

What have we learnt?



The aims were

- Review modern data analysis methods
- Investigate complementary graphical approaches
- Test methods on “real” data

Kinds of data (and variables)



- Observational, not from designed experiments
- Continuous and categorical (occasionally ordinal)
- Dependent and explanatory variables (occasionally just associations)
- Transformations (especially smoothing)

Methodological issues



- Bias and variance, overfitting
- Train, Validation, Test
- Outliers, robustness
- Crossvalidation
- Bootstrapping of various kinds
- Curse of dimensionality
- Tuning hyperparameters

Models?



- Why? (What are the aims?)
- Where? (What kind of data?)
- When? (What assumptions are made?)
- How do they work? (Methods)
- What do they offer? (Results)

- Are they robust?
- Do they work on large data sets?

Judging models



- Global criteria
 - AIC, BIC, deviance, test error,...
- Local criteria
 - residuals, diagnostics
- Flexibility and ease of use
 - need for tuning, software availability, interpretability of output
- Success in practice

Models and methods (1)



- Least squares
- Linear regression
(subset selection, shrinkage, lasso)
- Local regression and robustness
- PCR and PLS
- Splines and smoothing
- Kernel methods and Basis functions

Models and methods (2)



- Logistic regression
- B-splines and wavelets
- gam models
- Naive Bayes, Bayes Factor
- LDA, QDA, FDA, MDA
- SVM

Models and methods (3)



- Prototypes
 - k-means, LVQ, mixture models
- Nearest neighbour
- Association rules
- (Clustering)
- (Neural nets)

Software: Models in R



- `lm`, `rlm`, `glm`, `lasso` (in `lars`)
- `gam`
- `regsubsets` (in `leaps`)
- `lda`, `qda`, `fda`, `mda`
- `svm` (in `e1071`)
- ...
- clustering, neural networks, trees, ...

Gaps



- Theoretical
 - Quadratic Programming formulations of SVM
 - EM algorithm proof
 - LVQ model
- Practical
 - SVM outputs
 - MDA software

The future



- More thorough assessment of existing models
- (Development of theory)
- Unifying structure for many special cases
- More work with model ensembles
- Better graphical support