A Detailed Example of Ethnicity Prediction

Le Wang

Our problem

We would like to predict a voter's race. What is our procedure?

Our problem

We would like to predict a voter's race. What is our procedure?

Step 1. Figure out the distribution

$$\Pr[\text{Race}]$$

Our problem

We would like to predict a voter's race. What is our procedure?

Step 1. Figure out the distribution

$$\Pr[\text{Race}]$$

$$\Pr[\text{Race} = \text{white}], \Pr[\text{Race} = \text{black}], \Pr[\text{Race} = \text{hispanic}]$$
$$\Pr[\text{Race} = \text{asian}], \Pr[\text{Race} = \text{others}]$$

Our problem

We would like to predict a voter's race. What is our procedure?

Step 1. Figure out the distribution

$$\Pr[\text{Race}]$$

$$\Pr[\text{Race} = \text{white}], \Pr[\text{Race} = \text{black}], \Pr[\text{Race} = \text{hispanic}]$$
$$\Pr[\text{Race} = \text{asian}], \Pr[\text{Race} = \text{others}]$$

Step 2. Pick the most likely value, in this case, always **white**

More information

Now, suppose that I tell you this voter's last name is PIEDRA, what is your prediction for this person's race?

More information

Now, suppose that I tell you this voter's last name is PIEDRA, what is your prediction for this person's race?

Step 1. Figure out the conditional distribution, i.e., the distribution for this subgroup

$$\Pr[\text{Race} \mid \text{Surname}]$$

More information

Now, suppose that I tell you this voter's last name is PIEDRA, what is your prediction for this person's race?

Step 1. Figure out the conditional distribution, i.e., the distribution for this subgroup

$$\Pr[\text{Race} \mid \text{Surname}]$$

In this specific case:

$$\begin{aligned}
&\Pr[\text{race} = \text{white} &&\mid \text{surname} = \text{PIEDRA}] \\
&\Pr[\text{race} = \text{black} &&\mid \text{surname} = \text{PIEDRA}] \\
&\Pr[\text{race} = \text{hispanic} &&\mid \text{surname} = \text{PIEDRA}] \\
&\Pr[\text{race} = \text{asian} &&\mid \text{surname} = \text{PIEDRA}] \\
&\Pr[\text{race} = \text{others} &&\mid \text{surname} = \text{PIEDRA}]
\end{aligned}$$

More information

Now, suppose that I tell you this voter's last name is PIEDRA, what is your prediction for this person's race?

Step 1. Figure out the conditional distribution, i.e., the distribution for this subgroup

$$\Pr[\text{Race} \mid \text{Surname}]$$

In this specific case:

$$\begin{aligned}
&\Pr[\text{race} = \text{white} &&\mid \text{surname} = \text{PIEDRA}] \\
&\Pr[\text{race} = \text{black} &&\mid \text{surname} = \text{PIEDRA}] \\
&\Pr[\text{race} = \text{hispanic} &&\mid \text{surname} = \text{PIEDRA}] \\
&\Pr[\text{race} = \text{asian} &&\mid \text{surname} = \text{PIEDRA}] \\
&\Pr[\text{race} = \text{others} &&\mid \text{surname} = \text{PIEDRA}]
\end{aligned}$$

Step 2. Pick the most likely value and use it as our prediction.

How to do it?

```
head(FLvoters)
```

```
##       surname county VTD age gender  race
## 1      PIEDRA    115  66  58      f white
## 2        LYNCH    115  13  51      m white
## 4     LATHROP     115  80  54      m white
## 5      HUMMEL     115   8  77      f white
## 6 CHRISTISON     115  55  49      m white
## 7       HOMAN     115  84  77      f white
```

```r
subset <- subset(FLvoters, surname == "PIEDRA")
prop.table(table(subset$race))
```

```
##
##    asian    black hispanic   native    other    white
##        0        0        0        0        0        1
```

But in practice, if we do not have such information but have access to the census names data, what should we do?

```
subset <- subset(cnames, surname == "PIEDRA")
subset
```

```
##       surname count pctwhite pctblack pctapi pctaian pct2
## 8610  PIEDRA  3518     6.71      1.19   0.43    0.14
##       pctothers
## 8610       0.28
```

How to evaluate our results?

