ADM delivery1 - Naive Bayes Classification

Diego Garcia-Olano

March 11, 2014

Abstract

This paper is the beginning of an analysis of the data compiled for the Blue Islands project[1]. The data set is comprised of 2012 US presidential election data results by county along with socio economic data available from Measure Of America, a project of the Social Science Research Council. We will construct a Naive Bayses classifier on the dataset of 4587 counties and determine how well we can predict whether a county voted for the Democrat or Republican candidate. We will do so first on a subset of the data excluding the state within which a county is located, and the next we will do the classification on the full data set.

1 Classification of data without States

```
First we load in combined election/socio-economic data and give shorter variable names.
> edata <-read.csv("/opt/htdocs/electionmap/combineddata.csv", stringsAsFactors=FALSE)
> colnames(edata) =
    c("id","fips","id2", "county","dem","rep","isdem","year",
    "lesshs","hs","leastbach","graduate","enrollp","earnings","educationi",
    "incomei","white","afric","nativ","asian","other","latin",
    "povu6","povo65","popul","preschl","belowpov","gini","childpov",
    "wmanage","wservice","wsales","wfarm","wconstruct","wtransport")

Next we make our target variable categorical and remove extraneous variables for initial analysis.
> edata$isdem <- as.factor(edata$isdem)  #isdem will be the target variable
> levels(edata$isdem) <- c("no","yes")
> edata.clean <- edata
> edata <- edata[,c(7,9:35)]

We then look at the summary statistics for the 4587 counties and 28 variables.
> summary(edata)
```

Table 1: Summary Statistics

	isdem	lesshs	hs	leastbach	graduate	enrollp	earnings
1	no :2846	Min. : 0.70	Min. :47.90	Min. : 3.70	Min. : 0.000	Min. : 37.80	Min. : 5559
2	yes:1741	1st Qu.: 9.90	1st Qu.:81.30	1st Qu.:14.90	1st Qu.: 4.500	1st Qu.: 74.00	1st Qu.:23236
3		Median :13.00	Median :87.00	Median :19.80	Median : 6.600	Median : 76.80	Median :25744
4		Mean :14.97	Mean :85.03	Mean :22.46	Mean : 8.127	Mean : 76.63	Mean :26888
5		3rd Qu.:18.70	3rd Qu.:90.10	3rd Qu.:29.80	3rd Qu.:10.900	3rd Qu.: 79.70	3rd Qu.:29774
6		Max. :52.10	Max. :99.30	Max. :71.00	Max. :40.600	Max. :100.00	Max. :59672
	educationi	incomei	white	afric	nativ	asian	other
1	Min. :0.590	Min. :0.000	Min. : 2.80	Min. : 0.000	Min. : 0.000	Min. : 0.000	Min. : 0.000
2	1st Qu.:3.690	1st Qu.:3.260	1st Qu.:74.60	1st Qu.: 0.400	1st Qu.: 0.200	1st Qu.: 0.300	1st Qu.: 1.100
3	Median :4.400	Median :3.970	Median :89.70	Median : 1.300	Median : 0.300	Median : 0.600	Median : 1.400
4	Mean :4.502	Mean :4.132	Mean :82.06	Mean : 6.696	Mean : 1.245	Mean : 1.385	Mean : 1.664
5	3rd Qu.:5.320	3rd Qu.:4.980	3rd Qu.:95.10	3rd Qu.: 6.100	3rd Qu.: 0.500	3rd Qu.: 1.400	3rd Qu.: 1.900
6	Max. :9.400	Max. :9.800	Max. :99.20	Max. :85.400	Max. :94.100	Max. :43.000	Max. :35.000

[1] http://www.diegoolano.com/electionmap/.

