Data Mining
Model Selection

Bob Stine
Dept of Statistics, Wharton School
University of Pennsylvania
From Last Time

• Review from prior class
  • Calibration
  • Missing data procedures
    Missing at random vs. informative missing
  • Problems of greedy model selection
    Problems with stepwise regression.
    So then why be greedy?

• Questions
  • Missing data procedure: Why not impute?
    “Add an indicator” is fast, suited to problems with many missing.
    Imputation more suited to small, well-specified models.
    EG. Suppose every X has missing values. How many imputation models do you need to build, and which cases should you use?
Topics for Today

• Over-fitting
  • Model promises more than it delivers

• Model selection procedures
  • Subset selection
  • Regularization (aka, shrinkage)
  • Averaging

• Cross-validation
Model Validation

• Narrow interpretation
  • A predictive model is “valid” if its predictions have the properties advertised by model
  • Calibrated, right on average
  • Correct uncertainty, at least variance

• Must know process that selected model
  • Cannot validate a model from a static, “published perspective”
  • Stepwise model for S&P 500 looks okay, but...

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>29</td>
<td>0.00424379</td>
<td>0.000146</td>
<td>14.1056</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>Error</td>
<td>52</td>
<td>0.00053947</td>
<td>0.000010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>81</td>
<td>0.00478325</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Model Validation

- Fails miserably (as it should) when used to predict future returns
  - Predictors are simply random noise
  - Greedy selection overfits, finding coincidental patterns

\[ \pm 2 \text{ RMSE} \]
Over-Fitting

• Critical problem in data mining
  • Caused by an excess of potential explanatory variables (predictors)

• Claimed error rate steadily falls with size of the model

• “Over-confident”
  • Model claims to predict new cases better than it will.

• Challenge
  • Select predictors that produce a model that minimizes the prediction error without over-fitting.
Multiplicity

• Why is overfitting common?

• Classical model comparison
  • Test statistic, like the usual t-statistic
    Special case of likelihood ratio test
  • Designed for testing one, a priori hypothesis
  • Reject if $|t| > 2$, p-value < 0.05

• Problem of multiple testing (multiplicity)
  • What is the chance that the largest of $p$ z-statistics is greater than 2?

| p   | $P(\text{max } |z| > 1.96)$ |
|-----|-----------------------------|
| 1   | 0.05                        |
| 5   | 0.23                        |
| 25  | 0.72                        |
| 100 | 0.99                        |
Model Selection

• Approaches
  • Find predictive model without overfitting
  • Three broad methods

• Subset selection
  • Greedy $L_0$ methods like forward stepwise
  • Penalized likelihood (AIC, BIC, RIC)

• Shrinkage
  • Regularized: $L_1$ (lasso) and $L_2$ (ridge regression)
  • Bayesian connections, shrink toward prior

• Model averaging
  • Don’t pick one; rather, average several

Next week
Subset Solution

- Bonferroni procedure
  - If testing $p$ hypotheses, then test each at level $\alpha/p$ rather than testing each at level $\alpha$.
  - $\Pr(\text{Error in } p \text{ tests}) = \Pr(E_1 \text{ or } E_2 \text{ or } \ldots \text{ or } E_p) \leq \sum \Pr(\text{Error } i^{th} \text{ test})$
  - If test each at level $\alpha/p$, then $\Pr(\text{Error in } p \text{ tests}) \leq p(\alpha/p) = \alpha$

- Not very popular... easy to see why
  - Loss of power
  - Cost of data-driven hypothesis testing

<table>
<thead>
<tr>
<th>$p$</th>
<th>Bonferroni $z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2.6</td>
</tr>
<tr>
<td>25</td>
<td>3.1</td>
</tr>
<tr>
<td>100</td>
<td>3.5</td>
</tr>
<tr>
<td>100000</td>
<td>5.0</td>
</tr>
</tbody>
</table>
Discussion

- **Bonferroni is pretty tight**
  - Inequality is almost equality if tests are independent and threshold $\alpha/p$ is small

- **Flexible**
  - Don’t have to test every $H_0$ at same level
  - Allocate more $\alpha$ to ‘interesting’ tests
    - Split $\alpha=0.05$ with $\frac{1}{2}$ to $p$ linear terms and $\frac{1}{2}$ to all interactions

- **Process matters**
  - Look at model for stock market in prior class
  - Many predictors in model pass Bonferroni!
    - The selection process produces biased estimate of error $\sigma$
    - Use Bonferroni from the start, not at the end
Popular Alternative Rules

• Model selection criteria
  • AIC (Akaike information criterion, $C_p$)
  • BIC (Bayesian information criterion, SIC)
  • RIC (risk inflation criterion)

• Designed to solve different problems
  • “Equivalent” to varying p-to-enter threshold
  • AIC, $C_p$: Accept variable if $z^2 > 2$
    Equivalent to putting p-to-enter $\approx 0.16$
  • BIC: $z^2 > \log n$
    Aims to identify the “true model”
  • RIC: $z^2 > 2 \log p \approx$ Bonferroni
    The more you consider, the stiffer the penalty
Penalized Likelihood

• Alternative characterization of criteria

• Maximum likelihood in LS regression
  • Find model that minimizes -2 log likelihood
  • Problem: always adds more variables (max $R^2$)

• Penalized methods
  • Add predictors so long as
    
    $-2 \text{ log likelihood} + \lambda \ (\text{model size})$

    decreases

• Criteria vary in choice of $\lambda$
  • 2 for AIC, $(\log n)$ for BIC, $(2 \log p)$ for RIC
Example

