Principal component method. Preprocessing

Let's use the matrix representation of the regression model $\boldsymbol{X} \cdot \vec{\boldsymbol{w}} = \vec{\boldsymbol{Y}}$

(1)

$$\boldsymbol{X} = [\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k] = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \vec{\boldsymbol{w}} = [w_1, w_2, \dots, w_k]$$

Here we use the representation of the experiment matrix $\boldsymbol{X}$ in the form of column vectors, each column is characteristic, a descriptor of a multidimensional object, and each *i*-th row vector is the result of *i*-th experiment, which is described by the k-dimensional vector of properties $\vec{\boldsymbol{x}}^{(\boldsymbol{i})}$, observed in the experiment:

$$\vec{\boldsymbol{x}}^{(\boldsymbol{i})} = [x_{i1}, x_{i2}, \dots, x_{ik}]$$

Vector $\vec{\boldsymbol{Y}}$ is the response of the system and the basis for further training. Typically, these systems are modified to simplify; there are two main forms of preprocessing: standardization (in the case of heterogeneous features) and centering. Mainly the characteristics of the model have different nature (different units of measurements or orders and so on) this defines difficulties in finding solutions to such systems; they are removed by standardization of features.

<u>Standardization → Model 1</u>:

So, getting rid of dimensional quantities, standardize the experimental data (1). For convenience of formalization we will use Matlab notation and introduce *vector* $\vec{\mathbf{m}} = \text{mean}(\boldsymbol{X})$ – of averages and $\vec{\boldsymbol{s}} = \text{std}(\boldsymbol{X}) -$ standard deviations (SD); values of SD are calculated for each column of the matrix $\boldsymbol{X}$; whereas $\bar{\text{y}} = \text{mean}(\vec{\boldsymbol{Y}})$, $S = \text{std}(\vec{\boldsymbol{Y}})$ are scalars. System (1) after standardization takes the form

$$\widetilde{\boldsymbol{X}} \cdot \widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{Y}} \qquad (2)$$

$$\widetilde{\boldsymbol{X}} = \left(\tilde{x}_{ij}\right),\ \tilde{x}_{ij} = (x_{ij} - m_j)/s_j; \tag{3}$$

$$\widetilde{\boldsymbol{Y}} = (\tilde{y}_i),\ \tilde{y}_i = (y_i - \bar{y})/s;\ \ i = 1,..,n;\ j = 1,..k.$$

By the values of the average and SD dimensional data can be easily restored. Note that the system (2) does not have an exact solution; since it is overridden, the solution exists (in the absence of collinearity, the linear dependence of the strings in $\widetilde{\boldsymbol{X}}$) only as optimal, having a minimum value of the sum squares error:

$$\left(\widetilde{\boldsymbol{X}} \cdot \widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{Y}}\right)^{\boldsymbol{T}} (\widetilde{\boldsymbol{X}} \cdot \widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{Y}}) \rightarrow \min$$

The solution (2) is found by multiplying both parts of the system on the left by $\widetilde{\boldsymbol{X}}^T$:

$$\widetilde{\boldsymbol{X}}^T \cdot \widetilde{\boldsymbol{X}} \cdot \widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{Y}}$$

$$\widetilde{\boldsymbol{w}}^* = (\widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{X}})^{-1} \cdot \widetilde{\boldsymbol{X}}^T \cdot \widetilde{\boldsymbol{Y}} \tag{4}$$

As a result, we obtain solution vector $\widetilde{\boldsymbol{w}}^* = [\widetilde{w}_1{}^*,..,\widetilde{w}_\mathrm{k}{}^*]$; note that $\widetilde{\boldsymbol{w}}^*$ takes part in the prediction. Matrix

$$\boldsymbol{R} = \frac{1}{(\mathrm{n}-1)} (\widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{X}})_{k \times k} \tag{5}$$

is a correlation matrix of standardized k features of $\widetilde{\boldsymbol{X}}$, it is symmetric and positive definite.

