Chapter 11:

Predictive modeling:

Two classes/groups in machine learning:

Supervised learning: modeling a specific response variable as a function of some explanatory variables (e.g., regression)

Unsupervised learning: finding grouping in data where there is no clear response variable. The outcome in not measured in unsupervised learning, it is a search for unmeasured features of the data. This is in contrast to supervised learning, like regression, where there is a clear response/outcome variable.

Classification models have a categorical response, unlike regression which has a quantitative output.

Logistic regression: converts explanatory variables or features into a predicted probability. Models that output a binary outcome, (0,1), are known as classifiers.

For model evaluation, there is a bias-variance tradeoff. A more complex model will have less bias but more variance.

To navigate this tradeoff there is what is called regularization to balance between bias and variance. Two examples include ridge regression and the lasso.

Overfitting occurs if the model does well on the training dataset, but not the testing dataset. It is fitting too much to the noise in the data, and not the signal.

Training: The dataset you build your model with

Testing: Evaluate your model against data that it has not seen

2-fold Cross validation: Build your model with X, then test against Y. Then reverse the roles.

Measuring prediction error for quantitative responses:

RMSE is most common. Penalizes big misses. Same units as predicted (y) variable

MAE (mean abs. error): Similar to RMSE but does not penalize big misses as much

Correlation: linear association. Can also do rank correlation that looks at whether the ys are in the same order.

Coefficient of determination (R-squared)