hw2

Rob McCulloch

January 21, 2019

Problem 1

Using the training data (the first 4169 observations), the tables below give the counts for how often `Adult` and `Age` are in the documents. The tables with `1` are for the `Ham` data and the tables with 0 are for the `spam` data.

```
      smsAdult1
smsAge1   No  Yes
    No  3598    2
    Yes    5    0
       smsAdult0
smsAge0  No Yes
    No  549   3
    Yes  12   0
```

As in the notes, we will always use observed frequencies to estimate probabilities.

(a)

Using the tables check that the simple frequency estimate of check $p(age = yes|ham)$ =.00138 as in the notes.

(b)

Use the table and the Naive Bayes assumption to estimate $p(ham|adult = no, age = yes)$.

(c)

Use the table to estimate $p(ham|adult = no, age = yes)$ *without* assuming `Age` and `Adult` are independent given y=ham/spam.

(d)

What happens if we try to estimate $p(ham|adult = yes, age = yes)$ without the Naive Bayes assumption?

Problem 2

See the file do-nb-oos-loop.R on the webpage.

In this file we randomly do several train test splits for various choices of the tuning parameter "frequent words".

The file also illustrates uses a simple approach to parallel computing in `R`.

Play around with the file. Does increasing the cutoff from 5 (used in the notes) help?

In the notes we used 1 for the tuning parameter `laplace`. Use the oos (out-of-sample) loop approach to see if using a bigger value helps.