Università degli Studi di Milano
Master Degree in Computer Science

Information Management course

Teacher: Alberto Ceselli

Lecture 08: 28/10/2015

Data Mining: Methods and Models

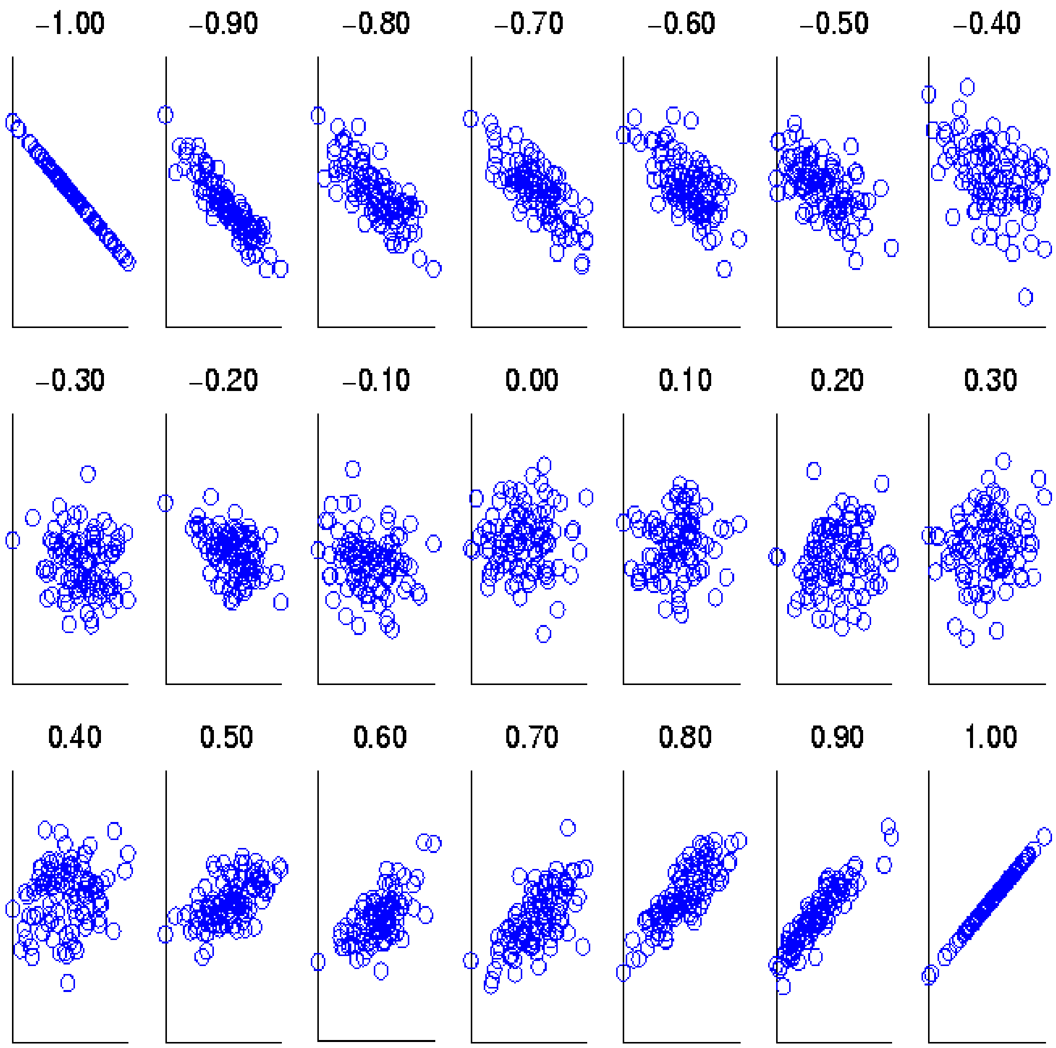
— Chapter 1 —

Daniel T. Larose

Data (Dimensionality) 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

Recall: Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

A minimal approach: user defined composites

- Sometimes correlation is *known to the data analyst* or *evident from data*
- Then, nothing forbids to *aggregate attributes by hand*!
- Example: say you have a “house” dataset
  - then housing median age, total rooms, total bedrooms and population can be *expected* to be strongly correlated as “block group size”
  - replace these four attributes with *a new attribute*, that is the *average* of them
    (possibly after normalization)

