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

Overview

First session

- introduction in learning (supervised and unsupervised learning, bias-variance trade-off, training-test error)

This session

- resampling methods (re-using the data set)
  - validation approach: test set prediction error
  - cross-validation (Leave One Out, LOOCV): test set prediction error
  - bootstrap: variability of an estimate (standard error, confidence intervals and significance testing)
  - permutation/randomization tests: testing for significance
- Classification methods

Materials

Materials

- Chapter 4 and Chapter 5 (ISLR)

Performance of a statistical method

How to get an idea about performance of a statistical method (e.g., prediction performance)

- get new independent samples from the population and test your predictions
- if method is good and results are generalizable
  - these predictions should be good
  - low test error
- estimate the test error

Estimating test error

Why do you want to estimate test error?

- training set error is too optimistic

  $\rightarrow$ becomes always better when having more complex models

- overfitting when having too complex models gives a worse test set error (i.e., no generalization possible)
  - adding more predictors in a regression model results in smaller residuals, but does not make the regression estimates generalizable to the population (i.e., an new independent sample from the population)
  - the same for determining the number of clusters in cluster analysis, the number of dimensions in PCA
  - growing a tree too large may undermine its generalizability

Estimating test error

Why do you want to estimate test error?

- overfitting: identifying peculiarities of your sample data (not good for generalization)
  - you are not really interested in your sample characteristics (it's just a sample from the population) but you want to know the population characteristics (overfitting will misguide you with regard to population characteristics)
  - reduces bias: more flexible model that is closer to the true model
    $\rightarrow$ deviation between the population quantity and the mean of all estimates (when considering all possible samples of equal size from the population)
  - but gives a large variance in your estimator (over independent new samples): ideosyncratic elememts will play no role in a new sample
    $\rightarrow$ large standard error: variability in the estimates when considering all possible samples (of the same size) from the population

Estimating test error

Why do you want to estimate test error?

- overfitting: identifying peculiarities of your sample data (not good for generalization)
  - when models fit equal, take the most parsimonious one (also gives a better interpretation)
  - remark: introducing bias may sometimes imply a large decrease in variance and thus a smaller MSE (e.g., penalty methods in Chapter 6)
  - call for parsimonious models (also better from an interpretational and substantive point of view)

Problem

Problem

- not always possible to get independent samples (means you need to collect extra data): you don't have a (large) test set
- solution: adjustment of training error to make it bigger (Mallow's CP-statistic, AIC, BIC) based on amount of fitting and the variance (i.e., a penalty for complexity)
- solution: resampling methods (use the same training sample over and over again): validation and (leave-one-out) cross-validation

Validation approach

Goals

- estimate test error of a method (to compare methods)
- determining optimal model complexity/flexibility (overfitting)

Procedure

- split data in 2 (equal) parts (training and validation set) and fit model to training data
- make predictions of data in validation set based on the fitted model for the training data
- compute prediction error for the validation set
  - regression (quantitative outcome): mean squared (prediction) error
  - classification (categorical outcome): misclassification error

Validation approach

- for model selection: take model with lowest prediction error
- for determining test error: take prediction error as estimate for test error

Evaluation of the validation method

- large bias: you only use half of the data
  - over-estimation of test error because of less data/information
  - is test error for half of the data
- only one estimate for test error with a large variance (variability) due to the split
  - is two-fold validation without cross-over (i.e., changing the role of training and validation set)

  $\rightarrow$ 2-fold cross-validation would be better (mean of 2 estimates)

Validation approach (exercise)

How to perform the validation approach in R

Exercise: which factors predict mathematics achievement?

