Statistical learning and prediction

Tom F. Wilderjans
Leiden University
Faculty of Social and Behavioral Sciences
Institute of Psychology
Methodology and Statistics Unit
Room 3B18
071-5276058
t.f.wilderjans@fsw.leidenuniv.nl

Overview

Today

- unsupervised methods (Chapter 10 of ISRL)
  - dimension reduction: Principal Components Analysis (PCA)
  - cluster analysis
- step back to prediction (regression) situation: combinations of supervised and unsupervised ideas (Section 6.3 of ISRL)
  - Principal Components Regression (PCR)
  - Partial Least Squares (PLS) regression

Supervised and unsupervised methods

Unsupervised vs supervised methods

- no criterion (Y) to supervise the learning
- searching for structure in the data (vs prediction)

  → groups of similar objects/variables, directions explaining most variance

- more explorative in nature
- more difficult to assess the performance of the method: how to do validation/cross-validation?
- interesting alternatives for high-dimensional problems
- sometimes used as pre-processing for supervised methods

  → identifying important variables when having many predictors

Unsupervised techniques

Different types of techniques

- dimension reduction techniques
  - Principal Component Analysis (PCA), Factor Analysis (FA), Independent Component Analysis (ICA), Nonnegative Matrix Factorization (NMF), Correspondence Analysis
  - Canonical Correlation Analysis: (cor)relation between sets of variables
- clustering techniques
  - one-mode clustering: K-means, mixture analysis, latent class analysis, hierarchical clustering
  - two-mode clustering: clustering objects and variables (biclustering)
- scaling techniques: multidimensional scaling (MDS)

Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

- first view: find (uncorrelated) linear combinations of the (correlated) variables with largest variance across the samples
  - summarizes (the variance in) the data into a small number of components (i.e., main directions in the data)
  - low-dimensional representation of the data set

- second view: line (1D) or subspace (2D/3D) closest to the data in terms of squared distances (i.e., least squares approximation)
  - (with centered variables): approximate x_{ij} with $\sum_{m=1}^{M} z_{im}\phi_{jm}$
  - find z_{im} and ϕ_{jm}'s such that $\sum_{i=1}^{N}\sum_{j=1}^{P}(x_{ij} - \sum_{m=1}^{M} z_{im}\phi_{jm})^2$ is minimal

  $\rightarrow$ with ϕ_1, ..., ϕ_m of length one and orthogonal to each other

Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

- three important concepts
  - component loadings (ϕ_{jm}): weight of each variable in the components (for interpretation of the components)
  - component scores (z_{im}): score of each case on the components (to see structure among the cases)
  - the variance of each component denotes the importance of that component

  $\rightarrow$ proportion explained variance of each component

Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

- considerations
  - always: centering of the variables (default option in many programs for PCA)
  - depending on the context: normalization of the variables (i.e., variance of one)
  - rotational freedom: rotate loadings to simple structure (varimax)

Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

- number of components
  - maximum number of components is $min(N, P)$
  - elbow in the scree plot

    → component number against proportion variance explained

    → component number against cumulative proportion variance explained
  - Kaiser's rule (do not use): eigenvalue larger than one
  - many procedures: parallel analysis, CHull

Improving least squares regression

Three methods to improve least squares linear regression model (see lecture 3)

- selection of predictors and fit least squares: subset selection
- shrinkage of regression coefficients (fit least squares with a constraint)
  - reduces the variance and can perform variable selection
- dimension reduction methods (fit least squares on derived predictors/features)
  - based on forming linear combinations of the original variables
  - no explicit selection of variables
  - not always easy to interpret the linear combinations
  - reduces the variance because some constraint on the coefficients is imposed (but may lead to bias)

    → penalty methods also constrain the coefficients

    → adding constraints is the only option when n $\ll$ p !!

Dimension reduction methods

Dimension reduction methods

- two step procedure
  - step 1: compute new variables as linear combinations of the original predictors (e.g., $\boldsymbol{z}_m$'s in PCA)
  - step 2: perform least squares regression with the new variables
- bias-variance trade off
  - the constraint increases the bias (it's a simpler model, less flexible)
  - but it may reduce the variance (especially when n « p)

Dimension reduction methods

Dimension reduction methods: some examples

- Principal Components Regression (PCR)
  - step 1: perform PCA (on standardized data) and take the first M components
    - principal components are linear combinations of the original variables that have the largest variance
    - when predictors are correlated: a few principal components will capture most of the data
    - later principal components are uncorrelated to former ones (no issue of multicollinearity)
    - when $M = P$: original least squares regression is obtained
    - larger M gives a smaller bias but a larger variance
  - step 2: perform least squares regression with these M components
  - use cross-validation to determine M

Dimension reduction methods

Dimension reduction methods: some examples

- Partial Least Squares (PLS) regression
  - PCR assumes that the direction of variation of the predictors is also the direction where the response is varying (i.e., the linear combinations are related to the response)
  - supervised way of selecting the linear combinations

    → simultaneously look for "components" that explain a lot of variance in the predictors and that are strongly related with the response
  - coefficients are obtained from univariate regressions: directions are strongly determined by variables having the largest correlation with the response

    → use standardized predictors and response

PCA, PCR and PLS: Lab and exercises

How to perform PCA, PCR and PLS in R

Exercises: *College* dataset from the *ISLR* package (the variable *Apps* is the dependent variable, so this variable should be left out for the first part of the question)

- perform PCA to detect the structure (i.e., relations between variables) in this data set (you probably want to use standardized data).
  - how many components would you extract?
  - which variables load high on each component?
  - give an interpretation of the components?
  - do you see some structure in the component scores?

