STA414/2104: Practice Problems 8

March, 2018

These practice problems are not for credit. Students may complete independently, in groups, or however they like. Treat these as representative of what might be on the tests.

Questions involving coding may be completed in any language. If you would like help regarding your language of choice from the course team, use R or Python. There will not be any *code*-related questions on the tests, but you will be asked about comuptational algorithms.

1. Suppose we have a training set with features $\mathbf{x} \in \mathbb{R}^p$, and we wish to fit a random forest. This involves choosing m, the number of features to consider at each split. One option is to cross validate for m directly. Another way: suppose that we thought we actually had $q < p$ truly predictive features, and we wanted to control for the probability that we select at least one predictive feature at each split.
   (a) Evaluate
   $$P(\text{choose at least one predictive feature as a candidate for a split})$$
   as a function of m, p, q.
   (b) Use this to derive the expected number of splits that have no predictive features- this is a statistic that can be used as a proxy for how noisy the resulting forest will be. *Hint*: start by deriving the expected number of splits in the entire forest- you have to make assumptions.
   (c) Discuss how you might go about determining q if faced with this problem on a real life dataset.