NaiveBayes: Analysis of Race on COVID-19 Conditional Death Probabilities

Collin Fabian

3/21/2021

We must first load in our data

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.6.2
```

```
patient_df <- read.csv('patient_cases_by_race.csv')
```

```
patient_df$hosp_yn <- factor(patient_df$hosp_yn)
head(patient_df, n = 10)
```

```
##       sex age_group hosp_yn icu_yn death_yn medcond_yn     Race
## 1  Female       0-9      No     No       No         No    White
## 2    Male       0-9      No     No       No         No Hispanic
## 3    Male       0-9      No     No       No         No Hispanic
## 4    Male       0-9     Yes     No       No        Yes Hispanic
## 5  Female       0-9      No     No       No         No    White
## 6  Female       0-9     Yes     No       No        Yes    White
## 7  Female       0-9      No     No       No         No Hispanic
## 8  Female       0-9      No     No       No         No Hispanic
## 9  Female       0-9     Yes     No       No        Yes Hispanic
## 10 Female       0-9     Yes    Yes      Yes        Yes    Black
```

```
train.index <- sample(c(1:dim(patient_df)[1]), dim(patient_df)[1]*0.6)
train.df <- patient_df[train.index, ]
valid.df <- patient_df[-train.index, ]
```

Now, we can use the Naive Bayes Classifier for Discrete Predictors to calculate the probability of death based on race.

```
race.nb <- naiveBayes(Race ~ death_yn, data = train.df)
race.nb$tables$death_yn
```

```
##                                 death_yn
## Y                                       No        Yes
##   AmericanIndian/NativeIslander  0.86937431 0.13062569
##   Asian                          0.93865350 0.06134650
```

```
##   Black                           0.90916347 0.09083653
##   Hispanic                        0.97160939 0.02839061
##   NativeHawaiian/PacificIslander 0.96091015 0.03908985
##   Other                           0.93805031 0.06194969
##   White                           0.93274712 0.06725288
```

As we can see, the Races ranked in order from highest percentage of death to lowest is:
1. AmericanIndian/NativeIslander
2. Black
3. White
4. Asian
5. Other
6. NativeHawaiian/PacificIslander
7. Hispanic

Now, lets use Naive Bayes again to get P(Hosp_yn | Race & death_yn = Yes). This will tell us the probability that a patient who died had access to a hospital given their Race.

```
race.nb2 <- naiveBayes(Race ~ hosp_yn, data = subset(train.df, death_yn=='Yes'))
race.nb2$tables$hosp_yn
```

```
##                                 hosp_yn
## Y                                     No        Yes
##   AmericanIndian/NativeIslander  0.05882353 0.94117647
##   Asian                           0.05296610 0.94703390
##   Black                           0.04745167 0.95254833
##   Hispanic                        0.05417186 0.94582814
##   NativeHawaiian/PacificIslander 0.07462687 0.92537313
##   Other                           0.12182741 0.87817259
##   White                           0.18267284 0.81732716
```

From our results, out of all the patients who died from COVID-19, the probabilities of getting treatment at a hospital based on Race are listed in order from lowest to highest:
1. White
2. Other
3. AmericanIndian/NativeIslander
4. NativeHawaiian/PacificIslander
5. Hispanic
6. Black
7. Asian