PCA, PCR and PLS: Lab and exercises

Exercises: *College* dataset from the *ISLR* package (the variable *Apps* is the dependent variable)

- perform PCR and PLS to this data set and use the variable *Apps* as response variable
  - use 10-fold cross-validation to determine the optimal number of components
  - do the first two PLS components differ much from the first two PCR/PCA components?
  - give an interpretation of these two components
  - compare both methods with respect to the amount of variance in the predictors and the response that is explained by the components
  - what are the regression weights associated with the first two components for both methods?
  - which method works (i.e., generalizes) best for this data set?

Cluster Analysis

Cluster analysis techniques

- goal is to find subgroups of samples that are similar to each other (within a group)
- partitional versus hierarchical methods
  - partition: each element belongs to one and only one cluster (K-means)
  - hierarchical: agglomerative (bottom up) versus divisive (top down)

K-means clustering

K-means clustering

- group similar objects such that within-cluster variations (WCV) are minimal
- $WCV_k = \frac{1}{|C_k|} \sum_{i,i' \in C_k} d(\ \boldsymbol{x}_i\ ,\ \boldsymbol{x}_{i'}\ )^2$

  $\rightarrow\ d(\ \boldsymbol{x}_i\ ,\ \boldsymbol{x}_{i'}\ )^2 = \sum_{j=1}^{P} (x_{ij} - x_{i'j})^2$
- $WCV_k = \frac{1}{|C_k|} \sum_{i,i' \in C_k} d(\ \boldsymbol{x}_i\ ,\ \boldsymbol{x}_{i'}\ )^2 = 2 \sum_{i \in C_k} d(\ \boldsymbol{x}_i\ ,\ \bar{\boldsymbol{x}}_k\ )^2$

  $\rightarrow\ \bar{\boldsymbol{x}}_k = \frac{1}{|C_k|} \sum_{i \in C_k} \boldsymbol{x}_i$

  $\rightarrow\ d(\ \boldsymbol{x}_i\ ,\ \bar{\boldsymbol{x}}_k\ )^2 = \sum_{j=1}^{P} (x_{ij} - \bar{x}_{k(i)j})^2$
- minimize total WCV by minimize $\sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{P} (x_{ij} - \bar{x}_{k(i)j})^2$

K-means clustering

K-means clustering

- algorithm
  - find partition and centroids leading to smallest total WCV
    - → a partition of the objects (i.e., a binary matrix with rows summing to 1)
    - → centroid vector for each cluster m
  - alternating algorithm
    - → compute centroid per cluster
    - → assign each sample to its closest cluster (in terms of Euclidean distance)
    - → the loss function (i.e., total WCV) will decrease in each step

K-means clustering

K-means clustering

- problem of local optima (non-convex optimization problem): always use a multi-start procedure
- normalization (and/or centering) of the variables (or the samples)?
- How to determine K: scree plot
  - K against total WCV
  - K against percentage explained

    → ratio of ($total\ variance\ in\ data - total\ WCV$) to ($total\ variance\ in\ data$)
- how robust is the clustering? Try a subset of the data and see whether you find the same clusters

Hierarchical clustering

Hierarchical clustering

- two types of procedures: agglomerative (bottom up) and divisive (top down)
- depends on the choice of distance measure and the type of linkage
  - distance measure: Euclidean or correlation-based
  - type of linkage: complete, single, average, centroid, Ward
- results are presented in a dendrogram: see the evolution of the clustering (vs K-means)
- cut the dendrogram at a certain height to get a partition of the samples
  - cutting at different heights gives you nested clusterings
- normalization (and/or centering) of the variables (or the samples)?

Clustering: Lab and exercises

How to perform K-means and hierarchical clustering in R

Exercises

- take the matrix below and calculate 2 iterations of the K-means algorithm with $K = 2$ by hand. Start with an initial clustering with clusters of the same size

$$\begin{pmatrix} 3 & 4 \\ 6 & 6 \\ 7 & 2 \\ 2 & 5 \end{pmatrix}$$

Clustering: Lab and exercises

Exercises

- take the matrix below (i.e., different last row) and calculate the dendrogram obtained by complete, single and average hierarchical clustering (using Euclidean distance) by hand.

$$\begin{pmatrix} 3 & 4 \\ 6 & 6 \\ 7 & 2 \\ 5 & 2 \end{pmatrix}$$

- prove that for two samples $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ (measured on $j = 1, \ldots, P$ variables) the squared (Euclidean) distance between the samples is (inversely) proportional to the correlation between both samples when the P scores for each sample are standardized (i.e., each sample has a mean of zero and a variance of one)

Clustering: Lab and exercises

Exercises

- take the data set *College* from the *ISLR* package
  - only use the following three variables
    - → (variable 5) *Top10perc* Percentage new students from top 10% of class
    - → (variable 12) *Personal* Estimated personal spending
    - → (variable 13) *PhD* Percentage of faculty staff with a Ph.D.
  - perform a K-means clustering
    - → which value for K in the K-means analysis would you suggest? Why?
    - → try to interpret the obtained clustering

Clustering: Lab and exercises

Exercises

- take the data set *College* from the *ISLR* package
  - perform a hierarchical clustering (select 20 cases at random)
    - → which dissimilarity measure will you use?
    - → which linkage method(s) will you use?
    - → how many clusters will you select?
    - → perform K-means clustering (with the chosen number of clusters) and compare both clusterings (hierarchical vs K-means)