<u>Centering→ Model 2:</u>

The second approach of preprocessing is data centering, it is used only in the case of the similar properties nature and is reduced to subtracting its average from each feature:

$$\widehat{\boldsymbol{X}} \cdot \widehat{\boldsymbol{w}} = \widehat{\boldsymbol{Y}} \tag{6}$$

$$\widehat{\boldsymbol{X}} = \left(\hat{x}_{ij}\right),\ \hat{x}_{ij} = \hat{y}_{ij} - \hat{m}_j; \tag{7}$$

$$\widehat{\boldsymbol{Y}} = (\hat{y}_i),\ \hat{y}_i = \hat{y}_i - \bar{\hat{y}};\ i = 1,..,n;\ j = 1,..k.$$

The solution of the system (2) in this case is similar to (4)

$$\widehat{\boldsymbol{w}}^* = (\widehat{\boldsymbol{X}}^T\widehat{\boldsymbol{X}})^{-1} \cdot \widehat{\boldsymbol{X}}^T \cdot \widehat{\boldsymbol{Y}} \tag{8}$$

Covariance matrix $\boldsymbol{C}$ is

$$\boldsymbol{C} = \frac{1}{(\mathrm{n}-1)}\widehat{\boldsymbol{X}}^T\widehat{\boldsymbol{X}} \tag{9}$$

Matrix $\boldsymbol{C}$ is symmetric and positive definite.

Principal components

Projection of the feature space onto the principal components

Based on *model 1* consider the matrix $\boldsymbol{R}$ it, by virtue of symmetry and positive certainty, can be factorized as follows:

$$\boldsymbol{R}_{k\times k} = \boldsymbol{V}_{k\times k} \cdot \boldsymbol{D}_{k\times k} \cdot \boldsymbol{V}^T_{k\times k} \tag{10}$$

Here $k \times k$ indicate the dimension of the matrices.

Matrix $\boldsymbol{V} = [\overline{\boldsymbol{v}}_1, \ldots, \overline{\boldsymbol{v}}_k]$, consists of $\overline{\boldsymbol{v}}_i$ – eigenvectors as columns of matrix $\boldsymbol{R}$ and is an orthogonal matrix ($\boldsymbol{V}^T = \boldsymbol{V}^{-1}$), and $\boldsymbol{D}$ – diagonal matrix such that $d_{ii} > 0$, $d_{ii} = \lambda_i$ – eigenvalues of matrix $\boldsymbol{R}$.

Given the type of matrix (5), we transform to (10) in this way

$$(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}) = \widetilde{\boldsymbol{V}}\boldsymbol{D}\widetilde{\boldsymbol{V}}^T \tag{11}$$

$$\boldsymbol{V}^{\mathrm{T}}\widetilde{\boldsymbol{X}}^{\mathrm{T}}\widetilde{\boldsymbol{X}}\boldsymbol{V} = (\widetilde{\boldsymbol{X}}\boldsymbol{V})^{\mathrm{T}}\widetilde{\boldsymbol{X}}\boldsymbol{V} \equiv \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} = \boldsymbol{D}, \quad (\boldsymbol{Z} = \widetilde{\boldsymbol{X}}\boldsymbol{V}) \tag{12}$$

This means that we have made an orthogonal transformation of the data $\widetilde{\boldsymbol{X}}$ in such a way that $\boldsymbol{D} = \widetilde{\boldsymbol{X}}\boldsymbol{V}$, and the new variables $\boldsymbol{Z}$ have a diagonal correlation matrix:

$$\boldsymbol{R_z} = \frac{1}{\mathrm{n-k}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}, \tag{13}$$

the diagonal nature of which proves the orthogonality of the features of the system in some new coordinates, i.e. $\left(\vec{\boldsymbol{Z}}^{(i)}, \vec{\boldsymbol{Z}}^{(j)}\right) = 0,\ i \neq j$.

Based on *model 2* consider the covariance matrix C, it can also be factorized similarly to (10):

$$\boldsymbol{C}_{k\times k} = \widehat{\boldsymbol{V}}_{k\times k} \cdot \widehat{\boldsymbol{D}}_{k\times k} \cdot \widehat{\boldsymbol{V}}^{T}_{k\times k}$$

Note that the matrices $\widehat{\boldsymbol{V}}$, $\widehat{\boldsymbol{D}}$, as in (10), respectively, are orthogonal and diagonal and consist of eigenvectors and eigenvalues of the matrix $\boldsymbol{C}$, while the eigenvalues coincide with the variances of the system features after projection onto the main components, i.e. $\hat{d}_{ii} = \hat{\lambda}_i = \hat{\sigma}_i^2$.

