HGDS 200 - Foundations of Data-Driven Analysis 2018/2019 - Handout 2

12/17/2018 - 1/7/2019

Name: Connor Voglewede

The following should be completed after reading Chapter 2 (A short tour of the predictive modeling process) of Applied Predictive Modeling.

1. In the “fuel economy” example, the concept of a linear and non-linear relationship between variables are introduced. Can you think of examples of some linear and non-linear relationships you would expect to see in data at Red Ventures?
   1. I wouldn’t presume to expect any shape of relationships between variables and outcomes we typically model for. However, one example of clearly non-linear relationship I have encountered is conversion rate by sales center agent tenure. After a time of rapid improvement during onboarding and the first 3 months, sales performance levels out and is flat over time. A linear relationship example would be between sales agent compensation and Red Ventures revenue. In fact, at last I checked-in, it’s a nearly perfectly linear relationship because the analysts have designed that way.
2. A decision tree model is a simple and popular type of model. In the case of a 1-dimensional feature space, the model has the form



where bi,ci are real numbers for all i.

If you already know what a decision tree is, convince yourself that it can be written in this way in the

1-dimensional setting.

Would you describe this as a linear or non-linear model?

1. This is a non-linear model made up of several segments. If b were plotted against f(x) in the xy-plane, I’d expect f(x) to look like a series of step functions as x increases between the bi cutoff points.
2. We were introduced to the notion of RMSE in this section. Does RMSE make sense for binary classification tasks? (That is, tasks where the label for each data point is 0 or 1) How else might you quantify performance of a classification model?
   1. RMSE doesn’t make as much sense because you lose the interpretability of how much the model *misses* by, on average. For binary classification models, like logistic regression, using a classification matrix to identify false positive/negatives rates would be helpful, as would evaluating the value of the area under the ROC (Receiving operating characteristic) curve to help understand performance across all possible classification cut-off points.