Computational Intelligence: Methods and Applications

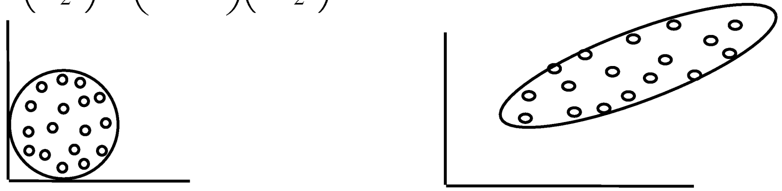
Lecture 6
Principal Component Analysis.

Włodzisław Duch
SCE, NTU, Singapore
Google: Duch

Linear transformations – example

2D vectors **X** uniformly distributed in a unit circle with mean (1,1);
Y = **A**∗**X**, **A** = 2x2 matrix

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$



The shape is elongated, rotated and the mean is shifted.

Invariant distances

Euclidean distance is not invariant to general linear transformations

$$\mathbf{Y} = \mathbf{A} \cdot \mathbf{X}$$

$$\left\| \mathbf{Y}^{(1)} - \mathbf{Y}^{(2)} \right\|^2 = \left(\mathbf{Y}^{(1)} - \mathbf{Y}^{(2)} \right)^{\mathrm{T}} \left(\mathbf{Y}^{(1)} - \mathbf{Y}^{(2)} \right)$$
$$= \left(\mathbf{X}^{(1)} - \mathbf{X}^{(2)} \right)^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{A} \left(\mathbf{X}^{(1)} - \mathbf{X}^{(2)} \right)$$

This is invariant only for orthonormal matrices $A^T A = I$
that make rigid rotations, without stretching or shrinking distances.

Idea: standardize the data in some way to create invariant distances.

Data standardization

For each vector component $\mathbf{X}^{(j)\mathrm{T}}=(X_1^{(j)}, \ldots X_d^{(j)}), j=1 .. n$
calculate mean and std: n – number of vectors, d – their dimension

$$\bar{X}_i = \frac{1}{n} \sum_{j=1}^{n} X_i^{(j)}; \quad \bar{\mathbf{X}} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{X}^{(j)}$$

Vector of mean feature values.

	$\mathbf{X}^{(1)}$	$\mathbf{X}^{(2)}$	$\cdots$	$\mathbf{X}^{(n)}$
$\bar{X}_1$	$X_1^{(1)}$	$X_1^{(2)}$	$\cdots$	$X_1^{(n)}$
$\bar{X}_2$	$X_2^{(1)}$	$X_2^{(2)}$	$\cdots$	$X_2^{(n)}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$\bar{X}_d$	$X_d^{(1)}$	$X_d^{(2)}$	$\cdots$	$X_d^{(n)}$

Averages over rows.

Standard deviation

Calculate standard deviation:

$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_i^{(j)}$$

Vector of mean feature values.

$$\sigma_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_i^{(j)} - \bar{X}_i\right)^2$$

Variance = square of standard deviation (std), sum of all deviations from the mean value.

Transform X => Z, standardized data vectors

$$Z_i^{(j)} = \left(X_i^{(j)} - \bar{X}_i\right)\Big/\sigma_i$$

Std data

Std data: zero mean and unit variance.

$$\bar{Z}_i = \frac{1}{n}\sum_{j=1}^{n} Z_i^{(j)} = \frac{1}{n}\sum_{j=1}^{n}\left(X_i^{(j)} - \bar{X}_i\right)\Big/\sigma_i = 0$$

$$\sigma_{Z,i}^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(Z_i^{(j)} - \bar{Z}_i\right)^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_i^{(j)} - \bar{X}_i\right)^2\Big/\sigma_i^2 = 1$$

Standardize data after making data transformation.

Effect: data is invariant to scaling only (diagonal transformation).

Distances are invariant, data distribution is the same.

How to make data invariant to any linear transformations?

