HW1 For Applied Data Mining
STAT W 3026-4026
Spring 2016
Columbia University

Robin Lee
rcl2136
QMSS MA

February 15, 2016

Abstract

This assignment is on the chapter 4 of Applied Predictive Modeling. The topics are over-fitting and model-tuning.

1 Exercise 4.1

Consider the music genre data set described in Sect. 1.4. The objective for these data is to use the predictors to classify music samples into the appropriate music genre.

1.1 What data splitting method(s) would you use for these data? Explain.

The outcomes of the music genre data are not balanced. Also the samples are not indepent. The sample size is not too small -"there were 12,495 music samples for which 191 characteristics were determined."

Since the sample size is large enough, I'd use simple 10-fold cross validation. Since the objective doesn't involve classifying newer music, I'd use random approach.

1.2 Using tools described in this chapter, provide code for implementing your approach(es).

```
> library(caret)
> cvSplit = createFolds(data$outcome, k=10, returnTrain = TRUE)
```

2 Exercise 4.3

Partial least squares (Sect. 6.3) was used to model the yield of a chemical manufacturing process (Sect. 1.4). The data can be found in the AppliedPredictiveModeling package and can be loaded using

```
> library(AppliedPredictiveModeling)
> data(ChemicalManufacturingProcess)
```

The objective of this analysis is to find the number of PLS components that yields the optimal R2 value (Sect. 5.1). PLS models with 1 through 10 components were each evaluated using five repeats of 10-fold cross-validation and the results are presented in the following table:

2.1 Using the "one-standard error" method, what number of PLS components provides the most parsimonious model?

The model with 3 components.

2.2 Compute the tolerance values for this example. If a 10% loss in R2 is acceptable, then what is the optimal number of PLS components?

The numerically optimal value is 54.5%. Anything above 49.05% works. I would choose 2 components (50%).

2.3 Several other models (discussed in Part II) with varying degrees of complexity were trained and tuned and the results are presented in Fig. 4.13. If the goal is to select the model that optimizes R2, then which model(s) would you choose, and why?

Random forest. It has the highest R2

2.4 Prediction time, as well as model complexity (Sect. 4.8) are other factors to consider when selecting the optimal model(s). Given each modelâĂŹs prediction time, model complexity, and R2 estimates, which model(s) would you choose, and why?

I would choose SVM, because it has the second to highest R2 and a low prediction time. SVM is not too complex.