2001-353-w06b-p-Resamp-PredictionErrorEstimates-FeatureSelection

Profs: R. H. French, L. S. Bruckman, P. Leu, K. Davis, S. Cirlos

TAs: W. Oltjen, K. Hernandez, M. Li, M. Li, D. Colvin

23 February, 2023

Contents

6.1.2.1 Reading, Homeworks, Projects, SemProjects

- Readings for next class:
  - For Today, ISLR7 Beyond Linear Models, (R4DS22-25)
  - For next Tuesday ISLR8 and DL08/DL09
- Laboratory Exercises:
  - LE3 due on Today
  - LE4 given out Tommorrow
- Office Hours: (Class Canvas Calendar for Zoom Link)
  - Wednesdays @ 4:00 PM to 5:00 PM
  - Saturdays @ 3:00 PM to 4:00 PM
  - **Office Hours are on Zoom, and recorded**
- Semester Projects
  - Office Hours for SemProjs: Mondays at 4pm on Zoom
  - DSCI 453 Students Biweekly Updates Due
    - Update # is Due ** **
  - DSCI 453 Students
    - Next Report Out # is Due ** **
  - All DSCI 353/353M/453, E1453/2453 Students:

        - Peer Grading of Report Out #1 is Due ** **
    - Exams
        - MidTerm: **Thursday March 9th**, in class or remote, 11:30 - 12:45 PM
            - **CWRU Spring Break is March 13th to March 17, so NO CLASS**
        - Final: **Thursday May 4th**, 2023, 12:00PM - 3:00PM, Nord 356 or remote

6.1.2.2 Textbooks

6.1.2.2.1 Introduction to R and Data Science

- For students new to R, Coding, Inferential Statistics
    - Peng: R Programming for Data Science
    - Peng: Exploratory Data Analysis with R
    - OIS = Diez, Barr, Çetinkaya-Runde: Open Intro Stat v4

6.1.2.3 Tidyverse Cheatsheets, Functions and Reading Your Code

- Look at the Tidyverse Cheatsheet
    - **Tidyverse For Beginners Cheatsheet**
        - In the Git/20s-dsci353-353m-453-prof/3-readings/3-CheatSheets/ folder
    - **Data Wrangling with dplyr and tidyr Cheatsheet**]

Tidyverse Functions & Conventions

- The pipe operator `%>%`
- Use `dplyr::filter()` to subset data row-wise.
- Use `dplyr::arrange()` to sort the observations in a data frame
- Use `dplyr::mutate()` to update or create new columns of a data frame
- Use `dplyr::summarize()` to turn many observations into a single data point
- Use `dplyr::arrange()` to change the ordering of the rows of a data frame
- These can be combined using `dplyr::group_by()`
    - which lets you perform operations "by group".
- The `%in%` matches conditions provided by a vector using the c() function
- The **forcats** package has tidyverse functions
    - for factors (categorical variables)
- The **readr** package has tidyverse functions
    - to read_…, melt_… col_…, parse_… data and objects

Reading Your Code: Whenever you see

- The assignment operator `<-`, think **"gets"**
- The pipe operator, `%>%`, think **"then"**

6.1.2.4 Syllabus

6.1.2.5 ISLR Chapter 5 Resampling Methods: Prediction Error Estimates

- Having procedures to evaluate different statistical methods is essential
    - Such as using Training and Testing Datasets
        - So that we can determine the prediction accuracy
        - Or the prediction error