Data standardization example

In slide 2 example Y=AX, assume all X means =1 and variances = 1

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

Transformation

$$\bar{\mathbf{X}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \bar{\mathbf{Y}} = \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Vector of mean feature values.

$$\boldsymbol{\sigma}_X^2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \boldsymbol{\sigma}_{\mathrm{Y}}^2 = \begin{bmatrix} 5 \\ 2 \end{bmatrix} = \mathrm{Diag}\left(\mathbf{A}\mathbf{A}^{\mathrm{T}}\right)$$

Variance check it!

$$\left\|\mathbf{Y}^{(1)} - \mathbf{Y}^{(2)}\right\|^2 = \left(\mathbf{X}^{(1)} - \mathbf{X}^{(2)}\right)^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\left(\mathbf{X}^{(1)} - \mathbf{X}^{(2)}\right)$$

How to make this invariant?

Covariance matrix

Variance (spread around mean value) + correlation between features.

$$C_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_i^{(k)} - \bar{X}_i\right)\left(X_j^{(k)} - \bar{X}_j\right); \quad i,j = 1\cdots d$$

C_X is d x d

$$\mathbf{C}_{\mathbf{X}} = \frac{1}{n-1}\sum_{k=1}^{n}\left(\mathbf{X}^{(k)} - \bar{\mathbf{X}}\right)\left(\mathbf{X}^{(k)} - \bar{\mathbf{X}}\right)^{\mathrm{T}} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^{\mathrm{T}}$$

where **X** is d x n dimensional matrix of vectors shifted to their means.

Covariance matrix is symmetric $C_{ij} = C_{ji}$ and positive definite.

Diagonal elements are variances (square of std), $\sigma_i^2 = C_{ii}$

Pearson correlation coefficient $r_{ij} = C_{ij}\big/\sigma_i\sigma_j \in [-1,+1]$

Spherical distribution of data has C_{ij}=I (unit matrix).

Elongated ellipsoids: large off-diagonal elements, strong correlations between features.

Mahalanobis distance

Linear combinations of features leads to rotations and scaling of data.

$$\mathbf{Y}=\mathbf{AX};\quad \bar{\mathbf{Y}}=\mathbf{A}\bar{\mathbf{X}};\quad \mathbf{C}_Y=\mathbf{A}\mathbf{C}_X\mathbf{A}^{\mathrm{T}}$$

Mahalanobis distance defined as: $\left\|\mathbf{X}\right\|^2_{C_X}=\mathbf{X}^{\mathrm{T}}\mathbf{C}_X^{-1}\mathbf{X}$

is invariant to linear transformations:

$$\begin{aligned}\left\|\mathbf{Y}^{(1)}-\mathbf{Y}^{(2)}\right\|^2_{C_Y}&=\left(\mathbf{Y}^{(1)}-\mathbf{Y}^{(2)}\right)^{\mathrm{T}}\mathbf{C}_Y^{-1}\left(\mathbf{Y}^{(1)}-\mathbf{Y}^{(2)}\right)\\&=\left(\mathbf{X}^{(1)}-\mathbf{X}^{(2)}\right)^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\left[\left(\mathbf{A}^{\mathrm{T}}\right)^{-1}\mathbf{C}_X^{-1}\mathbf{A}^{-1}\right]\mathbf{A}\left(\mathbf{X}^{(1)}-\mathbf{X}^{(2)}\right)\\&=\left\|\mathbf{X}^{(1)}-\mathbf{X}^{(2)}\right\|^2_{C_X}\end{aligned}$$

Principal components

How to avoid correlated features?

Correlations ⇔ covariance matrix is non-diagonal !

Solution: diagonalize it, then use the transformation that makes it diagonal to de-correlate features.

$$\mathbf{Y}=\mathbf{Z}^{\mathrm{T}}\mathbf{X};\quad \mathbf{C}_{\mathbf{X}}\mathbf{Z}^{(i)}=\lambda_i\mathbf{Z}^{(i)};\quad \mathbf{C}_{\mathbf{X}}\mathbf{Z}=\mathbf{Z}\mathbf{\Lambda}$$

$$\mathbf{C}_{\mathbf{Y}}=\mathbf{Z}^{\mathrm{T}}\mathbf{C}_{\mathbf{X}}\mathbf{Z}=\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\mathbf{\Lambda}=\mathbf{\Lambda}$$

In matrix form,
X, Y are d×n,
Z, C_X, C_Y are d×d

C – symmetric, positive definite matrix $X^{\mathrm{T}}CX > 0$ for $||X||>0$;

its eigenvectors are orthonormal: $\mathbf{Z}^{(i)\mathrm{T}}\cdot\mathbf{Z}^{(j)}=\delta_{ij}$

its eigenvalues are all non-negative $\lambda_i \geq 0$

Z – matrix of orthonormal eigenvectors (because C_X is real+symmetric), transforms X into Y, with diagonal C_Y, i.e. decorrelated.