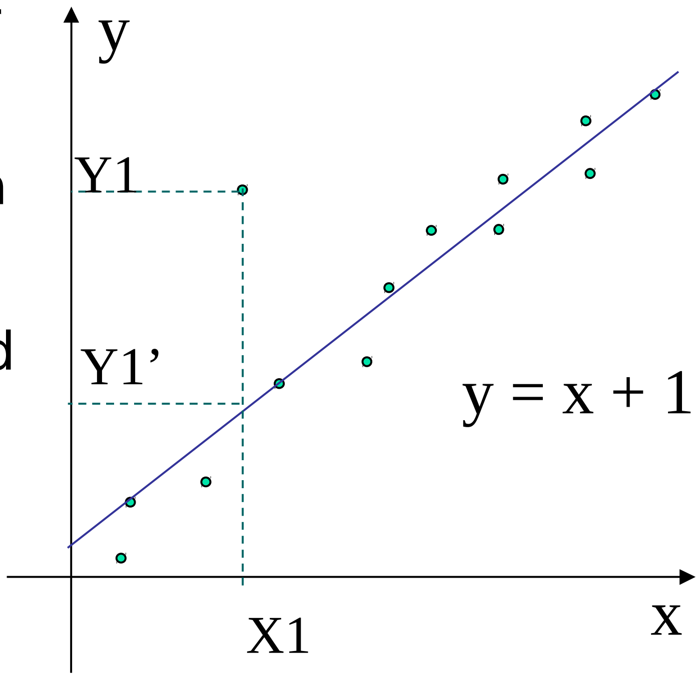
$$X^{m+1}_i = (X^1_i + X^2_i + X^3_i + X^4_i) / 4$$

Parametric Data Reduction: Regression and Log-Linear Models

- **Linear regression**
  - Data modeled to fit a straight line
  - Often uses the least-square method to fit the line
- **Multiple regression**
  - Allows a “response” variable Y to be modeled as a linear function of multidimensional “predictor” feature (variable) vector X
- **Log-linear model**
  - Approximates discrete multidimensional probability distributions

Regression Analysis

- Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values of a ***dependent variable*** (also called ***response variable*** or *measurement*) and of one or more *independent variables* (aka. ***explanatory variables*** or ***predictors***)
- The parameters are estimated so as to give a "**best fit**" of the data
- Most commonly the best fit is evaluated by using the ***least squares method***, but other criteria have also been used



- Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

Regress Analysis and Log-Linear Models

- Linear regression: $Y = w X + b$
  - Two regression coefficients, *w* and *b*, specify the line and are to be estimated by using the data at hand
  - Using the least squares criterion to the known values of Y_1, Y_2, …, X_1, X_2, ….
- Multiple regression: $Y = b_0 + b_1 X_1 + b_2 X_2$
  - Many nonlinear functions can be transformed as above
- Log-linear models:
  - Approximate discrete multidimensional prob. distributions
  - Estimate the probability of each point (tuple) in a multi-dimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations
  - Useful for dimensionality reduction and data smoothing

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: preprocess data

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

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j$$

  - Variance of each attrib:

$$\sigma_{jj}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i^j - \mu_j)^2$$

  - Covariance between two attrib:

$$\sigma_{kj}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i^k - \mu_k) \cdot (X_i^j - \mu_j)$$

  - Correlation coefficient:

$$r_{kj} = \frac{\sigma_{kj}^2}{\sigma_{kk} \sigma_{jj}}$$

PCA – Step 1: preprocess data

- Definitions
  - Standard Deviation Matrix:

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

  - (Symmetric) Covariance Matrix:

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

  - Correlation Matrix:

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

  - Standardization in matrix form:

$$Z=(X-\mu)(V^{1/2})^{-1} \qquad Z_{ij}=(X_i^j-\mu_j)/\sigma_{jj}$$

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

PCA – Step 2: compute eigenvalues and eigenvectors

- Eigenvalues of (mxm matrix) ρ are
  - scalars $\lambda_1 \ldots \lambda_m$ such that
  - $\det(\rho - \lambda I) = 0$
- Given a matrix ρ and its eigenvalue λ_j,
  - e^j is a corresponding (mx1) eigenvector if
  - $\rho\, e^j = \lambda_j e^j$
- Spectral theorem / symmetric eigenvalue decomposition (for symmetric ρ)
  - $$\rho = \sum_{j=1}^{m} \lambda_j e^j (e^j)^T$$
- We are interested in eigenvalues / eigenvectors of the *correlation matrix*

PCA – Step 3: compute principal components

- Consider the original (standardized, nxm) matrix Z, with columns Z^j
- Consider the (nx1 column) vectors
  - $Y^j = Z\, e^j$
  - e.g. $Y^1 = e^1{}_1\, Z^1 + e^1{}_2\, Z^2 + \ldots + e^1{}_m\, Z^m$
- Sort Y^j by value of variance:
  - $Var(Y^j) = (e^j)^T\, \rho\, (e^j)$
- Then
  1) Start with an empty sequence of principal components
  2) Select the vector e^j that
     1) maximizes $Var(Y^j)$
     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 column vector Z^j,
  - which equals the sum of the variances for each component,
  - which equals the sum of the eigenvalues,
  - Which equals the number of variables

$$\sum_{j=1}^{m} Var\left(Y^j\right) = \sum_{j=1}^{m} Var\left(Z^j\right) = \sum_{j=1}^{m} \lambda_j = m$$

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, $Corr(Y^k, Z^j) = e^k_j \, sqrt(\lambda_k)$

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

PCA – Experiment on real data

- Details on the dataset:
  - Block groups of houses (1990 California census)
  - Response: Median house value
  - Predictors:
    1) Median income
    2) Housing median age
    3) Total rooms
    4) Total bedrooms
    5) Population
    6) Households
    7) Latitude
    8) Longitude

PCA – Step 4: choose components

- How many components should we extract?
  - Eigenvalue criterion
    - Keep components having $\lambda > 1$ (they "explain" more than 1 attribute)
  - Proportion of the variance explained
    - Fix a coefficient of determination r
    - Choose the min. number of components to reach a cumulative variance > r
  - Scree plot Criterion
    - (try to barplot eigenvalues)
    - Stop just prior to "tailing off"
  - *Communality Criterion*

PCA – Profiling the components

- Look at principal components:
  - Comp. 1 is “explaining” attributes 3, 4, 5 and 6
    → block group size?
  - Comp. 2 is “explaining” attributes 7 and 8
    → geography?
  - Comp. 3 is “explaining” attribute 1
    → salary?
  - Comp. 4 ???
- Compare factor scores of components 3 and 4 with attributes 1 and 2

PCA – Communality of attributes

- Def: **communality** of an (original) attribute j is the sum of squared principal component weights for that attribute.
- When we consider only the first p principal components:

$$k(p,j) = corr(1,j)^2 + corr(2,j)^2 + \ldots + corr(p,j)^2$$

- Interpretation: communality is the fraction of variability of an attribute “extracted” by the selected principal components
- Rule of thumb: communality < 0.5 is low!
- Experiment: compute communality for attribute 2 when 3 or 4 components are selected

PCA – Final choice of components

- Eigenvalue criterion did not exclude component 4 (and it tends to underestimate when number of attributes is small)
- Proportion of variance criterion suggests to keep component 4
- Scree criterion suggests not to exceed 4 components
- Minimum communality suggests to keep component 4 to keep attribute 2 in the analysis

- → Let's keep 4 components