Università degli Studi di Milano
Master Degree in Computer Science

Information Management course

Teacher: Alberto Ceselli

Lecture 06: 24/10/2012

Data Mining: Methods and Models

— Chapter 1 —

Daniel T. Larose

Data (Dimension) Reduction

- In large datasets it is unlikely that all attributes are independent: multicollinearity
- Worse mining quality:
  - Instability in multiple regression (significant overall, but poor wrt significant attributes)
  - Overemphasize particular attributes (multiple counts)
  - Violates principle of parsimony (too many unnecessary predictors in a relation with a response var)
- Curse of dimensionality:
  - Sample size needed to fit a multivariate function grows exponentially with number of attributes
  - e.g. in 1-dimensional distrib. 68% of normally distributed values lie between -1 and 1; in 10-dimensional distrib. only 0.02% within the radius 1 hypersphere

Principal Component Analysis (PCA)

- Try to explain correlation using a small set of linear combination of attributes
- Geometrically:
  - Look at the attributes as variables forming a coordinate system
  - Principal Components are a new coordinate system, found by rotating the original system along the directions of maximum variability

PCA – Step 1: standardize data

- Notation (review):
  - Dataset with n rows and m columns
  - Attributes (columns): Xi
  - Mean of each attrib:

$$\mu_i = \frac{1}{n} \sum_{j=1}^{n} X_j^i$$

  - Variance of each attrib:

$$\sigma_{ii}^2 = \frac{1}{n} \sum_{j=1}^{n} \left(X_j^i - \mu_i \right)^2$$

  - Covariance between two attrib:

$$\sigma_{ij}^2 = \frac{1}{n} \sum_{k=1}^{n} \left(X_k^i - \mu_i \right) \cdot \left(X_k^j - \mu_j \right)$$

  - Correlation coefficient:

$$r_{ij} = \frac{\sigma_{ij}^2}{\sigma_{ii} \sigma_{jj}}$$

PCA – Step 1: standardize data

- Definitions
  - Standard Deviation Matrix:

$$V^{1/2}=\begin{bmatrix} \sigma_{11} & 0 & \ldots & 0 \\ 0 & \sigma_{22} & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \sigma_{mm} \end{bmatrix}$$

  - (Symmetric) Covariance Matrix:

$$Cov=\begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \ldots & \sigma_{1\mathrm{m}}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \ldots & \sigma_{2\mathrm{m}}^2 \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \sigma_{mm}^2 \end{bmatrix}$$

  - Correlation Matrix:

$$\rho=[r_{ij}]$$

  - Standardization in matrix form:

$$Z=(V^{1/2})^{-1}(X-\mu) \qquad Z_{ik}=(X_k^i-\mu_i)/\sigma_{ii}$$

- N.B. E(Z) = vector of zeros; Cov(Z) = ρ

PCA – Step 2: compute eigenvalues and eigenvectors

- Eigenvalues of ρ are
  - scalars $\lambda_1 \dots \lambda_m$ such that
  - $\det(\rho - \lambda I) = 0$
- Given a matrix ρ and its eigenvalue λ_i,
  - e_i is a corresponding eigenvector if
  - $\rho\, e_i = \lambda_i e_i$
- We are interested in eigenvalues / eigenvectors of the *correlation matrix*

PCA – Step 3: compute principal components

- Consider the vectors
  - $Y_i = e_i^T Z$
  - e.g. $Y_1 = e_{11} Z_1 + e_{12} Z_2 + \ldots + e_{1m} Z_m$
- Sort Y_i by value of variance:
  - $Var(Y_i) = e_i^T \rho e_i$
- Then
  1) Start with an empty sequence of principal components
  2) Select the vector e_i that
     1) maximizes $Var(Y_i)$
     2) Is independent from all selected components
  3) Goto (2)

PCA – Properties

- Property 1: The total variability in the standardized data set
  - equals the sum of the variances for each Z-vector,
  - which equals the sum of the variances for each component,
  - which equals the sum of the eigenvalues,
  - Which equals the number of variables

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

PCA – Properties

- Property 2: The partial correlation between a given component and a given variable is a function of an eigenvector and an eigenvalue.
  - In particular, $Corr(Y_i, Z_j) = e_{ij}\ sqrt(\lambda_i)$

- Property 3: The proportion of the total variability in Z that is explained by the ith principal component is the ratio of the ith eigenvalue to the number of variables,
  - that is the ratio λ_i/m

PCA – Experiment on real data

- Open R and read “cadata.txt”
- Keep first attribute (say 0) as response, remaining ones as predictors
- Know Your Data: Barplot and scatterplot attributes
- Normalize Data
- Scatterplot normalized data
- Compute correlation matrix
- Compute eigenvalues and eigenvectors
- Compute components (eigenvectors) – attribute correlation matrix
- *Compute cumulative variance explained by principal components*