Review: Training/Test/Validation split and Variable section

SML201: Introduction to Data Science, Spring 2019

Michael Guerzhoy

Training, Validation, and Test sets

- Training set: the set we use to fit the model
- Test set: the set we do not touch until the very end. The test set should ideally look just like new data would
- Performance on the test set approximates the performance on new data, since we don’t use the test data in any other way

Validation set

- Suppose we are considering using one of several models
  - Cannot select the model that produces the smallest cost on the training set
    - Overfitting means performance on the training set is not always representative of performance on new data
  - Cannot select the model that produces the smallest cost on the test set
    - If we try many models on the test set and select the one that produces the smallest cost, we cannot expect that the model will do as well on new data as on our test set, since we selected the model using the specific test set
- Try all the models on the validation set, select the model that works best. Then evaluate the model on the test set to get an unbiased estimate of how well we'll do on new data

Variable selection

- The problem of selecting which variables to put in the model
- Can think of choosing between several different models

$$y \approx a_0 + a_1$$
$$y \approx a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3$$
$$y \approx a_0 + a_1 x_1 + a_3 x_3$$

….

Should we use x_3?

- We have a model

$$y \approx a_0 + a_1 x_1 + a_2 x_2$$

- We are considering adding a third variable

$$y \approx a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3$$

- This will definitely not increase the training cost
  - `lm` can always set $a_3 = 0$, so that the new model is not different from the old model. If lm decides to set a_3 to something other than 0, it's because that helped decrease the training cost
- Using x_3 can increase the validation/test cost
  - Overfitting: decreasing the cost on the training set, at the expense of increasing the cost on new data
  - We care more about the cost on new data!

Strategies for selecting which variables to use

- Go variable by variable, and decide to use the ones that decrease the validation cost/increase the validation performance
  - What you should do for Project 1
- Try all possible combinations of available variables
- Add variables one-by-one, add a new variable only if it decreases cost