Projection rule which can discover in (12) $\boldsymbol{Z} = \widetilde{\boldsymbol{X}}\boldsymbol{V}$ for the second model transform to $\widehat{\boldsymbol{Z}} = \widehat{\boldsymbol{X}}\widehat{\boldsymbol{V}}$ centered matrices, similarly, the source system is converted to the form

$$\widehat{\boldsymbol{Z}} \cdot \widehat{\boldsymbol{w}} = \widehat{\boldsymbol{Y}} \tag{14}$$

and its solution is determined by least squares method

$$\widehat{\boldsymbol{w}}^{*} = (\widehat{\boldsymbol{Z}}^{T}\widehat{\boldsymbol{Z}})^{-1} \cdot \widehat{\boldsymbol{Z}}^{T} \cdot \widehat{\boldsymbol{Y}} \tag{15}$$

The matrix $\boldsymbol{C}_z$ is a covariancematrix of feature projected onto the main components.

$$\boldsymbol{C}_z = \frac{1}{\mathrm{n-k}}\widehat{\boldsymbol{Z}}^{T}\widehat{\boldsymbol{Z}}.$$

Diagonal form of $\boldsymbol{C}_z$ determines the orthogonality of the new variables-characteristics $\left(\vec{\boldsymbol{Z}}^{(i)}, \vec{\boldsymbol{Z}}^{(j)}\right) = 0$; note that the diagonal elements are positive eigenvalues of the covariance matrix $\boldsymbol{C}$ and the validity of the equation

$$\left(\vec{\boldsymbol{Z}}^{(i)}, \vec{\boldsymbol{Z}}^{(i)}\right) = \sigma_{ii}{}^2$$

It follows that the direction of the main axes coincides with the directions of the greatest variances of the system features.

Returning to the system in the principle components obtained on the basis of a standardized system, we have

$$\widetilde{\boldsymbol{Z}} \cdot \widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{Y}} \tag{16}$$

Get the solution similarly (15)

$$\widetilde{\boldsymbol{w}}^* = (\widetilde{\boldsymbol{Z}}^T \widetilde{\boldsymbol{Z}})^{-1} \cdot \widetilde{\boldsymbol{Z}}^T \cdot \widetilde{\boldsymbol{Y}}$$

Dimension reduction

On the one hand, we have the linear regression problem, using response in form of nominal class name or class number, or the linear classification problem, the problem of clustering and on the other hand – regression scaling. Identifying the immanent properties of the set – to unite into groups-clusters at this stage, the response is not involved in the process. Identification of clusters allows to obtain adjusted expert estimates of the objects belonging to a particular class (cluster). The most effective classification tool is the human eye, we recognize, classify, without difficulties, so the decrease in dimension is due, among other things, to the desire to consider and evaluate the cluster (or not) nature of the studied system.

We can estimate the prerequisites for clustering a set of points in 2D and 3D representations.

Choose from the available set of orthogonal $\boldsymbol{Z}$ features, such p, which provides an acceptable (close to one) coefficient of informativeness I_p.

We order the eigenvalues in descending order $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ and consider the coefficient of informativeness

$$I_p = \frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{k} \lambda_i} = \frac{\sum_{i=1}^{p} \sigma_{ii}{}^2}{\sum_{i=1}^{k} \sigma_{ii}{}^2};$$

The contribution of each feature to the system is determined by its variance, which is equal in magnitude to its eigenvalues. In case p=k $I_p = 1$.

Our goal is to select p in such a way as to exclude from the system the features that practically do not affect the variance explained by the model and leave the coefficient of informativeness close to one. This is the goal of reducing the size of the system to save computing resources.

The second goal is to evaluate multidimensional objects in terms of the linear separability of clusters that are inherent in this experimental set. Such a preliminary visual assessment will determine the strategy for further research: whether to use linear classification or clustering, or to use nonlinear approaches based, for example, on neural network technologies (Kohonen networks or multilayer perceptron with reverse error propagation).

Consider the case of p=2, p=3 and graphically construct the proposed system to estimate the prerequisites of linear classification (clustering) of sets.

The algorithm of the task.

(on the classification of irises based on PCA)

1. *Centering system (6) or standardizing case*
2. *Find the orthogonal transformation (projection rule) based on the covariance matrix (9)*
3. *Obtain a new matrix of the system of orthogonal features* $\widehat{\boldsymbol{Z}} = \widehat{\boldsymbol{X}}\widehat{\boldsymbol{V}}$ (or $\boldsymbol{Z} = \widetilde{\boldsymbol{X}}\boldsymbol{V}$)
4. *Estimate system losses in 2D and 3D cases (1-*I_p*)*
5. *Based on the assessment of losses to reduce the dimensionality*
6. *On the part of the set (the first 130 cases) to train the system and obtain* $\widehat{\boldsymbol{w}}^*$ (or $\widetilde{\boldsymbol{w}}^*$)
7. *Identify (predict) other objects (last twenty).*
8. *Need to compare the expert evaluation and the evaluation you received. Estimate the percentage of incorrectly classified objects. Graphically show objects with non-matching classification.*