MA4128 : Advanced Data Modelling

Kevin O'Brien

Section 1 Dimensionality Reduction

Principal Component Analysis and Factor Analysis

(Principal Components (PCs) and Factors are essentially the same thing.)

1.d

Rotation Methods
Simplify Interpretation of PCA by "spreading more evenly" the explanation across multiple PCs
Varimax is the most commonly used of the orthogonal procedures. There is also the oblique rotation procedures.

1.n

True structure of data is contained in some unobserved latent variables. These variables are unknown and are explored after the analysis is carried out (this is Factor Analysis - Structure Detection etc).
We dont know how many Latent variables there are. Use PCA to make a guess : build articifical PCs to represent these latent variables.

Importantly PCA is a very specific mathematical technique - used to estimate number of PCs. Factor Analysis includes this technique, but is broader in scope and what it sets out to do.

Section 2 Cluster Analysis

2.c

Dendrogram and verticle icicle plot.

Can assess appropriate number of clusters by when clusters get linked to each other (early or late in the schedule). Still a very subjective decision. No "right" answer.

Section 3 Linear Model

3.c

For two candidate models, one with lowest AIC is the preferred model.

3.h

Variance Inflation Factor and Tolerance

$$VIF = \frac{1}{\text{tolerance}}$$

3.i

Multicollinearity: Inflates the standard errors of the regression estimates.(i.e. very wide confidence intervals, and strange inaccurate p-values) Multicollinearity : Reduces predictive power of the model. (Multicollinearity is indicative of overfitting)

3.j

Find best set of independent variables. Considering overfitting and Law of Parsimony.

3.k

Variable selection procedures are used to determine which set of independent variables best describes the data. The variables are inserted or omitted according to the strength of their correlation with the response variable Describe

- Forward Selection
- Backward Selection
- Stepwise Selection

Can use the AIC as the method of determining improvement in model. Stepwise is a combination of the first two.

Section 4: Logistic Regression

4.C

The key aspect of this question is the nature of the response variable. With OLS regression, the outcome variable is assumed to be normally distributed continious variable (for example Height). With Logistic Regression the outcome variable is a binary value (either 0 or 1/ Success or Failure), essentially categorical, and by definition such an outcome is not normally distributed.

Section 5: Missing Data

5.E

Traditional techniques: Casewise deletion. If a case (set of observations) contains a missing value for some variable, then this case is discounted from the analysis.

We have seen in class a data set that contained approx 1000 observations, but only 300 or so were used to construct the model. Massively undermining the strenght of the model.

5.F

Missing Data can massively reduced the amount of knowledge that can be obtained from a data set.

Section 6 MANOVA and Discriminant Analysis

6.g

Discriminant Analysis requires a lot of assumption to be met in order to be a valid analysis. Multiinomial Logistic Regression can provide the same the type of analysis, but require less assumptions to be met.

6.h

Use of Training/Validation/testing procedures.

6.i

Example of a univariate outlier 85 in the following data set

$$\{14, 6, 12, 14, 9, 11, 15, 7, 15, 85\}$$

Bivariate outlier: unusual combination of values, even when the individual values are not unusual themselves Example: 6 foot tall person (1.83 meters) in height would not be unusal. A person weighing eight stone would not be unusual eithe, but a 6 foot tall person weighing 8 stone would be very unusual.

Mahalanobis Distance takes covariance (correlation) into account when computing distance from centre of a cluster of values. A value that does not fit in with the overall trend in the data would have a high mahalanobis distance.

Short person 8 stone - small Mahalanobis distance
Tall Person 17 stone - small Mahalanobis Distance
Tall Person - 8 stone - very high Mahalanobis Distance