For the white sample, we know that

$$\begin{aligned}
&\Pr[\text{race} = \text{white} &&\mid \text{surname} = \text{PIEDRA}] \\
&\Pr[\text{race} = \text{black} &&\mid \text{surname} = \text{PIEDRA}] \\
&\Pr[\text{race} = \text{hispanic} &&\mid \text{surname} = \text{PIEDRA}] \\
&\Pr[\text{race} = \text{asian} &&\mid \text{surname} = \text{PIEDRA}] \\
&\Pr[\text{race} = \text{others} &&\mid \text{surname} = \text{PIEDRA}]
\end{aligned}$$

$$\max(\cdots) = \Pr[\text{race} = \text{white} \mid \text{surname} = \text{PIEDRA}]$$

```
subset
```

```
##      surname count pctwhite pctblack pctapi pctaian pct2
## 8610  PIEDRA  3518     6.71     1.19   0.43    0.14
##      pctothers
## 8610      0.28
```

```
max(subset[, c("pctwhite","pctblack","pctapi","pctaian","pc
```

```
## [1] 91.39
```

```
max(subset[, c("pctwhite","pctblack","pctapi","pctaian","pc
```

```
## [1] FALSE
```

Further Information

What if I tell you one more piece of information that this voter's last name is PIEDRA living in county 115 and VTD 66

Further Information

What if I tell you one more piece of information that this voter's last name is PIEDRA living in county 115 and VTD 66

Step 1. Figure out the conditional distribution, i.e., the distribution for this subgroup

$$\Pr[\text{Race} \mid \text{Surname, County, VTD}]$$

Further Information

What if I tell you one more piece of information that this voter's last name is PIEDRA living in county 115 and VTD 66

Step 1. Figure out the conditional distribution, i.e., the distribution for this subgroup

$$\Pr[\text{Race} \mid \text{Surname, County, VTD}]$$

In this specific case:

$$\begin{aligned}
&\Pr[\text{race} = \text{white} &&\mid \text{surname} = \text{PIEDRA}, \text{county} = 115, \text{VTD} = 66] \\
&\Pr[\text{race} = \text{black} &&\mid \text{surname} = \text{PIEDRA}, \text{county} = 115, \text{VTD} = 66] \\
&\Pr[\text{race} = \text{hispanic} &&\mid \text{surname} = \text{PIEDRA}, \text{county} = 115, \text{VTD} = 66] \\
&\Pr[\text{race} = \text{asian} &&\mid \text{surname} = \text{PIEDRA}, \text{county} = 115, \text{VTD} = 66] \\
&\Pr[\text{race} = \text{others} &&\mid \text{surname} = \text{PIEDRA}, \text{county} = 115, \text{VTD} = 66]
\end{aligned}$$

Further Information

What if I tell you one more piece of information that this voter's last name is PIEDRA living in county 115 and VTD 66

Step 1. Figure out the conditional distribution, i.e., the distribution for this subgroup

$$\Pr[\text{Race} \mid \text{Surname, County, VTD}]$$

In this specific case:

$$\begin{aligned}
&\Pr[\text{race} = \text{white} &&\mid \text{surname} = \text{PIEDRA}, \text{county} = 115, \text{VTD} = 66] \\
&\Pr[\text{race} = \text{black} &&\mid \text{surname} = \text{PIEDRA}, \text{county} = 115, \text{VTD} = 66] \\
&\Pr[\text{race} = \text{hispanic} &&\mid \text{surname} = \text{PIEDRA}, \text{county} = 115, \text{VTD} = 66] \\
&\Pr[\text{race} = \text{asian} &&\mid \text{surname} = \text{PIEDRA}, \text{county} = 115, \text{VTD} = 66] \\
&\Pr[\text{race} = \text{others} &&\mid \text{surname} = \text{PIEDRA}, \text{county} = 115, \text{VTD} = 66]
\end{aligned}$$

