Validation and Interpretation

Bias Variance Tradeoff / Holdout Testing
Thus Far...

- This class has focused on developing an intuition around selecting appropriate methodologies for research questions.
  - Linear vs. Logistic Regression, Repeated Measures, Survival Analysis...
  - Z vs. T tests, 3 group statistics, multiple comparisons....
Thus Far...

• This class has focused on developing an intuition around selecting appropriate methodologies for research questions.
  – Linear vs. Logistic Regression, Repeated Measures, Survival Analysis...
  – Z vs. T tests, 3 group statistics, multiple comparisons....

• Along the way we have studied the assumptions of models and statistical techniques to assume the results they provide are reliable
  – Model assumptions, Multicollinearity
  – Normality and non-parametric statistics
Thus Far...

- This class has focused on developing an intuition around selecting appropriate methodologies for research questions.
  - Linear vs. Logistic Regression, Repeated Measures, Survival Analysis...
  - Z vs. T tests, 3 group statistics, multiple comparisons....

- Along the way we have studied the assumptions of models and statistical techniques to assume the results they provide are reliable
  - Model assumptions, Multicollinearity
  - Normality and non-parametric statistics

- We have also explored in great detail how these techniques can aid in better understanding data
  - Defining Statistical differences
  - Using linear models to identify significant relations between multiple variables (model inference)
What’s Missing

However we have yet to answer the question:  
*How good is this model?*
What’s Missing

However we have yet to answer the question: *How good is this model?*

• We ventured into this a bit with the $R^2$ statistic
  – However, this only tells us how well the model *fits the data*
  – It does not describe how well the model can be used to model yet unseen instances (i.e. a new patient)
This Week

• We are going to set out to answer some of the big remaining questions around model performance
  – Monday/Wednesday:
    • How do we obtain reliable performance estimates?
      – Motivation and Methods commonly used to assess model performance
  – Friday:
    • How do we measure the performance of a model?
      – Discussion of various performance metrics and their uses
  – Bonus: Time Permitting
    • How do we compare competing models?
A Theoretical Foundation

Bias Variance Tradeoff

![Bias Variance Tradeoff Diagram]

- Low Bias, Low Variance
- Low Bias, High Variance
- High Bias, Low Variance
- High Bias, High Variance
The Intuition Behind the Tradeoff

\[ d_{f,\theta}(y, t) = \text{Bias}_\theta + \text{Variance}_f + \text{Noise}_t \]

- \( f \) refers to the amount of force applied
- \( \theta \) denotes the angle of the launcher
- \( t \) corresponds to the location of the target
Bias-Variance in Classifiers

Quantifying model performance label can be analyzed using the same approach

*Predictions may be correct, while others can be way off the mark*

• **Bias:**
  – Design choices in the model introduce a bias analogous to that of the projectile launcher into the classifier (various fitting methods, i.e. those to deal with multicollinearity)

• **Variance:**
  – Different instances included in our data can impact lead to different decision boundaries (in the case of linear models parameter estimates).

• **Noise:**
  – The expected error (i.e., noise) is associated with the intrinsic noise in the data. That is, some instances with the same attributes may have different outcomes.
Remember...

- No classifier is inherently better than any other: you need to make assumptions to generalize

- Three kinds of error
  - **Inherent (noise):** unavoidable
  - **Bias:** How much the average model over all training sets differ from the true model?
    - Error due to inaccurate assumptions/simplifications made by the model
  - **Variance:** How much models estimated from different training sets differ from each other
    - Due to inability to perfectly estimate parameters from limited data
So, how does all this tradeoff impact us when building our models?

- Models with too few parameters are inaccurate because of a large bias (not enough flexibility).
- Models with too many parameters are inaccurate because of a large variance (too much sensitivity to the sample).
Generalization – More Formally

- **Underfitting:** model is too “simple” to represent all the relevant class characteristics
  - High bias and low variance
  - High training error and high test error

- **Overfitting:** model is too “complex” and fits irrelevant characteristics (noise) in the data
  - Low bias and high variance
  - Low training error and high test error
Underfitting

Low variance, high bias method
Overfitting

Low bias, high variance method
Estimating Model Performance

How do we estimate performance measures? Can I just compute distance of each point to the true value?
Estimating Model Performance

How do we estimate performance measures?
Can I just compute distance of each point to the true value?

No!

• These data points were used to create the line of best fit.
• Called resubstitution error.
• Not a good indicator of the performance on future data.

Simple solution

• Spit the available data into training and testing sets.
• This allows us to test “unseen” data without having to collect new data samples.
A Typical Model Setup
Bonus: Data Snooping

• It is important that the test data is not used in any way to create the classifier.

• Some learning schemes operating in two stages
  – Stage 1: builds the basic structure
  – Stage 2: optimizes parameter settings

• The test data cannot be used for parameter tuning.
  – Why?
Validation Data

Proper procedure uses three sets: training data, validation data, and test data.

- A validation dataset is a subset of the data used to tune parameters.
- Typically used when an appropriate model needs to be chosen from several rivaling approaches.
Model Evaluation Setup
Including Validation Data For Tuning

Data

Training Set

Test Set

Induction

Deduction

Learning Algorithm

Learn Model

Apply Model

Validation Set

Model
Validating & Interpreting Data

Preprocessing

Statistical Methods

Classification & Regression

Breaking Data into Test Sets

• Holdout
  – Single holdout
  – Stratified holdout
  – Repeated holdout

• Cross validation
  – Leave-one-out validation
  – $k$-fold validation
  – Stratified $k$-fold validation

• Bootstrap (Extra)
Breaking Data into Test Sets

• **Holdout**
  – Single holdout
  – Stratified holdout
  – Repeated holdout

• **Cross validation**
  – Leave-one-out validation
  – $k$-fold validation
  – Stratified $k$-fold validation

• **Bootstrap (Extra)**
Holdout Estimation

Key Idea:

Reserve a certain amount of data for testing and use the remainder for training.

Train Data

70%

Test Data

30%
Simple (Single) holdout

- In Simple holdout data is randomly selected from the original data and broken (often at a specific percentage) into a training and testing dataset.

- Often times it is suggested we first shuffle the data
  - Helps to prevent against any potential collection bias due to time or related instances
Stratified Holdout

• Generate holdout using *stratified sampling*.  
  – New subsets of instances have approximately equal proportions of classes

• Ensures classes are equally represented in samples.
Stratified Holdout

• Generate holdout using *stratified sampling*.  
  – New subsets of instances have approximately equal proportions of classes

• Ensures classes are equally represented in samples.
Stratified Holdout

• Generate holdout using *stratified sampling*.  
  – New subsets of instances have approximately equal proportions of classes

• Ensures classes are equally represented in samples.
Simple / Stratified Holdout Challenges

• For small or “unbalanced” datasets, instances might not be representative.
  – This is addressed by stratified holdout. However for small datasets the size of the “test” data may still be insufficient to determine if the model can generalize well.

• The data used for training and testing may vary significantly.
  – Since it is a single train-and-test experiment, the holdout estimate of error rate will be misleading if we happen to get an “unfortunate” split
Repeated Holdout

- Repeated holdout, improves the reliability of the holdout estimate by repeating the process with different subsamples.
  - Also known as “random subsampling.”

- For a set number of iterations, the same proportion of data is randomly selected for training utilizing either simple or stratified holdout method.
  - Error rates on different iterations averaged to yield overall error
Primary Challenge of Repeated Holdout

*Overlapping test sets.*
Next Class

Cross-Validation

Test Data ------- Training Data

Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5

Total Dataset