- JMP output
  - Osteo example
- Results
  - Add variables so long as BIC decreases
  - Fit extra then reverts back to best
- AIC vs BIC
  - AIC: less penalty, larger model

What happens if try either with stock market model?
Shrinkage Solution

- Saturated model
  - Rather than pick a subset, consider models that contain all possible features
  - Good start (and maybe finished) if $p << n$

- Shrinkage allows fitting all if $p > n$

- Shrinkage maximizes penalized likelihood
  - Penalize by “size” of the coefficients
  - Fit has to improve by enough (RSS decrease) to compensate for size of coefficients

  - Ridge regression: $\min \text{RSS} + \lambda_2 b'b$
  - LASSO regression: $\min \text{RSS} + \lambda_1 \sum |b_j|$

$\text{RSS}$ analogous to $-2 \log \text{likelihood}$

$\lambda = \text{regularization parameter, a tuning parameter that must be chosen}$

$p = \# \text{possible Xs}$
Lasso vs Ridge Regression

\[ \text{min RSS, } \Sigma |b_j| < c \]

Corners produce selection

\[ \text{min RSS, } \Sigma b_j^2 < c \]

Interpret \( \lambda \) as Lagrange multiplier.
Cross-Validation Solution

• Common sense alternative to criteria
  - Apply the model to new data
  - Estimate ‘hidden’ curve plot of over-fitting

• No free lunches
  - Trade-off
    More data for testing means less for fitting:
    Good estimate of the fit of a poorly estimated model.
    Poor estimate of the fit of a well estimated model.

• Highly variable
  Results depend which group was excluded for testing
  Multi-fold cross-validation has become common

• Optimistic
  Only place I know of a random sample from same population

• Multi-fold: leave out different subsets
Variability of CV

• Example
  • Compare ‘simple’ and ‘complex’ osteo models
    Need to fit both to the same CV samples… Not so easy in JMP
  • Evaluate one model

• Method of validation
  • Exclude some of the cases
  • Fit the model to others
  • Predict the held-back cases
  • Repeat, allowing missing data to affect results
  • Compare out-of-sample errors to model claims

• Is assessment correct?
  • Under what conditions?
Osteo Example

- CV 50 times, split sample
- Variability
  - If only did one CV sample, might think model would be 20% better or 15% worse than claimed!

Test cases look worse

Test cases look better
CV in Data Mining

• DM methods often require a three-way CV
  • Training sample to fit model
  • Tuning sample to pick special constants
  • Test sample to see how well final model does

• Methods without tuning sample have advantage
  • Use all of the data to pick the model, without having to reserve a portion for the choice of constants
  • Example: method that has “honest” p-values, akin to regression model with Bonferroni

• Caution
  • Software not always clear how the CV is done
  • Be sure CV includes the choice of form of model
Lasso

• Regularized regression model
  • Find regression that minimizes
    \[
    \text{Residual SS} + \lambda \sum |\beta_i| \]
    where \( \lambda \) is a tuning constant
  • Bayesian: double exponential prior on \( \beta \)
  • Scaling issues
    What happens if the \( \beta \)'s are not on a common scale?

• L₁ shrinkage
  • Shrink estimated parameters toward zero
  • Penalty determines amount of shrinkage
    Larger penalty (\( \lambda \)), fewer variable effects in model
  • Equivalent to constrained optimization
Lasso Example

• How to set the tuning parameter $\lambda$?

• Empirical: Vary $\lambda$ to see how fit changes
  • Cross-validation, typically 10-fold CV
  • Large values of $\lambda$ lead to very sparse models
    Shrinks all the way back to zero
  • Small values of $\lambda$ produce dense models
  • CV compares prediction errors for choices

• Implementations
  • Generalized regression in JMP Pro
  • glmnet package in R (See James et al, Ch 6)
    More “naked” software than JMP or Stata
Lasso Example

- Fit $L_1$ regression, Lasso
  - Plot estimated coefficients as relax penalty
  - Implemented in JMP as “generalized regression”

Where to stop adding features?

osteomodel
Lasso Example in R

- Follow script from James
  - See on-line document “Glmnet Vignette”
- Similar output
  - Less formatting, but more accessible details
Discussion of CV

• Use in model selection vs model validation
  • Shrinkage methods use CV to pick model
  • Validation reserves data to test final model

• Comments on use in validation
  • Cannot do selection and validation at same time
  • Flexible: models do not have to be nested
  • Optimistic
    Splits in CV are samples from one “population”
    Real test in practice often collected later than training data
  • Population drift
    Populations often change over time; CV considers a snapshot

• Alternatives?
  • Bootstrap methods
Take-Aways

• Overfitting
  • Increased model complexity often claims to produce a better fit, but in fact it got worse

• Model selection methods
  • Criteria such as AIC or p-value thresholds
  • Shrinkage methods such as lasso

• Cross validation
  • Multiple roles: validation vs model selection
  • Flexible and intuitive, but highly variable
Some questions to ponder...

• If you fit a regression model with 10 coefficients, what’s the chance that one is statistically significant by chance alone?
  • How can you avoid this problem?

• If you have a coefficient in your model that has a $t\approx 2$, what is going to happen to its significance if you apply split-sample CV?

• Why is cross-validation used to pick lasso models?

• Is further CV needed to validate a lasso fit?
Next Time

• Thursday    Newberry Lab
  • Hands-on time with JMP, R, and data
  • Fit models to the ANES data
    You can come to class, but I won’t be here!

• Friday    July 4th holiday