Step 2. Pick the most likely value and use it as our prediction.

```r
subset <- subset(FLvoters, surname == "PIEDRA" & county ==
prop.table(table(subset$race))
```

```
##
##    asian    black hispanic   native    other    white
##        0        0        0        0        0        1
```

What if this dataset is not available? We have access to only census names data and Florida census data?

```
head(cnames)
```

```
##     surname   count pctwhite pctblack   pctapi  pctaian p
## 1     SMITH 2376206 73.34267 22.21778 0.399960 0.849915
## 2   JOHNSON 1857160 61.55000 33.80000 0.420000 0.910000
## 3  WILLIAMS 1534042 48.52000 46.72000 0.370000 0.780000
## 4     BROWN 1380145 60.71607 34.54345 0.410041 0.830083
## 5     JONES 1362755 57.69000 37.73000 0.350000 0.940000
## 6    MILLER 1127803 85.80142 10.40896 0.419958 0.629937
##    pctothers
## 1  2.479752
## 2  2.730000
## 3  2.790000
## 4  2.690269
## 5  2.790000
## 6  1.939806
```

```
head(FLcensus)
```

Lets piece together information from two different datasets

$$\Pr[\text{race}|\text{surname,residence}]$$
$$= \frac{\Pr[\text{surname}|\text{race,residence}]\,\Pr[\text{race}|\text{residence}]}{\Pr[\text{surname}|\text{residence}]}$$

Lets piece together information from two different datasets

$$\Pr[\text{race}|\text{surname},\text{residence}]$$
$$= \frac{\Pr[\text{surname}|\text{race},\text{residence}]\,\Pr[\text{race}|\text{residence}]}{\Pr[\text{surname}|\text{residence}]}$$

$$\Pr[\text{race} = \text{white}|\text{surname}= \text{PEIDRA},\text{county}=115,\ \text{VTD}=66]$$
$$= \frac{\overbrace{\Pr[\text{surname}=\text{PIEDRA} \mid \text{race}=\text{white},\text{county}=115,\ \text{VTD}=66]}^{\textbf{Term A}}}{\underbrace{\Pr[\text{surname}=\text{PIEDRA}|\text{county}=115,\ \text{VTD}=66]}_{\textbf{Term C}}}$$
$$\cdot \overbrace{\Pr[\text{race}=\text{white}|\text{county}=115,\ \text{VTD}=66]}^{\textbf{Term B}}$$

Term B

```
subset <- subset(FLcensus, county == 115 & VTD == 66)
subset
```

```
##       county VTD total.pop     white       black  hispanic
## 8048     115  66       5699 0.7638182 0.06281804 0.1363397
```

```
B <- subset[,"white"]
```

Let me also calculate this for other races

```
B.black <- subset[,"black"]
B.hisp <- subset[,"hispanic"]
B.api <- subset[,"api"]
B.others <- subset[,"others"]
```

$$\Pr[\text{race} = \text{white} | \text{surname}= \text{PEIDRA}, \text{county}=115, \text{VTD}=66]$$

$$= \frac{\overbrace{\Pr[\text{surname}=\text{PIEDRA} \mid \text{race}=\text{white}, \text{county}=115, \text{VTD}=66]}^{\textbf{Term A}}}{\underbrace{\Pr[\text{surname}=\text{PIEDRA} | \text{county}=115, \text{VTD}=66]}_{\textbf{Term C}}}$$

$$\cdot \overbrace{\Pr[\text{race}=\text{white} | \text{county}=115, \text{VTD}=66]}^{\textbf{Term B}}$$

$$
\begin{aligned}
&\Pr[\text{surname=PIEDRA}|\text{county=115, VTD=66}] \\
&= \Pr[\text{surname=PIEDRA} \mid \text{race=white,county=115, VTD=66}] \cdot \Pr[\text{race=white} \mid \text{county=115, VTD=66}] \\
&+ \Pr[\text{surname=PIEDRA} \mid \text{race=black,county=115, VTD=66}] \cdot \Pr[\text{race=black} \mid \text{county=115, VTD=66}] \\
&+ \Pr[\text{surname=PIEDRA} \mid \text{race=hispanic,county=115, VTD=66}] \cdot \Pr[\text{race=hispanic} \mid \text{county=115, VTD=66}] \\
&+ \Pr[\text{surname=PIEDRA} \mid \text{race=asisan,county=115, VTD=66}] \cdot \Pr[\text{race=asian} \mid \text{county=115, VTD=66}] \\
&+ \Pr[\text{surname=PIEDRA} \mid \text{race=others,county=115, VTD=66}] \cdot \Pr[\text{race=others} \mid \text{county=115, VTD=66}]
\end{aligned}
$$

Term A

This is my **Term A**:

$$\Pr[\text{surname=PIEDRA} \mid \text{race=white,county=115, VTD=66}]$$

Term A

This is my **Term A**:

$$\Pr[\text{surname=PIEDRA} \mid \text{race=white,county=115, VTD=66}]$$

I will make the following **conditional independence assumption**

$$\Pr[\text{surname=PIEDRA} \mid \text{race=white,county=115, VTD=66}]$$
$$= \Pr[\text{surname=PIEDRA} \mid \text{race=white}]$$

Or similarly for every other racial category

$$\Pr[\text{surname=PIEDRA} \mid \text{race=black,county=115, VTD=66}]$$
$$= \Pr[\text{surname=PIEDRA} \mid \text{race=black}]$$

In general

$$\Pr[\text{surname} \mid \text{race, residence}] = \Pr[\text{surname} \mid \text{race}]$$

Does it make sense? Or, when will it be violated?

In general

$$\Pr[\text{surname} \mid \text{race, residence}] = \Pr[\text{surname} \mid \text{race}]$$

Does it make sense? Or, when will it be violated?

Conditional independence implies that once we know a voter's race, her residence location does not give us any additional information about her surname.

In general

$$\Pr[\text{surname} \mid \text{race, residence}] = \Pr[\text{surname} \mid \text{race}]$$

Does it make sense? Or, when will it be violated?

Conditional independence implies that once we know a voter's race, her residence location does not give us any additional information about her surname.

There is NO strong geographical concentration of certain surnames in Florida within a racial categroy. This will certainly be violated in the Chinese context.

$$\Pr[\text{surname}=\text{PIEDRA} \mid \text{race}=\text{white},\text{county}=115,\ \text{VTD}=66]$$
$$= \Pr[\text{surname}=\text{PIEDRA} \mid \text{race}=\text{white}]$$

$$\Pr[\text{race} = \text{white} | \text{surname} = \text{PEIDRA}, \text{county}=115, \text{VTD}=66]$$

$$= \frac{\overbrace{\Pr[\text{surname}=\text{PIEDRA} \mid \text{race}=\text{white}, \text{county}=115, \text{VTD}=66]}^{\textbf{Term A}}}{\underbrace{\Pr[\text{surname}=\text{PIEDRA} | \text{county}=115, \text{VTD}=66]}_{\textbf{Term C}}}$$

$$\cdot \overbrace{\Pr[\text{race}=\text{white} | \text{county}=115, \text{VTD}=66]}^{\textbf{Term B}}$$

$$\Pr[\text{race} = \text{white} | \text{surname= PEIDRA,county=115, VTD=66}]$$

$$= \frac{\overbrace{\Pr[\text{surname=PIEDRA} \mid \text{race=white}]}^{\textbf{Term A}}}{\underbrace{\Pr[\text{surname=PIEDRA} | \text{county=115, VTD=66}]}_{\textbf{Term C}}}$$

$$\cdot \overbrace{\Pr[\text{race=white} | \text{county=115, VTD=66}]}^{\textbf{Term B}}$$

Where do we have this information?

