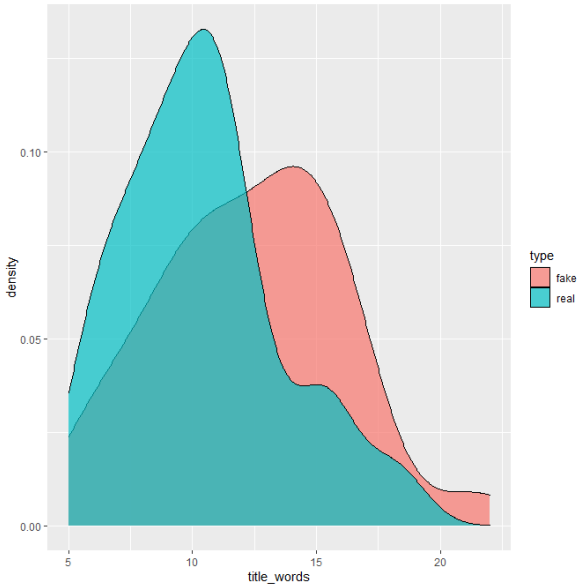
STAT 535 – HW 6 Example Solutions – Spring 2022

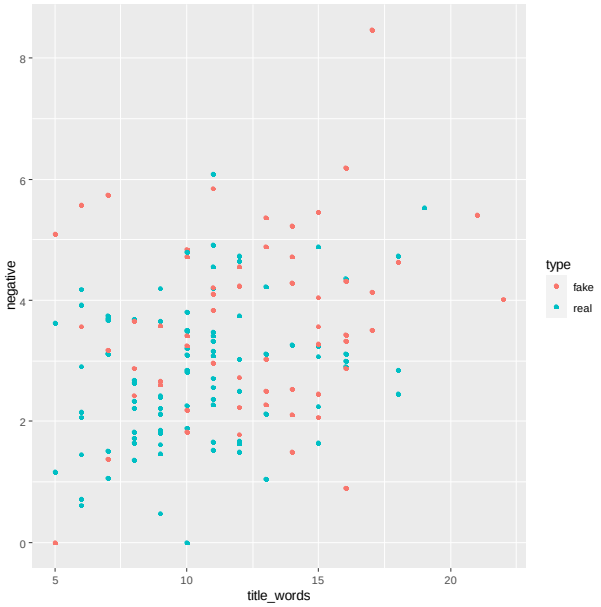
Exercise 14.2)
a) Naive Bayes, since logistic regression only works for a response with 2 categories, not 3.
b) Both, since the response has 2 categories (naive Bayes will work for 2 or more categories)
c) Both, since the response has 2 categories

Exercise 14.3)
a) Naive Bayes works for classification when the response has 2 OR MORE categories. Also, it is computationally simple, requiring no MCMC.
b) Naive Bayes assumes predictors act independently, which may not match reality. As implemented in the e1071 package, it also assumes that any continuous predictors are normally distributed, which may not be accurate.

Exercise 14.6)



It appears that fake news articles tend to have more words in their titles than real news articles do.



It appears that real news articles tend to have fewer words in their titles and be less negative (note the real news articles tend to fall in the lower left corner of the plot).

Exercise 14.7)

The posterior probability of the article being fake is 0.878 and of being real is 0.122, so we will classify it as “fake”.

Exercise 14.8) Cross-validated confusion matrices (numbers may vary slightly)

a)