	latin	povu6	povo65	popul	preschl	belowpov	gini
1	Min. : 0.00	Min. : 0.00	Min. : 0.0	Min. : 82	Min. : 0.00	Min. : 0.00	Min. :0.2070
2	1st Qu.: 1.30	1st Qu.: 14.50	1st Qu.: 7.3	1st Qu.: 16829	1st Qu.: 36.10	1st Qu.:10.00	1st Qu.:0.4110
3	Median : 2.70	Median : 21.90	Median : 9.2	Median : 42366	Median : 44.40	Median :13.00	Median :0.4300
4	Mean : 6.95	Mean : 23.00	Mean :10.4	Mean : 154854	Mean : 44.75	Mean :14.17	Mean :0.4321
5	3rd Qu.: 6.75	3rd Qu.: 29.85	3rd Qu.:12.4	3rd Qu.: 131219	3rd Qu.: 53.90	3rd Qu.:17.20	3rd Qu.:0.4510
6	Max. :95.70	Max. :100.00	Max. :47.4	Max. :9818605	Max. :100.00	Max. :53.50	Max. :0.6450
7					NA's :7		

	childpov	wmanage	wservice	wsales	wfarm	wconstruct	wtransport
1	Min. : 0.00	Min. :11.10	Min. : 0.00	Min. : 0.00	Min. : 0.000	Min. : 1.80	Min. : 1.20
2	1st Qu.:12.50	1st Qu.:27.00	1st Qu.:15.50	1st Qu.:21.40	1st Qu.: 0.400	1st Qu.: 8.80	1st Qu.:10.60
3	Median :17.60	Median :30.80	Median :17.30	Median :23.20	Median : 1.000	Median :10.30	Median :13.90
4	Mean :19.26	Mean :32.04	Mean :17.45	Mean :23.12	Mean : 1.854	Mean :10.85	Mean :14.69
5	3rd Qu.:24.65	3rd Qu.:36.50	3rd Qu.:19.00	3rd Qu.:25.30	3rd Qu.: 2.400	3rd Qu.:12.50	3rd Qu.:17.80
6	Max. :64.60	Max. :67.30	Max. :38.50	Max. :39.00	Max. :38.300	Max. :30.00	Max. :39.50

We note that there are 7 NA values for the "preschl" variable which corresponds to the enrollment ratio of children aged 3 and 4 in preschools within a given county. The missing data is as follows:

```
> edata[which(is.na(edata$preschl)),c(1,6,7,10:15,18:20,22)]
```

Table 2: counties with missing preschool enrollment data

	isdem	enrollp	earnings	white	afric	nativ	asian	other	latin	popul	preschl	belowpov	childpov
256	no	60.60	20529	95.20	0.30	0.60	0.10	0.80	2.90	712	NA	7.70	0.00
2589	no	69.40	17708	95.40	0.00	0.20	0.50	2.50	1.40	1179	NA	16.90	34.60
3210	no	73.70	22672	96.30	0.00	1.10	0.20	1.30	1.20	1321	NA	18.90	31.10
3752	no	76.20	31750	84.10	0.00	0.30	0.20	0.60	14.80	641	NA	4.30	9.20
3866	no	88.40	27778	20.70	0.20	1.40	0.20	0.70	76.70	416	NA	14.90	0.00
3870	no	68.60	37500	84.60	0.00	0.30	0.00	1.40	13.60	286	NA	0.00	0.00
3886	no	100.00	50556	73.20	0.00	4.90	0.00	0.00	22.00	82	NA	0.00	0.00

These are Republican counties with populations ("popul") smaller than the median composed of mainly white and latino voters. We impute the missing data, and observe that the new values appear to be tied to earnings, and the reported school enrollment percentage per county.

```
> library(mice)
> edata.post <- complete(mice(edata,m=1))
> edata.post[which(is.na(edata$preschl)),c(1,6,7,10:15,18:20,22)]
```

Table 3: imputed missing preschool enrollment data

	isdem	enrollp	earnings	white	afric	nativ	asian	other	latin	popul	preschl	belowpov	childpov
256	no	60.60	20529	95.20	0.30	0.60	0.10	0.80	2.90	712	17.30	7.70	0.00
2589	no	69.40	17708	95.40	0.00	0.20	0.50	2.50	1.40	1179	13.30	16.90	34.60
3210	no	73.70	22672	96.30	0.00	1.10	0.20	1.30	1.20	1321	34.70	18.90	31.10
3752	no	76.20	31750	84.10	0.00	0.30	0.20	0.60	14.80	641	40.40	4.30	9.20
3866	no	88.40	27778	20.70	0.20	1.40	0.20	0.70	76.70	416	42.40	14.90	0.00
3870	no	68.60	37500	84.60	0.00	0.30	0.00	1.40	13.60	286	60.10	0.00	0.00
3886	no	100.00	50556	73.20	0.00	4.90	0.00	0.00	22.00	82	69.90	0.00	0.00