Resampling Methods also do the same thing

- Such as Cross-validation
- Bootstrap

Day:Date	Foundation	Practicum	Readings(optional)	Due(optional)
w01a:Tu:1/17/23 w01b:Th:1/19/23	Markov Cluster Stat. Learning, Approach	R, Rstudio IDE, Git Bash, Git, Class Repo	ISLR1,2 (R4DS-1-3)	(LE0)
w02a:Tu:1/24/23 w02b:Th:1/26/23 w02Pr:Fr:1/27/23	Lin. Regr. Bias-Var. Train/Test, Bias vs. Vari. **ADD DROP**	SemProjs; Regr. Ovrvw Tidyverse Review **DEADLINE**	ISLR3,(R4DS-4-6) DL01 DL02 (R4DS-7,8)	**(LE0:Due)** LE1 **453 Update 1**
w03a:Tu:1/31/23	Logistic Regr. Classif	Pred. Analytics, Regr.	DL03,ISLR4	
w03b:Th:2/2/23	LDA/QDA	ggPlot2, Code Expect.	DL04, DL05	**LE1:Due**, LE2
w03:Sa:2/4/23				**LE1:Due**
w04a:Tu:2/7/23 w04b:Th:2/9/23 w04Pr:Fr:2/10/23	Resample Cross-Valid. DL, ML Overview	ggplot Multilevel Mod.	ISLR5 ISLR6 (R4DS9-16)	 **453 Update 2**
w05a:Tu:2/14/23	Resampling: Bootstrap	Bootstrap Mixed Effects	DL2R1, DL06,07	**LE2:Due**, LE3
w05b:Th:2/16/23	Subset Selec., Shrink.	Dim. Red. PCA	DLwR2	
w05Pr:Fr:2/17/23				**453 Rep. Out 1**
w06a:Tu:2/21/23	ML with NNs	ggplot, clustering	DLwR3	
w06b:Th:2/23/23	Beyond Linear Modls	Feature Select., Caret	ISLR7 (R4DS22-25)	**LE3:Due**, LE4
w06Pr:Fr:2/24/23				**453 Update 3**
w07a:Tu:2/28/23 w07b:Th:3/2/23	Dec. Trees, Rand. Forest MidTerm Review, SVM	Tidy Modeling SVM, SVR, ROC	ISLR8, DL08,09 ISLR9 (R4DS26-30)	 **Peer Review 1**
w08a:Tu:3/7/23	ML Overview	, Keras/TF2, Torch	ISLR10	
w08b:Th:3/9/23	**MIDTERM EXAM**		DL10,11	**LE4:Due** LE5
w08Pr:Fr:3/10/23				**453 Update 4**
Tu:3/14/23 Th:3/16/23	**SPRING** **SPRING**	**BREAK** **BREAK**	ISLR10 DL12,13	
w09a:Tu:3/21/23	Deep Learning	TF2 Keras Intro	Pocket Perceptron	ISLR10, DLR3
w09b:Th:3/23/23	Computer Vision, CNN	CNN w/TF2, Overfit	DLR4	
w09Pr:Fr:3/24/23				**453 Rep. Out 2**
w10a:Tu:3/28/23	Deep Learn Intro	NN Types	DLR5	
w10b:Th:3/30/23 w10Pr:Fr:3/31/23	DL CNN,RNN ImageNet	NN Types, CNN wTF2	Hinton ImageNet	 **453 Upd.5 & PrRev 2**
Sa:4/1/23				**LE5:Due** LE6
w11a:Tu:4/4/23 w11b:Th:4/6/23	Fitting NNs NLP, Graphs & ML	AUC,Prec,Recall Fruit	 LeCun DL Rev. 2015	
w12a:Tu:4/11/23	Graphs & ML	NLP with sequences	DLR6	
w12b:Th:4/13/23	NLP w attention	Graph Repr Proc Wrkflw		**LE6:Due** LE7
w13a:Tu:4/18/23	DL Frameworks	Explaining DL w Lime		
w13b:Th:4/20/23	Linux Distros XGBoost	Explain Preds	Deep Dream	
w13Pr:Fr:4/21/23				**453 Rep. Out 3 Due**
w14a:Tu:4/25/23	Tranformers			
w14b:Th:4/27/23	Final Exam Review	Torch NN & DeepLearn		**LE7:Due**
w14Pr:Fr:4/28/23				**Peer Rev 3 Due**
	FINAL EXAM	**Th. 5/4/23, 12-3pm**	Nord 356 & Zoom	
	453 Final PDF Report	**Fr. 4/29, 11:59pm**		

Table 1: DSCI353-353M-453 Weekly Syllabus. R4DS-x.y, OISx.y, ISLRx.y, DLGBx.y refers to chapters and sections assigned as reading in our textbooks. DLx are deep learning articles.

Figure 1: IT Fundamentals: Applied Data Science with R, Syllabus

- Leave-one-out cross-validation

In addition there are other metrics of statistical significance

We've seen

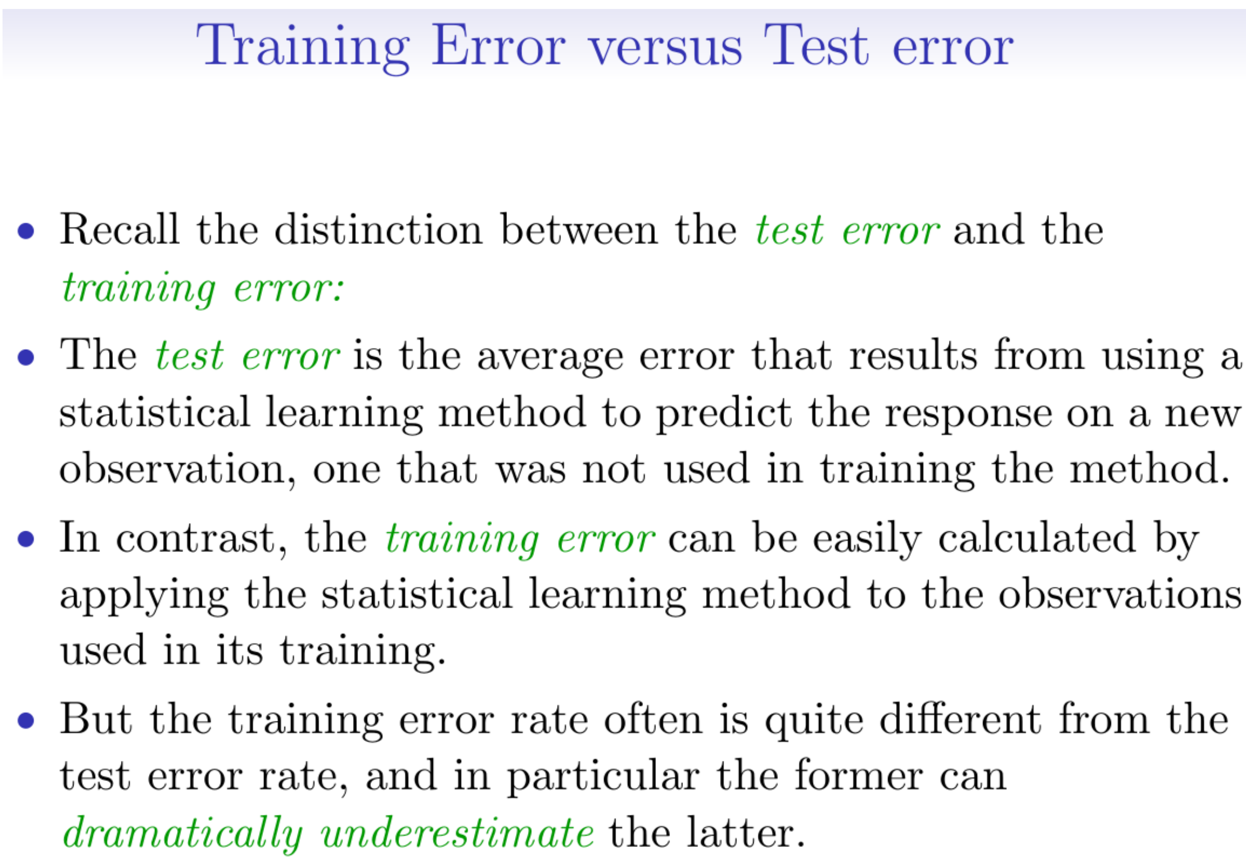
- R^2 and $adj.R^2$
- p-values for null hypothesis testing

There are also C_p statistic, AIC and BIC and others

- Malloy's C_p statistic
- Akaike Information Criteria, AIC
- Bayesian Information Criteria, BIC

