Statistical testing in latent variable models

Justyna Kowalska

LVMs

Main idea

$$Y=BL+E$$

where
Y - observed data
B - matrix of unknown parameters of interest
L - manifestation of latent variables
E - independent random variation

Types

Manifested variables / Latent variables	Continuos	Categorical
Continuos	Factor analysis (FA)	Latent trait analysis
Categorical	Latent profile analysis	Latent class analysis

Feature Selection

Purpose

finding variables (features) that have the greatest impact on the model's predictions

Methods

1. Filter methods

The filter method employs a feature ranking function to choose the best features. The ranking function gives a relevance score based on a sequence of examples. Intuitively, the more relevant the feature, the higher its rank

- mRMR
- RelieF
- CFS
- Information Gain

2. Wrapper methods

Wrapper methods search the space of feature subsets, testing performance of each subset using a learning algorithm. The feature subset that gives the best performance is selected for final use. Clearly, if there are m features in total, then there are 2^m possible subsets to search.

- RFE-SVM
- **Jackstraw**
- Boruta

3. Embedded methods

These methods use an information measure or loss function to choose the best features. Other learning algorithms have been developed with embedded feature selection in mind. For example, Littlestone’s WINNOW algorithm is an adaptation of the Perceptron algorithm that uses multiplicative weight updates instead of additive. This has the effect of rapidly degrading the weights on irrelevant features quickly, leaving only a relatively few features with nonzero weights.

- LASSO
- Elastic Net
- Stability Selection

Which one to choose?

For large data sets → Filter methods (fast and model independent)
For smaller data sets → Wrapper methods (more accurate but slower)
For integrating feature selection into the model → Embedded methods

Association tests

Purpose

The aim is to identify which manifested variables (features) have statistically significant associations with underlying latent variables that cannot be directly measured

Approach

Association testing between data and manifestation of latent variables
*naive approach is unreliable in this context because of constructing components directly from the data what leads to overestimation (null p-values)

Methods

1. Jackstraw[1]

- provides a resampling strategy and testing scheme to estimate statistical significance of association between the observed data and their systematic patterns of variation. The jackstraw methods learn over-fitting characteristics inherent in unsupervised learning, where the observed data are used to estimate the systematic patterns and to be tested again
- R package made by Neo Christopher Chung, John D. Storey, Wei Hao, Alejandro Ochoa
- applications
  - gene expression data
  - SNPs
  - scRNA-seq
  - proteomics

2. Boruta[2]

- It finds all relevant features by comparing original attributes' importance with importance achievable at random, estimated using their permuted copies (shadows) and progressively eliminating irrelevant features to stabilise that test. It’s capable of working with any classification method that output variable importance measure (VIM)
- by default Boruta uses Random Forest
- R package made by Miron Bartosz Kursa and Witold Remigiusz Rudnicki

[1] https://github.com/ncchung/jackstraw
[2] https://gitlab.com/mbq/Boruta/

Sources:

- Bellotti, T., Nouretdinov, I., Yang, M., & Gammerman, A. (2014). Chapter 6 - Feature Selection. In V. N. Balasubramanian, S.-S. Ho, & V. Vovk (Eds.), *Conformal Prediction for Reliable Machine Learning* (pp. 115–130). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-398537-8.00006-7
- Chung, N. C., & Storey, J. D. (2015). Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*, *31*(4), 545–554. https://doi.org/10.1093/bioinformatics/btu674
- Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the **Boruta** Package. *Journal of Statistical Software*, *36*(11). https://doi.org/10.18637/jss.v036.i11
- “Statistical tests and feature selection in unsupervised learning” – presentation by Neo Christopher Chung