Matrix form

Eigenproblem for C matrix in matrix form: $\mathbf{C}_X\mathbf{Z}=\mathbf{Z}\mathbf{\Lambda}$

$$\begin{pmatrix}C_{11}&C_{12}&\cdots&C_{1d}\\C_{21}&C_{22}&\cdots&C_{2d}\\\vdots&\cdots&\cdots&\vdots\\C_{d1}&C_{d2}&\cdots&C_{dd}\end{pmatrix}\begin{pmatrix}Z_{11}&Z_{12}&\cdots&Z_{1d}\\Z_{21}&Z_{22}&\cdots&Z_{2d}\\\vdots&\cdots&\cdots&\vdots\\Z_{d1}&Z_{d2}&\cdots&Z_{dd}\end{pmatrix}=$$

$$\begin{pmatrix}Z_{11}&Z_{12}&\cdots&Z_{1d}\\Z_{21}&Z_{22}&\cdots&Z_{2d}\\\vdots&\cdots&\cdots&\vdots\\Z_{d1}&Z_{d2}&\cdots&Z_{dd}\end{pmatrix}\begin{pmatrix}\lambda_1&0&\cdots&0\\0&\lambda_2&\cdots&0\\\vdots&\cdots&\cdots&\vdots\\0&0&\cdots&\lambda_d\end{pmatrix}$$

Principal components

PCA: old idea, C. Pearson (1901), H. Hotelling 1933

$$\mathbf{Y}=\mathbf{Z}^{\mathrm{T}}\mathbf{X};$$

Y – principal components, or vectors X transformed using eigenvectors of C_X

$$\mathbf{C}_Y=\mathbf{Z}^{\mathrm{T}}\mathbf{C}_X\mathbf{Z}=\mathbf{\Lambda}$$

Covariance matrix of transformed vectors is diagonal => ellipsoidal distribution of data.

Result: PC are linear combinations of all features, providing new uncorrelated features, with diagonal covariance matrix = eigenvalues.

Small λ_i ⇔ small variance ⇔ data change little in direction Y_i

PCA minimizes **C** matrix reconstruction errors:

Z_i vectors for large λ_i are sufficient to get: $\mathbf{Z}\mathbf{\Lambda}\mathbf{Z}^{\mathrm{T}}=\mathbf{C}_X$

because vectors for small eigenvalues will have very small contribution to the covariance matrix.

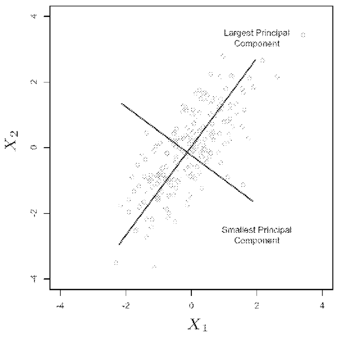
Two components for visualization

Diagonalization methods: see Numerical Recipes, www.nr.com

New coordinate system:
axis ordered according to variance
= size of the eigenvalue.

First *k* dimensions account for

$$V_k = \sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{d} \lambda_i$$



fraction of all variance (please note that λ_i are variances);
frequently 80-90% is sufficient for rough description.

PCA properties

PC Analysis (PCA) may be achieved by:

- transformation making covariance matrix diagonal
- projecting the data on a line for which the sums of squares of distances from original points to projections is minimal.
- orthogonal transformation to new variables that have stationary variances $\sigma_Y(W)$ – around max. variance change is minimal.

True covariance matrices are usually not known, they have to be estimated from data.

