3.4: Model Validation

Dr. Bean - Stat 5100

1 Why Model Validation?

Recall that there are two, distinct, goals of linear modeling and we don't always care about both at the same time:

- Inference: Is there a significant, linear relationship between X_k and Y, after accounting for the effect of a set of other X variables?
  - Example: Do students who use the tutor center see a significant positive affect to their GPA after accounting for study time and demographics?
- Prediction: Given a set of variables that are *easy* to measure, can I predict a variable that is hard to measure?
  - Use car weight (easy to measure) to predict car safety (hard to measure).

For **prediction**, there are a lot of alternatives to linear regression for which measures such as AIC, SBC, $C(p)$, and even R^2 are not relevant.

We need an *objective* way to compare the effectiveness of models with incomparable forms.

Why is the data we are using to fit our models not a fair measure of model effectiveness?

We ultimately want a model that can predict well on new data. Complex models have incentive to overfit the current data at the sacrifice of good predictions on new data.

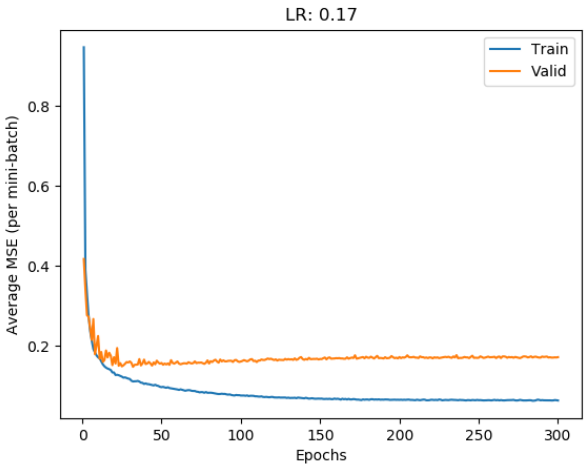

Figure 1: Comparison of accuracy on training and test sets for a neural network.

2 Validation Details

Terminology

- **Training set:** the data that is used to fit each model.
- **Test set:** data not used in model fitting that is used to compare model accuracy.
- **Validation set (optional):** If you perform too many comparisons with the test set, you run the risk of overfitting the test data. A validation set is a third set of data that is also withheld and only used to validate the best one or two models based on the test set.

Example in SAS: proc surveyselect can randomly assign observations to training and test sets.

3 Cross Validation

Whenever you have enough data, withholding a subset of the data prior to model building is ideal.

However, collecting new data can be very expensive such that creating a "test set" is not feasible.
Cross Validation: is a method that tries to estimate test set error using training data.

The process:

- Randomly separate our data into k-groups (usually five or ten).
- Treat all but one of the groups as a training set, the remaining group as a test set.
- Fit a model using the training data, predict for the test data.
- Repeat the process, each time treating a different group as the test data until all observations have a prediction.

SAS does not have an easy method for performing custom cross validation. For this purpose, we will stick to validation accuracy from a test set in our projects.

However, certain procedures use cross validation as a means of performing variable selection such as proc glmselect.

Cautions and Considerations:

- *Any* variable selection techniques or other forms of training must be included as part of the cross validation process. In other words, you can't use all of the data to select variables, then act "blind" to that same data in the model validation step.
  - The consequence of such a move is that you will likely overestimate your model's predictive capability.
  - Trying to embed variable selection into cross validation is extraordinarily difficult and not necessarily stable.
  - Check out this optional video for a more detailed explanation: `https://www.youtube.com/watch?v=r64tRyHFAJ8&list=PLOg0ngHtcqbPTlZzRHA2ocQZqB1D_qZ5V&index=23`
- The more groups you create, the more models you must fit, which can get computationally expensive.

- Too many groups makes it hard to estimate the true "test set" error.
  - Less groups, more bias, less variance in the test set error estimation. Try to select a number of cross validation groups that balance the bias and variance (usually five or ten groups).
- Check out chapter 5 in this book for more details on cross validation and other forms of model validation: `http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf`