Fall 2023

Wharton
UNIVERSITY *of* PENNSYLVANIA

Statistics 4220
Predictive Analytics

Professor Robert Stine
317 Academic Research Building
stine@upenn.edu

Overview

This half-semester course introduces predictive techniques that extend beyond regression analysis introduced in courses such as Stat 201 or Stat 621. These extensions select the most predictive model, classify observations into categories, and partition data into homogeneous subsets. The emphasis is conceptual, focused on the motivation for and properties of various methodologies.

Illustrative applications use real data, typically with more cases and variables than encountered in intro courses. The datasets are "tabular", with features organized into columns and cases into rows as if sampled, independent observations from some population. Other courses cover unstructured data such as text and sequential data such as time series.

This course divides into four segments. The first two segments emphasizing the role of statistical inference using tests and p-values to judge the predictive performance of models. The first segment reviews least squares, emphasizing interpretation, nonlinear and categorical features, and graphical diagnostics. The second segment introduces logistic regression designed for classification, as in predicting the purchase choice made by a consumer. Examples use held-back data to illustrate how statistical tests anticipate results when fitted models are applied to new data.

The last two segments adopt the spirit of machine learning, relying on held-back data in place of statistical tests. The third segment covers model selection. Models often look better when built than realized in practice, particularly when the modeler chooses from among many potentially predictive features. Statistical techniques avoid such over-confidence by following a theoretical "model selection criterion" whereas machine learning techniques require held-back data. Key among these is the lasso. The fit from a lasso is nonetheless familiar, an equation with estimated coefficients.

The final segment introduces models that avoid the explicit, equation-based structure of regression. Tree-based models partition the sample, collecting observations into batches whose averages serve as predictions. Enhanced versions of this approach that combine many trees have established a strong reputation in practice and continue to rival the best neural networks when building predictive models from tabular data.

Prerequisites

With one quarter to cover these topics, this course moves at a brisk pace and presumes students are familiar with inferential statistics, such as the fundamental terminology and use of hypothesis tests, confidence intervals, and *p*-values. Students should also be familiar with applied multiple regression analysis, specifically the use and interpretation of least squares regression (including R^2, RMSE, various *t*- and *F*-tests, confidence/prediction intervals, and residual diagnostics). Prior exposure to logistic regression and Bayesian statistics is helpful but not required.

The course presumes some familiarity with the statistics package JMP. This is not a programming course and instead relies on the built-in capabilities of this software to illustrate the practical use of the covered methods. Students are encouraged to use on-line resources, particularly ChatGPT, for guidance on how to run this software.

Materials

Class notes provide background for the methods and examples covered in lectures; these will be distributed via Canvas. These notes cover some details but rely on the required textbook for further explanation:

> *An Introduction to Statistical Learning, 2nd Edition* (2021, abbreviated ISL)
> James, Witten, Hastie, and Tibshirani
> https://hastie.su.domains/ISLR2/ISLRv2_website.pdf
>
> Datasets and supplemental materials for this book
> https://www.statlearning.com/resources-second-edition

This text uses the R statistical modeling environment; as noted, we'll use JMP to avoid the overhead of learning how to program in R while also learning the methodology. JMP is available via Canvas.

Further background reading that covers multiple regression may be found in the following casebook and text (used in some Wharton classes).

> *Business Analysis Using Regression*. Foster, Stine, and Waterman; Springer, 1998.
> *Business Statistics* (BS) Stine and Foster, Addison-Wesley, 2010.

Office hours and TA

A schedule for TA hours will be posted in Web Café.

My office hours are on Monday and Wednesday afternoons from 3:30 to 5:30 p.m. Otherwise the best way to contact me is by e-mail.

Requirements

Total grade = 75% assignments + 25% cumulative test

All relevant University policies regarding Academic Integrity must be followed in regard to assignments and tests.

Assignments

These consist of weekly data analyses problem sets that reinforce classroom discussion and develop hands-on familiarity with how to use the methods to build predictive models**.**

- I expect each student to submit solutions to an assignment expressed in his or her own words. It is okay to discuss assignments with classmates, but the assignments are to be written up independently.
- Assignments must be turned in during class by the submitting student.
- All questions about the grading of an assignment must be resolved within one week of the date on which the graded assignment is returned.

Cumulative Test

The final test will be closed book (one page of notes), during the final class period (Wednesday, October 18).

ChatGPT

Within this class, you are welcome to use AI models in an unrestricted fashion at no penalty. You should be aware that tools based on large language models such as GPT may make up incorrect facts and fake citations. You will be responsible for inaccurate, biased, offensive, or otherwise unethical content you submit regardless of whether it comes from you or an AI program. Having said this, the use of an AI program is encouraged, as it may make it possible for you to submit assignments with higher quality, in less time. The university's policy on plagiarism still applies to any uncited or improperly cited use of work by other human beings, or submission of work by other human beings as your own.

