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

Materials

Materials

- Part of Chapter 6 (not part on PCR and PLS) of ISRL

Model selection and regularization

Linear regression is great (so not immediately go to more complex models)

- easy interpretation (which predictors are most important/related to the outcome)

- often quite ok for prediction (although being a not very flexible method)
  - linear regression performs well when the true relation between outcome and predictors is more or less linear (small bias) and n is much larger than p (i.e., a large number of data points per parameter, which prevents overfitting and leads to a small variance)

Model selection and regularization

Especially when $p > n$, we can do better in prediction (control the variance) and interpretation (feature selection)

Three methods to improve least squares fitting of the linear regression model

- selection of predictors: subset selection (best or stepwise)
- shrinkage of regression coefficients: regularization/penalisation methods
  - reduce the variance and can perform variable selection
- computing linear combinations of variables and using these as predictors in a regression (see last session)
  - no explicit selection of variables
  - not always easy to interpret the linear combinations

Subset selection

Selection of predictors

- best subset selection: try all possible subsets of predictors
  - within a subset: use training SSE to select best model
  - between subsets: look at test error (different options) and take difference in complexity into account
- problems with best subset selection
  - computational intensive for large p
  - overfitting: "fit the data too hard" (increase in variance)

Subset selection

Selection of predictors

- solution: stepwise methods
  - only explore a restricted set of models: nested models
  - based on training SSE to find best model with a given complexity (i.e., next model in the sequence)
  - based on test SSE to find the final model
  - forward, backward and hybrid forms (not exactly the same as in SPSS)

    $\rightarrow$ when $p > n$: only forward subset selection is possible
  - when predictors are correlated, it may not find the best subset in terms of training error as it is a "greedy approach" (but this is not necessarily bad for test error)
- stepwise models (also) depend on peculiarities in the data (i.e., which variable to add/remove)

Subset selection

Choosing the optimal model (step 3 in best subset and stepwise approaches)

- look at the model with lowest test error (i.e., taking differences in complexity into account)

- two approaches
  - indirectly estimate test error: adjust training error (i.e., adjust model for its complexity)

    $\rightarrow$ Mallow's C_p, AIC, BIC, adjusted R^2

  - directly estimate test error

    $\rightarrow$ validation approach and cross-validation

    $\rightarrow$ perform subset selection on the training data (and not on the full data)

    $\rightarrow$ one-standard error rule: look for a more parsimonious model

Regularization

Shrinkage methods (regularization, penalty methods)

- no least squares fit but fit a penalized least squares function
  - penalty parameter λ balances model fit (RSS) and size of the coefficients
  - penalty will shrink the coefficients to 0 (cf. constrained optimization: having a tight budget)
- as λ increases (less flexibility): decrease in variance at the cost of some bias (i.e., underestimation of coefficients)
  - to decrease bias: de-biased lasso

Regularization

Shrinkage methods (regularization, penalty methods)

- two penalties used
  - ridge regression (L2-norm): coefficients never become 0 (no variable selection)
  - lasso (L1-norm): for large enough λ coefficients become zero (variable selection)
  - when the true model is sparse: lasso outperforms ridge regression
  - other options: elastic net, bridge, ...
- use standardized variables in regularization methods and do not shrink the intercept
- cross-validation needed to find optimal λ

  $\rightarrow$ use same split to fit models for a grid of λ's

Subset selection and regularization (exercises)

Exercise

```
> student_full$G3bin = student_full$G3>8 # to make a dichotomous outcome
> student_full = read.csv2( "student-mat.csv" )
```

- use the student data but now with all variables. The outcome measure now is the mathematics achievement at moment 3 (*G3*). Use all other variables as predictors (don't include *G1bin*).

- fit ridge regression and the lasso and determine the optimal value of λ by means of 10-fold cross-validation

- make a dichotomous variable *G3bin*

Subset selection and regularization (exercises)

- fit a ridge logistic regression and a lasso penalised logistic regression (predict *G3bin* and use all other variables as predictors, also *G3*). Determine the optimal λ for both methods by means of 10-fold cross-validation. Determine (also with CV) the test error for both methods (using the model with optimal complexity). Which method performs best for this data set?

- Which variable is the most important in predicting *G3bin*?