TRIED Etude de Cas 1

OQAI1 Indoor Air Quality Study

Nelson Fernandez & Carl Robinson

15th February 2018

Data Pretreatment

- Renamed, sorted, and split data into qualitative and quantitative datasets
- Quantitative variables were usually mean centered and reduced.
- The best qualitative variables were determined by using chi-square analysis against Formaldehyde, and choosing those with P-value < 0.01
- Variable quantisation to transform quantitative variables into qualitative
  - K-means clustering to convert Formaldehyde to 4 categories
  - Quantiles used to convert various variables into 3 or 4 categories
- Sparse PCA coordinates were calculated in an attempt to both reduce the number of variables and improve their quality.

Methods

Analysis

- Spearman Correlation
- Chi-square independence tests, P-value
- Principal Component Analysis (PCA) including Sparse PCA
- Multiple Correspondence Analysis (MCA)
- Linear discriminant analysis (LDA)
- Agglomerative Hierarchical Clustering / Dendrogram

Prediction

- Gradient Boosted Classification Trees & Standard Decision Trees
- Multiple Regression: evaluation with residual graphs and correlation coefficient
- Multilayer Perceptron: using one-hot encoded qualitative variables to predict formaldehyde category

Formaldehyde Prediction Performance

Gradient Boosted Classification Trees

Accuracy of classification into 4 Formaldehyde categories (low / med / high / v.high): Raw data: [43% 10-fold cv acc]. CAH clustered: [Clust1: 56%, Clust2: 38%, Clust3: 46% acc]

Multiple Regression:

R^2 (goodness of fit) = [0.136] using all variables to predict quantitative formaldehyde value. Residual graph shows some skewing from normal distribution.

Multilayer Perceptron

Using only the qualitative variables with chisq P-value < 0.01 to predict 3 Formaldehyde categories based on quantiles: [45% 10-fold CV acc]

Decision Trees using Sparse PCA coordinates

Using only the sparse PCA principal coordinates: [35% 10-fold CV acc]