

偏最小二乘法概述

- 偏最小二乘法（Partial Least Squares, PLS）
- PCA方法提取出的前若干个主成分携带了原输入变量矩阵的大部分信息，消除了相互重叠部分的信息。并没有考虑主成分对输出变量的解释能力，**方差贡献率很小但对输出变量有很强解释能力的主成分将会被忽略掉**，这无疑会对校正模型的性能产生一定的影响。偏最小二乘法（PLS）可以很好地解决这个问题。
- PLS的基本思路是逐步回归，逐步分解输入变量矩阵和输出变量矩阵，并综合考虑提取的主成分对输入变量矩阵和输出变量矩阵的解释能力，直到满足性能要求为止。
- 设标准化的输入变量矩阵和输出变量矩阵分别为

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots\cdots & x_{1k} \\ x_{21} & x_{22} & \cdots\cdots & x_{2k} \\ & & \cdots\cdots & \\ x_{n1} & x_{n2} & \cdots\cdots & x_{nk} \end{bmatrix} \qquad Y=\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}$$

偏最小二乘法概述

- 要求构造变量 $\boldsymbol{t_1}$ 和 u_1 满足：

$$\begin{cases} t_1 = Xw_1 \ , \ \|w_1\| = 1 \\ u_1 = Yc_1 \ , \ \|c_1\| = 1 \end{cases}$$

- 同时，应满足以下三个条件：

① $\boldsymbol{t_1}$ 应尽可能大地携带输入变量矩阵 X 的信息；

② u_1 应尽可能大地携带输出变量矩阵 Y 的信息；

③ $\boldsymbol{t_1}$ 和 u_1 应具有尽可能大的相关程度。

- 由主成分分析原理可知，条件（1）和条件（2）等价于要求 $\boldsymbol{t_1}$ 和 u_1 的方差尽可能地大，即

$$\begin{cases} Var(t_1) \to \max \\ Var(u_1) \to \max \end{cases}$$

偏最小二乘法概述

- 根据典型相关分析，条件（3）可以转换为使得 t_1和 u_1 的相关系数尽可能地大，即

$$r(t_1,u_1)\to \max$$

- 上述问题可以转化为计算 t_1和 u_1 的协方差的最大值，即

$$Cov(t_1,u_1)=t_1^{'}u_1=w_1^{'}X^{'}Yc_1\to \max$$
$$st:\|w_1\|=1,\|c_1\|=1$$

- 构造拉格朗日函数 $L=w_1^{'}X^{'}Yc_1-\lambda_1(w_1^{'}w_1-1)-\lambda_2(c_1^{'}c_1-1)$

$$\begin{cases}\frac{\partial L}{\partial w_1}=X^{'}Yc_1-2\lambda_1w_1=0\\ \frac{\partial L}{\partial c_1}=w_1^{'}X^{'}Y-2\lambda_2c_1=0\\ \frac{\partial L}{\partial \lambda_1}=-(w_1^{'}w_1-1)=0\\ \frac{\partial L}{\partial \lambda_2}=-(c_1^{'}c_1-1)=0\end{cases}$$

$$\theta=2\lambda_1=2\lambda_2=w_1^{'}X^{'}Yc_1=Cov(t_1,u_1)$$

$$\begin{cases}X^{'}Yc_1=\theta w_1\\ Y^{'}Xw_1=\theta c_1\end{cases}\qquad \begin{cases}X^{'}YY^{'}Xw_1=\theta^2w_1\\ Y^{'}XX^{'}Yc_1=\theta^2c_1\end{cases}$$