Where do we have this information?

$$\Pr[\text{surname}=\text{PIEDRA} \mid \text{race}=\text{white}]$$

Where do we have this information?

$$\Pr[\text{surname}{=}\text{PIEDRA} \mid \text{race}{=}\text{white}]$$

$$= \frac{\overbrace{\Pr[\text{race}{=}\text{white} \mid \text{surname} = \text{PIEDRA}]}^{\textbf{Term A.1}}}{\underbrace{\Pr[\text{race} = \text{white}]}_{\textbf{Term A.2}}} \cdot \overbrace{\Pr[\text{surname}{=}\text{PIEDRA}]}^{\textbf{Term A.3}}$$

```
A.1 <- cnames[which(cnames$surname=="PIEDRA"), "pctwhite"]/
A.1
```

```
## [1] 0.0671
```

$$\Pr[\text{surname}{=}\text{PIEDRA} \mid \text{race}{=}\text{white}]$$

$$\Pr[\text{surname}{=}\text{PIEDRA} \mid \text{race}{=}\text{white}]$$

$$= \frac{\overbrace{\Pr[\text{race}{=}\text{white} \mid \text{surname} = \text{PIEDRA}]}^{\textbf{Term A.1}}}{\underbrace{\Pr[\text{race} = \text{white}]}_{\textbf{Term A.2}}} \cdot \overbrace{\Pr[\text{surname}{=}\text{PIEDRA}]}^{\textbf{Term A.3}}$$

Term A.2

```
head(FLcensus)
```

```
##   county VTD total.pop     white     black   hispanic
## 1      1  46      4482 0.6137885 0.1871932 0.12092816 0.
## 2      1  31      5470 0.6742230 0.1201097 0.10859232 0.
## 3      1  55      3165 0.6315956 0.2085308 0.10110585 0.
## 4      1  49      3458 0.7111047 0.1917293 0.05060729 0.
## 5      1  56      1937 0.6319050 0.2426433 0.07279298 0.
## 6      1  60      3103 0.4166935 0.4785691 0.05156300 0.
```

```
race.prop <- apply(FLcensus[,c("white", "black", "api", "hi
                   2,
                   weighted.mean,
                   weights = FLCensus$total.pop)
A.2 <- race.prop["white"]
A.2
```

```
##     white
## 0.6045159
```

$$\Pr[\text{surname}{=}\text{PIEDRA} \mid \text{race}{=}\text{white}]$$

$$\Pr[\text{surname}{=}\text{PIEDRA} \mid \text{race}{=}\text{white}]$$

$$= \frac{\overbrace{\Pr[\text{race}{=}\text{white} \mid \text{surname} = \text{PIEDRA}]}^{\textbf{Term A.1}}}{\underbrace{\Pr[\text{race} = \text{white}]}_{\textbf{Term A.2}}} \cdot \overbrace{\Pr[\text{surname}{=}\text{PIEDRA}]}^{\textbf{Term A.3}}$$

Term A.3

```
head(cnames)
```

```
##     surname   count pctwhite pctblack   pctapi  pctaian p
## 1     SMITH 2376206 73.34267 22.21778 0.399960 0.849915
## 2   JOHNSON 1857160 61.55000 33.80000 0.420000 0.910000
## 3 WILLIAMS 1534042 48.52000 46.72000 0.370000 0.780000
## 4     BROWN 1380145 60.71607 34.54345 0.410041 0.830083
## 5     JONES 1362755 57.69000 37.73000 0.350000 0.940000
## 6    MILLER 1127803 85.80142 10.40896 0.419958 0.629937
##    pctothers
## 1  2.479752
## 2  2.730000
## 3  2.790000
## 4  2.690269
## 5  2.790000
## 6  1.939806
```

```r
total.count<- sum(cnames$count)
A.3 <- cnames[which(cnames$surname == "PIEDRA"),"count"]/to
```