This works well on single-cluster data;

more complex structure may require local PCA: the PCA transformation should then be done separately for each cluster or neighborhood of a query vector X.

Some remarks on PCA

PC results obviously depend on the initial scaling of the features, therefore one should standardize the data first to make it independent of scaling or measurement units. Example: Heart data.

Assume that the data matrix X has been standardized, show that:

$$\overline{Y_i} = 0 \quad \sigma^2\left(Y_i\right) = \lambda_i$$

that is the mean stays as zero and the variance of principal components is equal to the eigenvalues. Therefore rejecting Y_i components with small variance leads to small errors in reconstruction of $\mathbf{X} = \mathbf{ZY}$, where rejected components are replaced by zero values.

PC is useful for:

finding new, more informative, uncorrelated features;

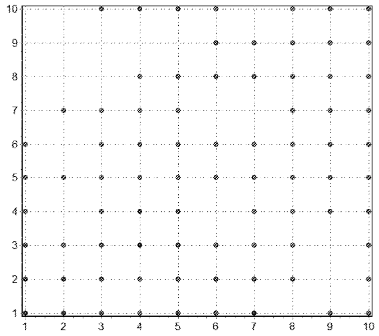
reducing dimensionality: reject low variance features,

reconstructing original data from lower-dimensional projections.

PCA Wisconsin example

Wisconsin Breast Cancer data:

- Collected at the University of Wisconsin Hospitals, USA.
- 699 cases, 458 (65.5%) benign (red), 241 malignant (green).
- 9 features: quantized 1, 2 .. 10, cell properties, ex:
  Clump Thickness, Uniformity of Cell Size, Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses.



2D scatterograms do not show any structure no matter which subspaces are taken!

Example cont.

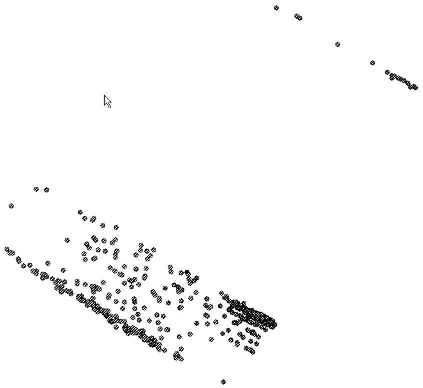
PC gives useful information already in 2D.

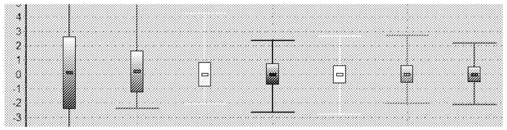
Taking first PCA component of the standardized data:

If (Y_1>0.41) then benign
else malignant

18 errors/699 cases = 97.4%

Transformed vectors are not standardized, std's are below.





Eigenvalues decrease to zero slowly, but classes are well separated.

PCA disadvantages

Useful for dimensionality reduction but:

- Largest variance determines which components are used, but does not guarantee interesting viewpoint for clustering data.
- The meaning of features is lost when linear combinations are formed.

Analysis of coefficients in Z_1 and other important eigenvectors may show which original features are given much weight.

PCA may be also done in an efficient way by performing singular value decomposition of the standardized data matrix.

PCA is also called Karhuen-Loève transformation.

Many variants of PCA are described in A. Webb, Statistical pattern recognition, J. Wiley 2002.

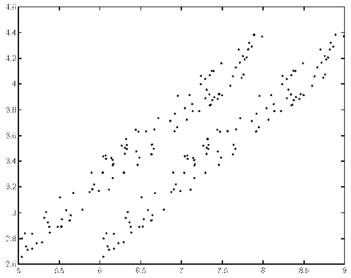
2 skewed distributions

PCA transformation for 2D data:

First component will be chosen along the largest variance line, both clusters will strongly overlap, no interesting structure will be visible.

In fact projection to orthogonal axis to the first PCA component has much more discriminating power.

Discriminant coordinates should be used to reveal class structure.


4.6
4.4
4.2
4
3.8
3.6
3.4
3.2
3
2.8
2.6
5
5.5
6
6.5
7
7.5
8
8.5
9