We then construct our Naive Bayes classifier on our clean data set, using 75 percent of it for training and the remaining 25 percent to test it.

```
> library(e1071)
> trainingdata <- edata.post[1:3440,]
> classifier<-naiveBayes( trainingdata[,2:28], trainingdata$isdem )
```

We note our training data contains 2150 republican counties (isdem = no) and 1290 democrat ones.
We use our naive classifier to predict on our training data to see how it works.

```
>results <- table(predict(classifier,trainingdata[,2:28]), trainingdata$isdem, dnn=list('predicted','actual'))
              actual
   predicted   no  yes
         no  1785  554
         yes  365  736
```

Our Naive Bayes classifier correctly labeled 1785 republican counties and mislabeled 554 democrat ones as republican. On the other end, it correctly labeled 736 democrat counties and mislabled 365 republican ones as democrat. Looking at the proportions, we see that the classifier identifies Republican counties correct 76.3 percent of the time and Democrat ones 66.8 percent of the time.

```
> diag(prop.table(results, 1))
         no       yes
   0.7631466 0.6684832
```

Overall, our classifier is 73.28 percent accurate on the training data without state info.

```
> sum(diag(prop.table(results))) #.7328 ...
```

We now try our classifier on the unseen test data.

```
> testdata <- edata.post[3441:4587,]
> results2 <- table(predict(classifier, testdata[,2:28]), testdata$isdem, dnn=list('predicted','actual'))
              actual
   predicted  no  yes
          no  622 224
         yes  74  227
> diag(prop.table(results2, 1))
         no       yes
   0.7352246 0.7541528
```

Our Naive Bayes classifier does slightly worse for republican predictions, but substantially better on democrat ones. The total percent of correct predictions on the test data is 74 percent which is almost identical to the training data.

```
> sum(diag(prop.table(results2))) #.7402
```

2 Classification of data which contains state information

We now combine state information back into our prior data set as follows:

```
> statedata <-read.csv("/opt/htdocs/electionmap/smaller2012president.csv", na.strings = '')
> tmp = cbind(statedata$CountyNumber,statedata$FIPSCode,
              as.character(statedata$CountyName),
              as.character(statedata$StatePostal) )
> colnames(tmp) = c("CountyNumber","FIPSCode","CountyName", "State")
> head(tmp)
```

	CountyNumber	FIPSCode	CountyName	State
1	1	0	Alaska	AK
2	2001	2000	Alaska	AK
3	1	0	Alabama	AL
4	1001	1001	Autauga	AL
5	1002	1003	Baldwin	AL
6	1003	1005	Barbour	AL

We note from above that there are entries with FIPSCode = 0 and CountyNumber = 1, which correspond to vote totals per state, and thus we remove them before merging.

```
> statetotals = tmp[which(tmp[,2] == 0),]
> tmp2 <- tmp[which(tmp[,2] != 0),]
> states <- as.data.frame(tmp2)

> merged = merge(edata.clean,states,by.x="county",by.y="CountyNumber")
> merged$preschl = edata.post$preschl

Now that our data is merged, we can construct a new Naive Bayes classifier and
run our predictions on training and then test data.
> mtrainingdata <- merged[1:3440,]
> mclassifier<-naiveBayes( mtrainingdata[,2:37], mtrainingdata$isdem )

Prediction using testdata.
> mresults <- table(predict(mclassifier,mtrainingdata[,2:37]),
                    mtrainingdata$isdem,dnn=list('predicted','actual'))
            actual
predicted   no  yes
      no  1959  225
      yes  191 1065

This time we do way better, and get 89.6 percent correct hits for Republicans
and 84.8 percent for democrats.
> diag(prop.table(mresults, 1))
       no       yes
0.8969780 0.8479299

Our total success rate is then 87.9 on training data.
> sum(diag(prop.table(mresults))) #.879 ...

Now we try our classifier on the unseen test data.
> mtestdata <- merged[3441:4587,]
> mresults2 <- table(predict(mclassifier, mtestdata[,2:37]), mtestdata$isdem, dnn=list('predicted','actual'))
          actual
predicted  no yes
      no  672 131
      yes  24 320

On test data, we still perform well however this time our prediction for Republican success drops to 83.6%
and our Democrat prediction goes up to 93 %
> diag(prop.table(mresults2, 1))
         no       yes
  0.8368618 0.9302326

Our final total is 86.48% accurracy with a Naive Bayes classifier
> sum(diag(prop.table(mresults2))) # 0.8648
```

