# Lecture 5

# Last time

- Models fit training data better than new data
- Models chosen by optimizing training performance (e.g. deviance) will be overfit
- Methods for approximating performance (log-likelihood, deviance, accuracy, etc.) on new data:
  - Train/test splits
  - $k$-fold cross-validation
  - leave-one-out cross-validation

**Next step:** A way to systematically search through models

# General model search idea

1. Consider a set of potential models
2. Search through set

- As we search, fit models and calculate an optimality criterion

3. Choose the model with the best optimality criterion

**Question:** What is our set of potential models, and how do we search through that set?

# Best subset selection

- Set of potential models: *all* possible combinations of the available explanatory variables
- Calculates optimality criterion for *every* potential model in the set

**Question:** Are there any potential issues with best subset selection?

# Forward stepwise selection

# Backward stepwise selection

Similar to forward stepwise selection, but in the other direction:

1. Specify the largest model we are willing to consider
2. Consider each term in the model; remove the term which most improves the optimality criterion
3. Repeat Step 2 until the optimality criterion can no longer be improved

# Some issues with model selection

- Best subset selection is computationally prohibitive with too many variables
- Stepwise selection algorithms are greedy, and generally won’t return the best model
- Optimality criteria involving cross-validation can be computationally expensive when calculated for many models in a search procedure

# Alternative optimality criteria

# AIC and BIC

# Class activity

https://sta712-f23.github.io/class_activities/ca_lecture_5.html