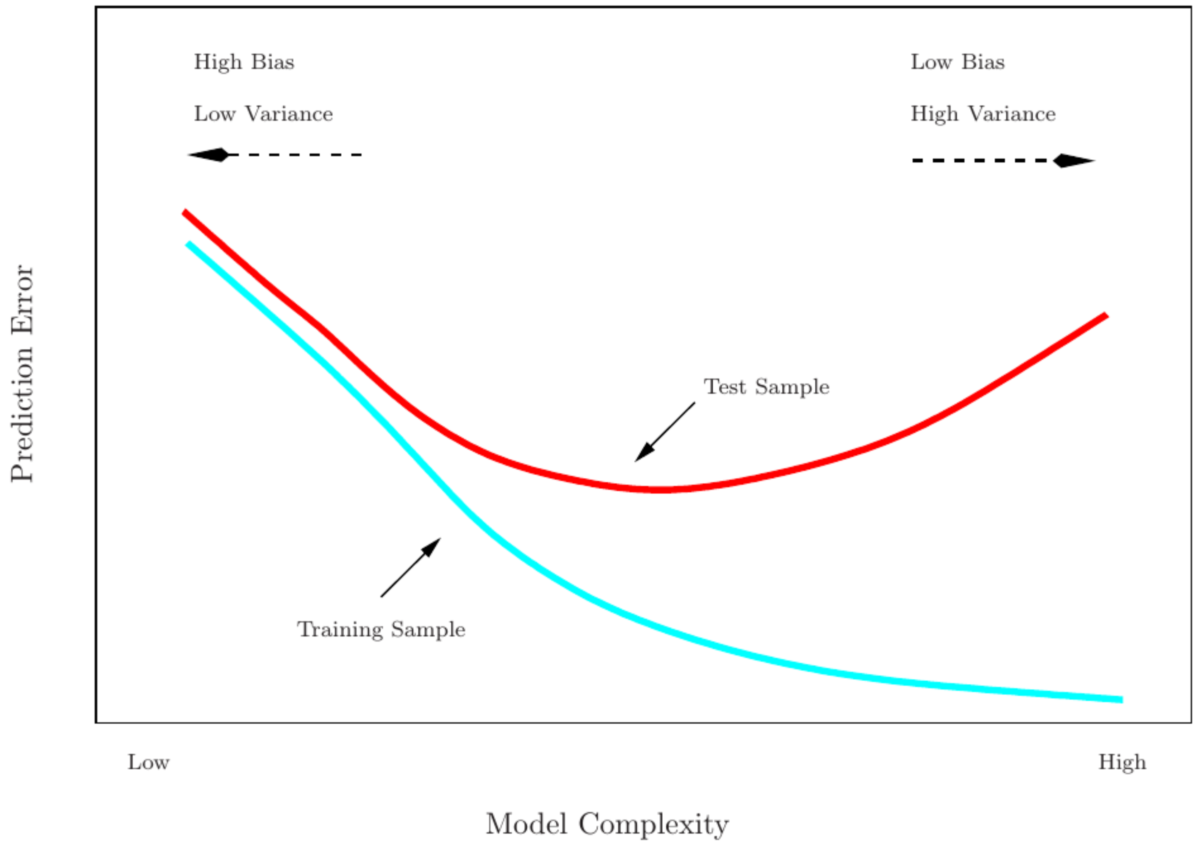
6.1.2.6 Training and Testing for Prediction Error Determination

- Training vs. Testing Error



Training vs. Testing Performance

Training- versus Test-Set Performance



6.1.2.7 Cross-validation and Bootstrap

- Cross-validation and Bootstrap

Cross-validation and the Bootstrap

- In the section we discuss two *resampling* methods: cross-validation and the bootstrap.
- These methods refit a model of interest to samples formed from the training set, in order to obtain additional information about the fitted model.
- For example, they provide estimates of test-set prediction error, and the standard deviation and bias of our parameter estimates

6.1.2.8 Other Prediction Error Estimates

- Other Prediction Error Estimates

More on prediction-error estimates

- Best solution: a large designated test set. Often not available
- Some methods make a *mathematical adjustment* to the training error rate in order to estimate the test error rate. These include the *Cp statistic*, *AIC* and *BIC*. They are discussed elsewhere in this course
- Here we instead consider a class of methods that estimate the test error by *holding out* a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations

6.1.3 ISLR6 Linear Model Selection and Regularization

6.1.3.1 Feature or Variable Selection

- These are all parts of the general topic of Feature Selection

In machine learning and statistics, feature selection,

- also known as variable selection,
  - attribute selection or variable subset selection,
- is the process of selecting a subset of relevant features
  - (variables, predictors) for use in model construction.

