

Statistical Learning, Spring 11

Homework exercise 4

Due date: 1 June 2011

1. Categorical splitting algorithm for CART

Prove the favorable property mentioned in ESL 9.2.4: if we are splitting on a categorical variable X_j with q values and looking for the optimal split in terms of either squared error reduction (for regression) or Gini index (for 2-class classification), then an optimal split out of the possible 2^{q-1} splits is always one of the $q - 1$ splits defined by:

- Sort the q groups by their average response: $\bar{y}_{(1)} \leq \bar{y}_{(2)} \leq \dots \leq \bar{y}_{(q)}$.
- Consider only splits along this sequence, i.e., ones which divide according to whether $X_j \in \{g_{(1)}, g_{(2)}, \dots, g_{(k)}\}$ for some $k < q$.

Hint: Prove this by negation, showing that you can improve a split that does not comply with this condition by “switching” values of X_j between the splits.

2. Playing around with trees

Run a variety of tree-based algorithms on our competition data and show their performance. Compare:

- Small tree without pruning
- Large tree without pruning
- Large tree after pruning with 1-SE rule
- Bagging small trees (100 iterations)
- Bagging large trees (100 iterations)

Do this under five-fold cross validation on our competition training set, and use the results of the five different folds to calculate confidence intervals for performance. Plot all the results in a reasonable way (e.g. using `boxplot()`) and comment on them. Explain your choices of “small” and “large”.

Hints: a. Start early since bagging may take a while to run. b. Use as a basis the code from class which implements much of this.

3. **ESL 7.5: Degrees of freedom of Nearest Neighbors** Prove that in the standard i.i.d error model (which the book calls “additive error”), the effective degrees of freedom of k -NN with N observations is N/k .
4. **ESL 7.8 (7.9 in 2nd edition): Trying out model selection methods**
 The use of BIC is optional.
 Tip: for all-subset modeling in R, you can use the function `leaps()` in the package of the same name, which you may need to download from the CRAN repository¹ and install.
5. **Questions about Yehuda Koren’s presentation on 4 May (available on class homepage)**
 - (a) Consider the “collaborative filtering” model on slide 25. Assume we want to use this method on Netflix data. Explain why either an intercept or overall standardization of the r_{ui} ’s is required rather than using the raw ratings in $\{1, 2, 3, 4, 5\}$ (you may find it useful to think about the vectors p, q , as embedded points as on slide 18).
 - (b) On slide 40 we see better prediction for “active” users. This can conceivably be due to two distinct reasons: the nature of the users (“low bias”) or the use of more information to better model active users (“low variance”). Explain briefly the two options and suggest how the two can be differentiated with a simple modeling exercise.
 - (c) Slides 35, 36 give two views of temporal behavior. How can 35 can be used to explain 36?
 - * Extra credit: Some claims have been made that the results on Slide 36 make no sense, because the grade for each movie tends to decrease over time, rather than increase. Assuming that this is true, an overall pattern like in slide 36 can still occur, despite the apparent contradiction. Can you suggest an explanation?

¹<http://cran.r-project.org/web/packages/>