- 394 students
- dependent variable (*G1*): score on a mathematics exam
- three predictors
  - *studytime*: amount of study time
  - *schoolsup*: extra support by school (yes/no)
  - *romantic*: involved in a romantic relationship (yes/no)
- questions
  - use the validation approach to determine the test error for a model including the main effects and interaction between study time and school support

Validation approach (exercise)

- questions
  - select an optimal (not too complex) regression model with the validation approach: choose between the following models
    - only an intercept
    - only a main effect of study time
    - main effects of study time and school support
    - main effects and interaction between study time and school support
    - the previous model with an added main effect of romantic
    - all effects between study time, school support and romantic
  - repeat the analysis with another split in training and validation set. Will you select the same model?
  - do you think that the validation approach is a good or a bad approach? Give arguments using a bias-variance perspective.

Cross-validation

Same goals as the validation-approach: (1) estimating test error and (2) determining model complexity

Procedure of K-fold cross-validation

- divide data in K (equal) parts and perform K times the validation approach
  - crossed over: each part in turn is the validation test (and other parts constitute the training set)
  - compare with (2-fold) validation (CV is kind of validation with different splits)
- take the average of estimated test error over the K folds (weighted by size of validation set)
  - it is better to take an average than a single estimate of test error (as in the validation approach)

Cross-validation

Jackknife (LOOCV): Leave-One-Out Cross-Validation

- is cross-validation with $K = N$
  - the validation set consists of a single data point, all other cases are in the training set
  - almost no bias: training on $N - 1$ data points
- take the average estimated test error over $K = N$ folds
  - would work great if the training samples would be independent from each other
  - however large variance: resampled training sets are very much alike (correlated)

    $\rightarrow$ large dependence is not good because new samples do not add much extra information (compare with adding correlated predictors in regression or adding multiple measurements of the same person)

Cross-validation

Bias-variance trade off in choosing K

- bias: depends on the size of the training set (the larger, the better)
  - training on less data gives an upwards bias (less information/data gives larger estimates for test error)
  - LOOCV: small bias
  - smaller K (working with less data): larger bias
- variance: depends on the dependence between the samples on which the estimated test errors are based (the more independent, the better)
  - LOOCV: large variance due to a high dependency between training samples
  - smaller K: smaller variance
- what for $K = 2$?

Cross-validation

Bias-variance trade off in choosing K

- what for $K = 2$?
  - large bias, small variance
- optimal K is 5 or 10

Cross-validation

Remarks on Cross-Validation

- CV the good and wrong way: include pre- and post-processing steps
  - pre: picking predictors, centering and scaling
    $\rightarrow$ the steps will introduce some extra variability that you need to include to get a good test error estimated for the procedure
  - post: model selection procedure (i.e., selecting λ in regularization methods)

- sometimes the model with lowest MSE (based on CV) is only a little bit better than a more parsimonious model
  - the CV estimate for MSE is only an estimate (it has uncertainty)
    $\rightarrow$ in videos: formula for SE for CV estimate of MSE
  - 10% (or 1 SE) rule: select the most parsimonious model amongst all models that have a CV error that is less than 10% larger than the CV error of the optimal model (or within 1 SE)

Cross-validation

Remarks on Cross-Validation

- when using cross-validation to estimate test error, while also deciding on model complexity (i.e., number of predictors in the model): a combination of cross-validation (inner step for model selection) and validation (outer step for estimating test error)
  - step 1: split the data in test set and training set
  - step 2: perform K-fold cross-validation on the training set (i.e., for each fold make a validation set and a -smaller- training set) and decide on the model complexity
  - step 3: for the model complexity determined with cross-validation, fit it on the full training data (see step 1)
  - step 4: predict the test set (see step 1) based on the model in step 3 and compute the MSE for the test set

    $\rightarrow$ it is also possible to do a cross-validation at the outer step: double cross-validation

Cross-validation (exercise)

How to perform cross-validation and LOOCV in R

Exercise

- select an optimal model (amongst the six models defined earlier) by using LOOCV, 5-fold and 10-fold cross-validation
- write a function to repeat K-fold cross-validation S times. Display the S CV-error curves in a single plot (and plot the mean of the S curves)
- what would be a good choice for K in the cross-validation procedure? Argue in terms of the bias-variance trade off

Bootstrap

Why performing bootstrap analysis?

