Cross Validation

STA303/STA1002: Methods of Data Analysis II, Summer 2016

Michael Guerzhoy

First attempt at figuring out how well the model fits the data

- The larger the log-likelihood, the better the model
- Problem: larger models will *always* have larger log-likelihoods than models nested inside them
  - Adding an interaction term will always increase the likelihood by at least a little bit
- Solution: only prefer larger models if they substantially increase the likelihood

Idea: k-fold Cross-Validation

- A model is good if it predicts *new* data well
- Split the dataset into k folds.
  - Fit the model k times, with k-1 folds as the training set, and the remaining fold as the validation set
  - Compute some measure of how well you're doing on the validation set
- K-fold cross-validation with k=N (the number of datapoints) is called "leave-one-out cross-validation"
- AIC approximates leave-one-out cross-validation

Measuring how well are we doing on the new data

- For linear regression, we minimize the Mean Squared Error
- For logistic regression, we minimize the Negative Log Likelihood (equivalently, deviance)
- We can also measure how well we are predicting new data
  - Depends on what we want to achieve

Statistical Inference

- If we find the best model using cross-validation, can we then make inferences about its parameters?
  - Not really – model selection is itself a form of data snooping
- However, the best predictive model we can get is the one with the best cross-validation performance