Feature selection techniques are used for three reasons:

- simplification of models to make them easier to interpret by researchers/users,
- shorter training times,
- enhanced generalization by reducing overfitting
  - (formally, reduction of variance)

The central premise when using a feature selection technique is

- that the data contains many features that are either redundant or irrelevant,
- and can thus be removed without incurring much loss of information.

Redundant or irrelevant features are two distinct notions,

- since one relevant feature may be redundant
- in the presence of another relevant feature
  - with which it is strongly correlated.

Feature selection techniques should be distinguished from feature extraction.

- Feature extraction creates new features from functions of the original features,
- whereas feature selection returns a subset of the features.

Feature selection techniques are often used in domains

- where there are many features
- and comparatively few samples (or data points).

Archetypal cases for the application of feature selection include

- the analysis of written texts and
- DNA microarray data,

where there are many thousands of features,

- and a few tens to hundreds of samples.

6.1.3.2 Prediction Error Estimates and Optimality Criteria

- And Optimality Criteria

Three common ones are the following

- Malloy's C_p Statistic
- Aikake Information Criteria
- Bayesian Information Criteria

6.1.3.2.1 Malloy's C_p Statistic

- In statistics, Mallows's C_p,
  - named for Colin Lingwood Mallows,

    - is used to assess the fit of a regression model
        * that has been estimated using ordinary least squares.

It is applied in the context of model selection,

- where a number of predictor variables are available for predicting some outcome,
    - and the goal is to find the best model involving a subset of these predictors.
- A small value of C_p means that the model is relatively precise.

Mallows's C_p has been shown to be equivalent

- to Akaike information criterion
    - in the special case of Gaussian linear regression.

6.1.3.2.2 Aikake Information Criteria

- The Akaike information criterion (AIC)

    - is a measure of the relative quality of statistical models
        * for a given set of data.
        * based on Shannon's information theory and information entropy

Given a collection of models for the data,

- AIC estimates the quality of each model,
    - relative to each of the other models.
- Hence, AIC provides a means for model selection.

AIC is founded on information theory:

- it offers a relative estimate of the information lost
    - when a given model is used to represent the process that generates the data.
- In doing so, it deals with the trade-off between
    - the goodness of fit of the model
    - and the complexity of the model.
- The Akaike information criterion (AIC)
    - is a measure of the relative quality of statistical models
    - for a given set of data.
- Given a collection of models for the data,
    - AIC estimates the quality of each model,
    - relative to each of the other models.
- Again, AIC provides a means for model selection.

AIC does not provide a test of a model

- in the sense of testing a null hypothesis; i.e.

AIC can tell nothing about the quality of the model in an absolute sense.

- If all the candidate models fit poorly,
- AIC will not give any warning of that.

6.1.3.2.3 Bayesian Information Criteria

- In statistics, the Bayesian information criterion (BIC)

    - or Schwarz criterion (also SBC, $SBIC$)
    - is a criterion for model selection
        * among a finite set of models;
        * the model with the lowest BIC is preferred.

It is based, in part, on the likelihood function

- and it is closely related to the Akaike information criterion (AIC).

When fitting models,

- it is possible to increase the likelihood by adding parameters,
- but doing so may result in overfitting.

Both BIC and AIC resolve this problem

- by introducing a penalty term for the number of parameters in the model;
- the penalty term is larger in BIC than in AIC.

The BIC was developed by Gideon E. Schwarz

- and published in a 1978 paper,
- where he gave a Bayesian argument for adopting it.

6.1.3.3 Cites

- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning: With Applications in R. 1st ed. 2013, Corr. 5th printing 2015 edition. Springer Texts in Statistics. New York: Springer, 2013.
- https://en.wikipedia.org/wiki/Statistics