- goal is to get an idea of variability of an estimator and constructing bootstrap intervals (and significance testing)

  $\rightarrow$ PCA (confidence interval for loadings), statistics with no formula for the standard error

Procedure

- sampling with replacement from original (training) data
- standard deviation of the estimates is an estimate for the standard error of the estimator (is parametric version)
- non-parametric bootstrap estimate for the standard error: middle 95% values of the estimates
- different procedure to construct bootstrap confidence intervals (e.g., percentile method)

Bootstrap

- bootstrap procedure depends on the problem under study
  - regression analysis: 2 ways of performing the bootstrap based on whether predictors are considered fixed or random
  - time series regression: as there is correlation in the data points (data points are not iid draw from the population), it is better to use a block bootstrap
- bootstrap is good for determination of the standard error of a statistic
- bootstrap is not good for estimating prediction test error
  - test error estimate will be biased (i.e., too positive view) due to the overlap between bootstrap samples and between each bootstrap sample and the original data set
  - using not-sampled data points for each bootstrap sample as validation data is a complicated method that does not give better results than cross-validation

Bootstrap

Literature: bootstrap

- (introductory) Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping. A nonparametric approach to statistical inference* (Series: Quantitative Applications in the Social Sciences). Londen, UK: Sage Publications, Inc.

- (applied) Chernick, M. R. (2008). *Bootstrap methods. A guide for practitioners and researchers* (second edition). Hoboken, NJ, USA: John Wiley & Sons, Inc.

- (more theoretical) Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY, USA: Chapman & Hall

Permutation tests

Permutation tests

- distribution-free statistical test (when assumptions are violated and non-parametric methods are not OK)
- calculate the probability of getting a value as extreme or more extreme than an obtained value of a test statistic under a null hypothesis by recalculating the test statistic for all or many rearrangements or divisions of the data
- idea is to test a specific null-hypothesis
  - independent samples t-test
  - H_0 (no association): means in both groups are equal (can also be used for medians, variances)

Permutation tests

Permutation tests

- generate data under the null-hypothesis
  - take all observed values and randomly divide them into two groups (with the original sample sizes)
  - compute the difference in means
  - repeat this 1000 (or more) and build a sampling distribution for the mean difference
  - determine p-value (can be one- or two-sided)
- can be used when sample sizes are small

Permutation tests

Permutation tests

- testing whether a correlation is zero
  - permute the cases at random in one of both variables and compute the correlation
  - repeat 1000 times and build a sampling distribution for the correlation
  - compute p-value
- sounds a bit like bootstrapping but it is different
  - no interest in standard error of a statistic
  - you specify a specific null-distribution and generate data under this hypothesis
- different applications: correcting for multiple comparisons (fMRI: testing group differences for many voxels), testing for treatment differences in single case designs, ...

Permutation tests

Literature: permutation tests

- Ernst, M. D. (2004). Permutation methods: A basis for exact inference. *Statistical Science, 19*(4), 676-685. doi:10.1214/ 08834230400000396

- practical applications of permutation tests

  $\rightarrow$ video on $http://videolectures.net/as2010_good_pap/$

- use and misuse of permutation tests

  $\rightarrow$ see *http://statcourse.com/misuse.htm*

- Welch, W. J. (1990). Construction of permutation tests. *Journal of the American Statistical Association, 85*, 693-698.

Bootstrap and permutation tests (exercise)

How to perform bootstrap in R

Exercise

- determine the standard error for the main effect of romantic and the interaction effect between time spent studying and school support by means of a bootstrap (using model 6). Compare the bootstrap estimate for the standard error with the standard error provided by the regression analysis.

- test whether the correlation between the score on the mathematics test and the time spent studying significantly differs from zero (determine the two-sided p-value of the test)

- test whether students without school support have a significantly better score on the mathematics test than students with school support (determine the two- and one-sided p-value of the test)