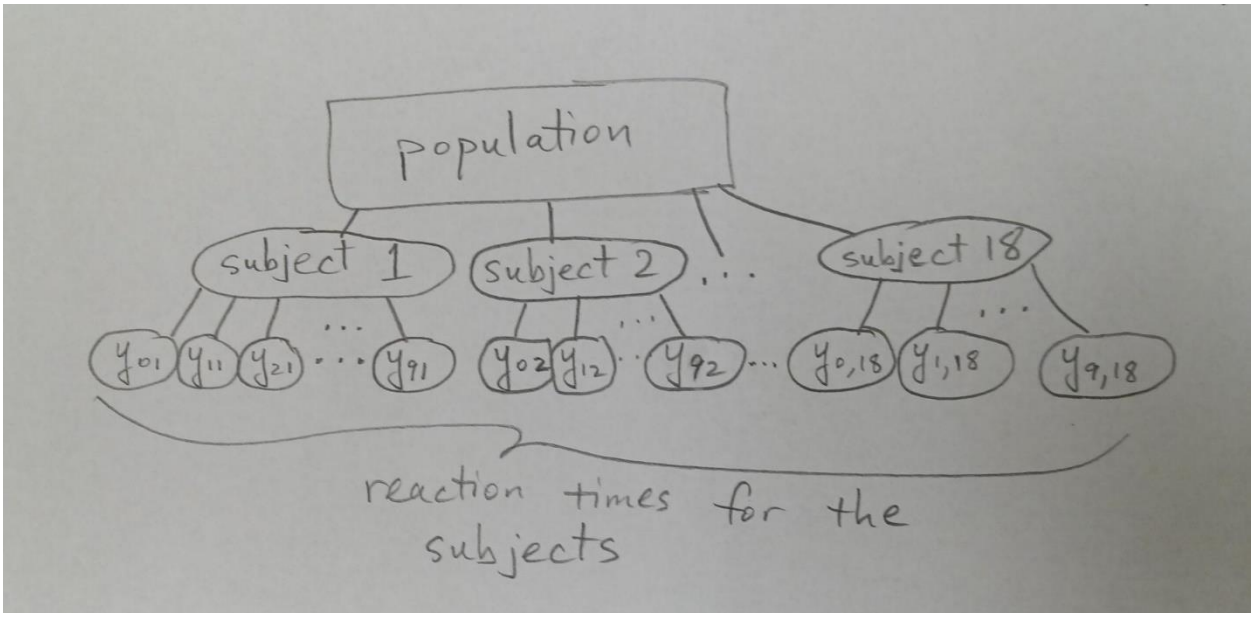
```
> cv_model_1$cv
 type        fake         real
 fake 26.67% (16) 73.33% (44)
 real  2.22%  (2) 97.78% (88)
```

c)

```
> cv_model_2$cv
 type        fake         real
 fake 36.67% (22) 63.33% (38)
 real 14.44% (13) 85.56% (77)
>
> cv_model_3$cv
 type        fake         real
 fake 41.67% (25) 58.33% (35)
 real 20.00% (18) 80.00% (72)
>
> cv_model_4$cv
 type        fake         real
 fake 48.33% (29) 51.67% (31)
 real 14.44% (13) 85.56% (77)
```

We see when an article is really fake, using Model 4 (which uses all three predictors) classifies it as fake 48.8% of the time, better in this respect than any of the other models.

Exercise 15.3)



Exercise 15.5)
The complete-pooling model assumes all the measured reaction times are independent (but reaction times within the same subject will be correlated, since a subject's reaction times will have some similarity to each other). It also assumes all the subjects share the same mean reaction time, but the mean reaction times may differ across subjects in reality (since some subjects may be naturally quicker or slower reactors).

Exercise 15.7)
With the no-pooling model, we cannot use information from any other subject to estimate a particular subject's mean reaction time (it may be useful to use other subjects' data for this, especially if the number of measurements on a particular subject is small). Also, we cannot use the no-pooling model to generalize, i.e., to predict the reaction time for a subject who wasn't in the original sample.

Exercise 16.3)
a) The multiple typing times for each typist are not independent; they are correlated. Accounting for the grouping structure allows each typist to have his or her own distribution, rather than assuming that all the measurements are a single sample from the population.
b) μ_j is the mean typing time for typist j
μ is the global mean of all the μ_j's, which can be interpreted as the overall mean typing time for the broader population of typists.
σ_y measures the variation in typing times within a particular typist (the within-group variation).
σ_μ measures the variability in the μ_j's, how different the mean times for the various typists are (the between-group variation).