Term A

```
A <- A.1*A.3/A.2
A
```

```
##         white
## 1.612791e-06
```

Repeat this process for all other racial groups

```
A.1.black <- cnames[which(cnames$surname=="PIEDRA"), "pctbl
A.1.hisp <- cnames[which(cnames$surname=="PIEDRA"), "pcthis
A.1.api <- cnames[which(cnames$surname=="PIEDRA"), "pctapi"
A.1.others <- cnames[which(cnames$surname=="PIEDRA"), "pcto

A.2.black <- race.prop["black"]
A.2.hisp <- race.prop["hispanic"]
A.2.api <- race.prop["api"]
A.2.others <- race.prop["others"]

A.black <- A.1.black*A.3/A.2.black
A.hisp <- A.1.hisp*A.3/A.2.hisp
A.api <- A.1.api*A.3/A.2.api
A.others <- A.1.others*A.3/A.2.others



A.black
```

```
##       black
```

$$\Pr[\text{race} = \text{white} | \text{surname}= \text{PEIDRA}, \text{county}=115, \text{VTD}=66]$$

$$= \frac{\overbrace{\Pr[\text{surname}=\text{PIEDRA} \mid \text{race}=\text{white}, \text{county}=115, \text{VTD}=66]}^{\textbf{Term A}}}{\underbrace{\Pr[\text{surname}=\text{PIEDRA} | \text{county}=115, \text{VTD}=66]}_{\textbf{Term C}}}$$

$$\cdot \overbrace{\Pr[\text{race}=\text{white} | \text{county}=115, \text{VTD}=66]}^{\textbf{Term B}}$$

Term C

$$
\begin{aligned}
&\Pr[\text{surname=PIEDRA}|\text{county=115, VTD=66}] \\
&= \Pr[\text{surname=PIEDRA} \mid \text{race=white,county=115, VTD=66}] \cdot \Pr[\text{race=white} \mid \text{county=115, VTD=66}] \\
&+ \Pr[\text{surname=PIEDRA} \mid \text{race=black,county=115, VTD=66}] \cdot \Pr[\text{race=black} \mid \text{county=115, VTD=66}] \\
&+ \Pr[\text{surname=PIEDRA} \mid \text{race=hispanic,county=115, VTD=66}] \cdot \Pr[\text{race=hispanic} \mid \text{county=115, VTD=66}] \\
&+ \Pr[\text{surname=PIEDRA} \mid \text{race=asisan,county=115, VTD=66}] \cdot \Pr[\text{race=asian} \mid \text{county=115, VTD=66}] \\
&+ \Pr[\text{surname=PIEDRA} \mid \text{race=others,county=115, VTD=66}] \cdot \Pr[\text{race=others} \mid \text{county=115, VTD=66}]
\end{aligned}
$$

```
residence <- subset(FLcensus, county== 115 & VTD ==66)
residence
```

```
##      county VTD total.pop     white      black  hispanic
## 8048    115  66      5699 0.7638182 0.06281804 0.136339
```

```
C <- A*residence["white"] + A.black*residence["black"] + A
C
```

```
##              white
## 8048 9.905006e-06
```

```
A*B/C
```

```
##           white
## 8048 0.1243693
```

```r
cond.prob <- c(A*B/C, A.black*B.black/C, A.hisp*B.hisp/C, A
names(cond.prob) <- c("white", "black", "hispanic", "api",
cond.prob
```

```
## $white
## [1] 0.1243693
##
## $black
## [1] 0.007865472
##
## $hispanic
## [1] 0.8589315
##
## $api
## [1] 0.005162938
##
## $others
## [1] 0.003670794
```