偏最小二乘法概述

- 可见，w_1 是矩阵$X^{'}YY^{'}X$ 的特征向量，对应的特征值为 θ^2 。
- 使 t_1和 u_1 的协方差最大的 w_1 是对应于矩阵 $X^{'}YY^{'}X$ 最大特征值的单位特征向量。
- 同理，c_1 是矩阵$Y^{'}XX^{'}Y$ 的特征向量，对应的特征值为 θ^2 。
- 使 t_1和 u_1 的协方差最大的 c_1 是对应于矩阵$Y^{'}XX^{'}Y$最大特征值的单位特征向量。

$$\begin{cases} X = t_1 p_1^{'} + X^{*} \\ Y = t_1 r_1^{'} + Y^{*} \end{cases}$$

- 若回归方程的精度已经满足要求，则停止；否则，利用残差矩阵 X^{*} 和 Y^{*}，计算第二主成分，并重新建立回归方程。以此类推，直到回归方程的精度满足要求为止。

案例分析

主成分回归分析

偏最小二乘法

重点函数解读

- **princomp**

  COEFF = princomp(X) performs principal components analysis (PCA) on the n-by-p data matrix X, and returns the principal component coefficients, also known as loadings.

- **regress**

  b = regress(y,X) returns a p-by-1 vector b of coefficient estimates for a multilinear regression of the responses in y on the predictors in X.

- **plsregress**

  [XL,YL] = plsregress(X,Y,ncomp) computes a partial least-squares (PLS) regression of Y on X, using ncomp PLS components, and returns the predictor and response loadings in XL and YL, respectively.

特征选择方法概述

- **Filter vs. Wrapper**
  - ✓ Filter无需利用学习模型的性能，即可进行特征选择，主要依赖一些评价准则，如：相关系数、互信息、信息熵等
  - ✓ Wrapper需要建立学习模型，通过模型的性能进行评价特征的优劣。
- **搜索法**
  - ✓ 随机搜索
  - ✓ 启发式搜索
- **正则化方法（L1范数、LASSO）**
- **…**

遗传算法特征选择概述

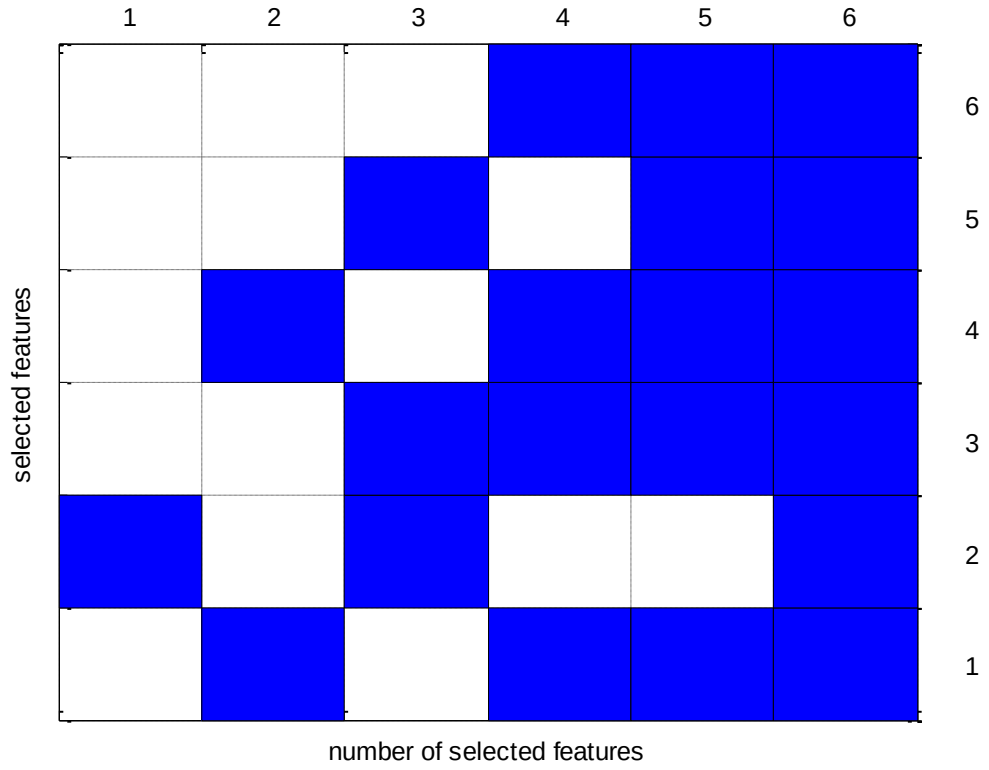
- 将N个输入变量用一个长度为N的染色体表示，染色体的每一位代表一个输入变量。
- 每一位的基因取值只能是“1”和“0”两种情况。
- 如果染色体的某一位值为“1”，表示该位对应的输入变量被选中，参与模型建立。
- 反之，如果染色体的某一位值为“0”，则表示对应的输入变量未被选中，不参与模型建立。