3 What factors are important for classifying counties?

```
First, lets do an average profile of counties which voted republican vs democrat.
> merged.continous.vars = merged[,c(5,6,9:35)]

Republican counties:
> reps = merged.continous.vars[as.vector(which(merged$isdem=="no")),]
> avgrep = colMeans(reps)
> rep = round(avgrep,2)  #average rep

Democrat counties:
> dems = merged.continous.vars[as.vector(which(merged$isdem=="yes")),]
```

```
> avgdem = colMeans(dems)
> dem = round(avgdem,2) #average dem

> diff = rep - dem
> profiles = t(data.frame(republicans=rep,democrats=dem,diff=diff))
```

Table 4: Average Republican/Democrat County Profiles and their difference

	dem	rep	lesshs	hs	leastbach	graduate	enrollp	earnings	educationi	incomei
republicans	7316.42	11983.14	16.28	83.72	19.08	6.42	75.80	25895.94	4.14	3.89
democrats	23772.37	14178.30	12.83	87.17	28.00	10.92	77.97	28509.75	5.09	4.53
diff	-16455.95	-2195.16	3.45	-3.45	-8.92	-4.50	-2.17	-2613.81	-0.95	-0.64
	white	afric	nativ	asian	other	latin	povu6	povo65	popul	
republicans	83.75	5.73	1.18	0.90	1.53	6.91	23.95	10.78	92911.22	
democrats	79.30	8.27	1.35	2.18	1.88	7.01	21.44	9.80	256110.71	
diff	4.45	-2.54	-0.17	-1.28	-0.35	-0.10	2.51	0.98	-163199.49	
	preschl	belowpov	gini	childpov	wmanage	wservice	wsales	wfarm	wconstruct	wtransport
republicans	42.38	14.61	0.43	20.14	30.25	17.18	22.87	2.18	11.56	15.96
democrats	48.63	13.44	0.44	17.83	34.96	17.89	23.52	1.33	9.68	12.62
diff	-6.25	1.17	-0.01	2.31	-4.71	-0.71	-0.65	0.85	1.88	3.34

```
We can plot this table as follows:
> profilesdf = profiles[1:2,c(3:29)]  #do not include dem/rep columns
> proportional_df = round(prop.table(profilesdf,2),4)
> proportional_df = proportional_df[,order(proportional_df[1,],decreasing=TRUE)]
> library(reshape)
> library(ggplot2)
> dfm <- melt(proportional_df)
> names(dfm) = c("counties_who_voted_for","socio_economic_indicators","proportional_value")
> dfm$socio_economic_indicators =
      factor(dfm$socio_economic_indicators, unique(dfm$socio_economic_indicators))

> p <- theme_update(axis.text.x = theme_text(angle = 90, hjust = .25, colour="black", size = 20),
        axis.text.y = theme_text(angle = 0, hjust = 1, colour="black", size = 18),
        panel.grid.major = theme_line(colour = "gray90"),
        panel.grid.minor = theme_blank(), panel.background = theme_blank(),
        axis.ticks = theme_blank(), legend.position = "right", legend.text=theme_text(size=20))

> p <- ggplot(dfm, aes(socio_economic_indicators, fill=counties_who_voted_for, y=proportional_value)) +
      geom_bar(position="dodge",stat="identity") +
      coord_cartesian(ylim=c(0,1)) +
      coord_flip() +
      scale_fill_manual(values = c("blue","red")) +
      guides(fill=FALSE)

Figure 1 shows that Democrat counties comparable to Republican counties, in decreasing order, have:
    larger population sizes,
    larger percentage Asian American population,
    higher percentage of people with graduate degrees,
    higher percentage of people with at least a bachelors degree,
    larger percentage African American population,
    larger percentage of "other" ethnicity,
    larger income index,
    larger amount of workers within "Managerial/Professional" category,
    larger amount of preschool enrollment,
    larger percentage Native American population,
    and slightly larger mean earnings, "service" industry workers, high school attainment, etc.
```

Figure 1: Socio-Economic indicators for counties won by Democrats(blue)/Republican(red)