Overview of Course Topics

We won't spend a day on each module; those later in the course will likely take more time. Consider sections of these notes not covered in class to be supplemental, as if from portions of the course textbook that you skim rather than read carefully.

Be aware: I will occasionally revise these notes that are posted on-line to correct errors or omissions and to elaborate topics that arise in class discussion. Check the revision date on the first page of the notes.

Multiple Regression ***ISL, Ch 1-3***

Module 1.0 Review example

The methods illustrated in this regression analysis should be familiar.

Module 1.1 Multiple regression

Further details of multiple regression, emphasizing interpretation, inference, and diagnostics.

Module 1.2 Categorical explanatory variables

Regression models make heavy use of group information in the form of categorical variables and their interactions with other features. Questions of multiplicity arise.

Module 1.3 Calibration

Every predictive model should be calibrated. It is easy to check, "low-hanging fruit". There's no reason not to.

Readings

Skim chapters 1-2 of ISL. We'll not use R, so skip those portions such as Sections 2.3 and 3.6 unless you're interested. We will not cover K-means clustering, but it's easy to understand for the sake of reading the comparisons offered in Section 3.5. BTW, JMP makes it easy to construct 3-D figures like Figure 3.4 or 3.5.

Classification ***ISL, Ch 4***

Module 2.1 Simple logistic regression

When the response that you want to predict is categorical, you're in the domain of generalized regression. We start with models having one predictor to figure out how to fit and interpret logistic regression. No more least squares.

Module 2.2 Multiple logistic regression

Logistic regression can use more than one predictor, just like least-squares regression. Once we allow several predictors, how are we supposed to decide which features belong in a multiple logistic regression? Once we identify them, what do they mean?

Readings

Focus on Section 4.3 that introduces logistic regression. ISL includes an alternative classifier known as linear discriminant analysis (LDA) which is basically least squares regression fit to a dummy variable. Skim the details of LDA and focus on the use of logistic regression. Notice that important methods for evaluating a classifier (*e.g.* ROC curve, confusion matrix) are introduced in the discussion of LDA (pages 150-152). If you go further into "machine learning", you're likely to run into naïve Bayes (Section 4.4.4) so don't skip this part. Skim (or omit if you're pressed for time) the discussion of QDA and Poisson regression in Section 4.6. The discussion of QDA alerts you to assumptions of LDA; Poisson regression illustrates a type of regression designed for counts.

Model Selection

Module 3.1 Over-fitting *ISL, Ch 6, Sec 7.1*

It's easy to pick the model if you only have a few predictive features, but what if you have thousands? Once you recognize the importance of interactions, the set of possible models grows geometrically. Greedily picking the best apparent choice using an procedure such as stepwise regression typically leads to over-fitting: the fitted model looks better than when applied to new data.

Module 3.2 Cross validation *ISL, Ch 5*

Regression models often claim to do better than they actually can. The problem is most apparent when we test a model using the same data that we used to pick the model. Cross-validation avoids this problem by setting aside data for testing the model. Variations on this theme include repeating the process and recognizing that cross-validation itself makes an optimistic assessment of the model.

Module 3.3 Model selection techniques *ISL, Sec 6-6.2*

Cross-validation helps, but becomes slow (and inconsistent) when applied more generally. Instead, penalty methods (ridge and lasso) and selection criteria (AIC, BIC) filter down the collection of competitive models with less computation. For example, the Bonferroni method limits the model to predictors that provide assured gains in accuracy.

Readings

Concerns about model selection run throughout the text, starting from Chapter 2 onward (*e.g.* peek back at Section 2.2). We won't spend time with bootstrap resampling (Section 5.2), so skim this discussion so you can appreciate the analysis in Section 5.3. We won't cover PCA or partial least squares (Section 6.3), but you may run into these in other courses.

Partitioning *ISL, Sec 2.2, Ch 8*

Module 4.1 Trees

In a general sense, all statistical models work by averaging. The question is which observations to average. Regression models use an equation to identify which observations should be used to predict new data. Alternatives based on partitioning the data are more direct. Just look and see which observations are close to each other. Decision trees do this partitioning recursively.

Module 4.2. Forests and boosted trees

Trees are natural, but often unstable (illustrating the bias/variance trade-off introduced in Chapter 2). We gain stability (at the cost of interpretation) by averaging over many trees. Boosting combines lots and lots of little trees (*a.k.a.,* shrubs, perhaps taking the analogy too far). A specific implementation of this approach (XGBoost) has become a reference method in many fields of application.

Readings

JMP simplifies the construction of decision trees, so you can incrementally see (and control) how the method partitions the sample. We won't cover Bayesian additive regression trees (Sec 8.2.4), so skim that material.