STATISTICS 452/652: Statistical Learning and Prediction

November 30, 2020

Final Review Outline

(Reading: ISLR Sections 3.1–3.4)

1 Regression Review

1.1 Intro Stat

- Distributions and models
- Sampling distributions and the effects of sampling on statistics
- Standard errors and relationship to n

1.2 Simple Linear Regression

- Model and parameters
- Least squares
- Sampling distribution of regression estimates
- Sampling variability in predicted values

1.3 Multiple Linear Regression

- Model and parameters
- Least Squares
- Hyperplane
- Multicollinearity

1.4 How the Universe Works

- $Y = g(\mathbb{X}) + \delta$
  - True structure, universal function, universal predictor
  - Irreducible error
- Approximating truth with model $f(X)$
- **Bias-Variance Tradeoff (BVT)**
  - Bias of a model
  - Variance of a model
- Modeling is managing BVT

2 Evaluating Models

- Sums of squares, sMSE
- Prediction error, MSPE
- Overfitting
- Sample re-use, resampling, data splitting
  - Training/validation/test sets
  - Random splits
  - Cross-validation
  - Bootstrap
- Using resampling methods to select models
- relative error

3 Extensions of Variables

- Categorical explanatories
  - Binary indicator/dummy variables
  - How R does it
- Feature Engineering
  - Transformations
    - Polyniomials
  - Interactions
  - Arbitrary functions of multiple variables

4 Simplifying Models

- Relationship between model complexity and BVT

4.1 Subset Selection

- All Subsets regression
- Stepwise selection
  - Forward
  - Backward
  - Hybrid methods
- Criteria
  - MSPE
  - Information Criteria

4.2 Shrinkage

- Using something other than least squares
  - Penalized least squares
- Ridge regression
  - Shrinkage penalty
  - Tuning parameter
    - GCV
- LASSO
  - Simultaneous shrinkage and variable selection
  - LASSO penalty
    - Choosing tuning parameter with CV

4.3 Dimension Reduction

- Reduce complexity of models
- Allow $n < p$
- Principal Components Analysis
  - Rotate axes to account for variance in X

  - Project data onto PC

- Principal Components Regression
  - Choose smaller number of PCs, $M < p$, for model
  - Ignores Y

- Partial Least Squares
  - Similar to PC, but uses Y to help select components
  - Choose smaller number than p to work with

5 Flexible Regression Models

- Regions and indicator variables
- Step Functions
- Basis Function

5.1 Splines

- Regression/Basis/Cubic Splines
  - Piecewise polynomials
  - Knots
  - Smoothness constraints

- Natural Splines

- Smoothing Splines
  - Natural splines with lots of knots and shrinkage
  - Equivalent Degrees of Freedom

- Local Polynomial Regression (LOESS)
  - Kernel-weighted function within a neighbourhood

- Mostly limited to 1 dimension

5.2 Spline Applications in Higher Dimensions

- Generalized Additive Models
  - Like linear regression, but with splines in each dimension
  - Limited ability to model interactions
- Projection Pursuit
  - Create components optimally
  - Fit spline to component
  - Repeat on residuals

6 Modern Statistical Learning Machines

6.1 Neural nets

- Hidden layer(s) of Hidden nodes
- Weights
- Activation function
- Decay/shrinkage
- Poorly identified parameters
  - Slow Convergence
  - Sub-optimal minimum
  - Multiple re-starts
- Tuning!
- Pre-process data for `nnet()`

6.2 Trees

- Decision tree concept
- Splitting/partitioning data
  - recursively applied to resulting subsets
- Stopping rules
- Pruning
- Properties of predictions

6.3 Ensembles

- Bagging
    - Bootstrap aggregation
    - Refitting learners to resamples
    - Averaging across resamples
    - Properties
- Random Forest
    - Bagging regression trees
    - Added tweak of subsampling variables at each split
    - Variable Importance
    - Tuning
- Boosting
    - Fitting small trees in sequence to residuals
    - Incrementing prediction function by small amount
        - Shrinkage
    - Tuning

7 Classification

7.1 Problem of Classification

- Categorical response variable
    - K possible classes at each x
    - Discrete distribution for $P(Y = k|X = x)$
    - "True" class
    - Irreducible error
- Goal is to predict "true" class (most likely class) at each x
    - Classifier is a machine $f(X)$ that guesses true class
    - Bayes classifier
- Misclassification rate
- Confusion matrix
- Decision boundaries
- BVT

K Nearest Neighbour Classifier

- Predictions based on most likely class among M neighbours
- M controls BVT
- Not great in higher dimensions

8 Linear Classifiers

8.1 Logistic regression

- Model log-odds (logit) as linear regression
- Estimates $P(Y = k|X = x)$ for each k
- Multi-response (baseline) logits for multiple classes
- Linear decision boundaries

8.2 Discriminant Analysis

- Multivariate normal distribution for X within each class
- Linear discriminant analysis
  - Equal variances and correlations across groups
  - Linear decision boundaries.
- Quadratic Discriminant Analysis
  - Unequal variances and correlations across groups
  - Quadratic decision boundaries
- Choice is BVT

9 Nonlinear extensions

- Generalized Additive model
  - Extension of logistic regression
  - Uses splines for each variable instead of linear terms
  - Created flexible decision boundaries
  - No interactions
- Naive Bayes

  - Extension of discriminant analysis
  - Assumes correlations are all 0
  - Kernel density estimate or normal
  - PCA rotation or not

10 Tree-based classifiers

- Classification tree
  - Splits to increase node purity
  - Prediction is larges class in terminal node
  - Pruning
- Random forests
  - Bagging classification trees
  - Subset of variables for each split
  - Trees vote on class
  - Variable importance
- Boosting
  - Slowly build machine
  - Lots of small trees
  - R function doesn't work for $K > 2$

11 Modern Machines

- Neural Nets
  - Response indicators
  - Estimating means
  - Softmax function
  - Classifier is highest score
  - Same tuning as for regression
- Support Vector Machines
  - Just for classification
  - Optimal separating hyperplane
  - Support vectors

- Margin
- Maximal margin hyperplane
- Add slack and penalize for nonseparable cases
  - Cost is a tuning parameter
- Expand dimension for better separation
- Linear SVM becomes nonlinear in original space
- Kernel functions
  - Gaussian Radial Basis
  - Polynomial
  - Tuning parameters on each
- Multiclass SVM
  - Series of 1 vs 1 SVMs
  - Class with most votes wins.