Lecture 5

Last time

- When evaluated on the *training* data, a large model will have a lower deviance than any of its sub-models
  - Prediction metrics like AUC are also often higher, even if a reduced model is correct
- Often prefer the simpler model (easier to interpret, less variability, etc.) if model performance is similar, *even if* a hypothesis test would choose the larger model
- We expect model performance (deviance, AUC, etc.) to be better on training data than on a new test set

Data splitting

How should we assess and compare model performance if we can’t sample new data (e.g., in the dengue scenario)?

- Randomly divide available data into two groups: training and test
  - E.g. 70% training, 30% test
- Fit the model on the training sample
- Evaluate the model on the test sample

Downsides of train/test splits

- We get less data for training
- Performance measure depends on the (random) split

Alternative: cross-validation

Cross validation

- Divide data into k groups (*folds*)
- For each fold $i = 1, \ldots, k$:
  - Train model on the remaining $k - 1$ folds
  - Evaluate on fold i
- Average performance across the k folds

Key take-aways

- Don’t choose a model based solely on training performance
  - Will bias towards more complex models
- Train/test splits and cross-validation give better estimates of model performance

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

1. Start with a small model
2. Consider each unused variable; add the variable which most improves the optimality criterion
3. Repeat Step 2 until the optimality criterion can no longer be improved

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