前向选择法vs.后向选择法概述

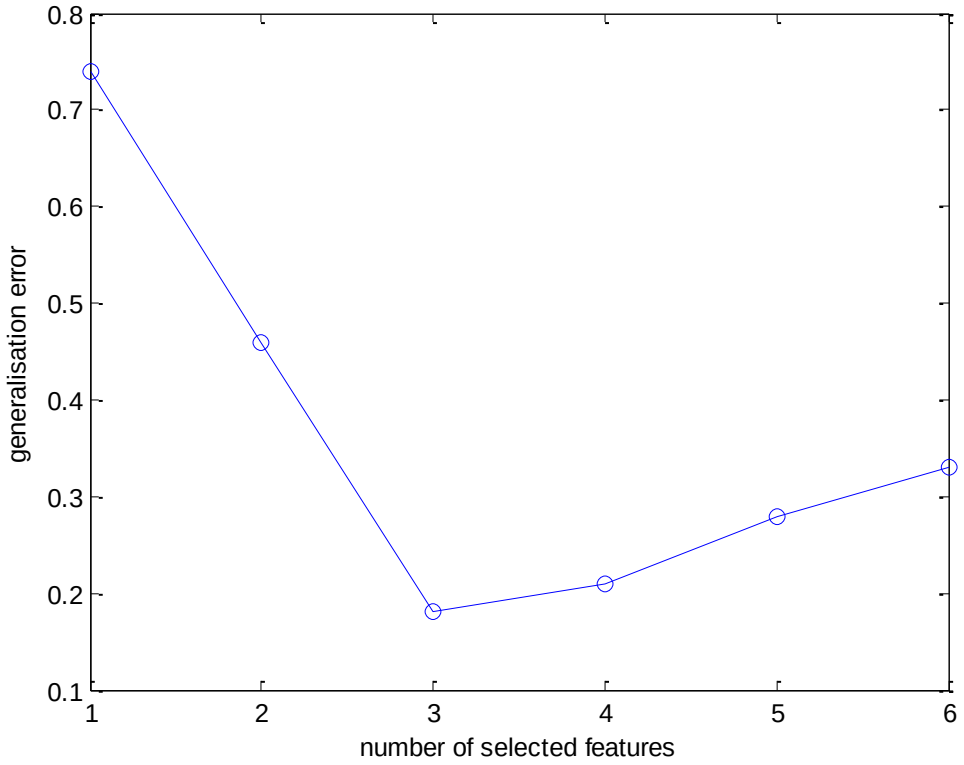
- 前向选择法
  - ✓ 自下而上的选择方法，又称集合增加法
  - ✓ 特征集合初始化为一个空集
  - ✓ 每次向特征集合中添加一个输入变量，当新加入的变量致使模型性能更优时，则保留该输入变量，否则不保留。
- 后向选择法
  - ✓ 自上而下的选择方法，又称集合缩减法
  - ✓ 特征集合初始化为全部的输入变量
  - ✓ 每次从特征集合中剔除一个输入变量，当剔除后致使模型性能更优时，则剔除该输入变量，否则保留该输入变量。
- 广义方法
  - ✓ 一次增加或剔除多个变量
  - ✓ 区间法

FSP Plot and SET Curve

Feature Selection Path Plot



Sparsity-Error Trade-off Curve



炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间