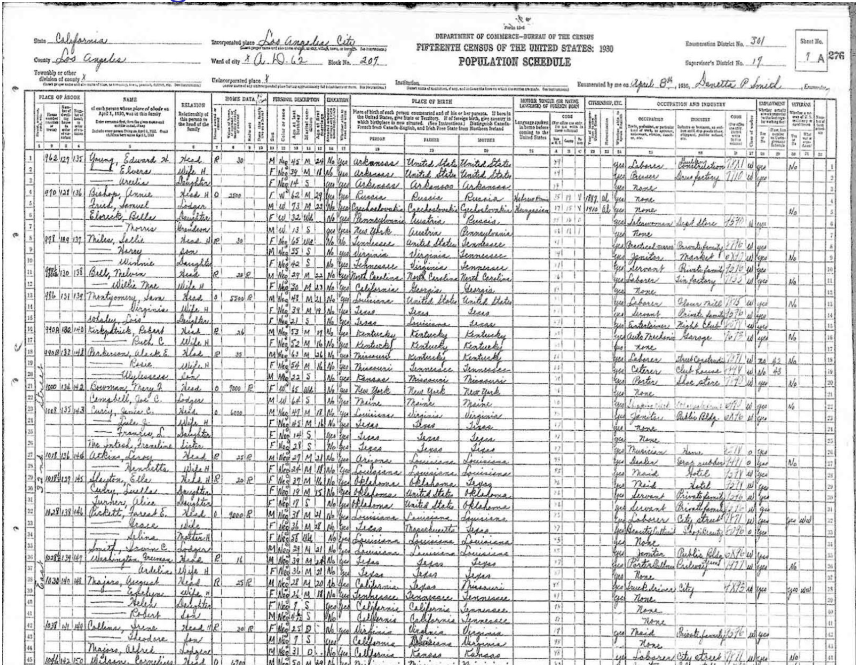
Letter Classification

Matt Oehler and Jackson Curtis

Stat 536

October 26, 2018

Problem Background


DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS
FIFTEENTH CENSUS OF THE UNITED STATES: 1930
POPULATION SCHEDULE

State: California
County: Los Angeles
Township or other division of county: X
Incorporated place: Los Angeles City
Ward of city: X A.D. 62
Block No.: 207
Unincorporated place: X
Institution:
Enumeration District No. 301
Supervisor's District No. 17
Sheet No. 7 A 276
Enumerated by me on April 8th, 1930, Danetta P. Smith, Enumerator.

Line	House no.	Dwelling	Family	Name	Relation	Sex	Race	Age	Marital	Birthplace	Father	Mother	Occupation	Industry
1	962	127	135	Young, Edward H.	Head	M	Neg	45	M	Arkansas	United States	United States	Laborer	Building Construction
2				Elvera	Wife H	F	Neg	39	M	Arkansas	United States	United States	Presser	Dress factory
3				Ureelia	Daughter	F	Neg	14	S	Arkansas	Arkansas	Arkansas	None	
4	970	128	136	Bishop, Annie	Head H	F	W	62	M	Russia	Russia	Russia	None	
5				Fried, Samuel	Lodger	M	W	73	M	Czechoslovakia	Czechoslovakia	Czechoslovakia	None	
6				Elsrock, Belle	Daughter	F	W	32	Wd	Pennsylvania	Austria	Russia	None	
7				Morris	Grandson	M	W	13	S	New York	Austria	Pennsylvania	Saleswoman	Dept. store
8	978	129	137	Miles, Sallie	Head H	F	Neg	65	Wd	Tennessee	United States	Tennessee	Practical nurse	Private family
9				Harvey	Son	M	Neg	35	S	Virginia	Virginia	Tennessee	Janitor	Market
10				Winnie	Daughter	F	Neg	42	S	Tennessee	Virginia	Tennessee	Servant	Private family
11	982½	130	138	Bell, Melvin	Head	M	Neg	29	M	North Carolina	North Carolina	North Carolina	Laborer	Tin factory
12				Willie Mae	Wife H	F	Neg	30	M	California	Georgia	Georgia	None	
13	986	131	139	Montgomery, Sam	Head	M	Neg	48	M	Louisiana	United States	United States	Laborer	Flour mill
14				Virginia	Wife H	F	Neg	39	M	Texas	Texas	Texas	Servant	Private family
15				Ashley, Lois	Daughter	F	Neg	21	S	Texas	Louisiana	Texas	Entertainer	Night Club
16	990A	132	140	Kirkpatrick, Robert	Head	M	Neg	53	M	Kentucky	Kentucky	Kentucky	Auto Mechanic	Garage
17				Ruth C.	Wife H	F	Neg	52	M	Kentucky	Kentucky	Kentucky	None	
18	990B	133	141	Perkerson, Alexk E.	Head	M	Neg	63	M	Missouri	Kentucky	Kentucky	Laborer	Street Construction
19				Rosie	Wife H	F	Neg	54	M	Missouri	Tennessee	Tennessee	Caterer	Club house
20				Ulylessess	Son	M	Neg	22	S	Kansas	Missouri	Missouri	Porter	Shoe store
21	1000	134	142	Bowman, Mary F.	Head	F	W	65	Wd	New York	New York	New York	None	
22				Campbell, Joe C.	Lodger	M	W	68	S	Maine	Maine	Maine	[illegible]	[illegible]
23	1008	135	143	Curry, Jenie C.	Head	M	Neg	47	M	Louisiana	Virginia	Virginia	Janitor	Public Bldg.
24				Lula J.	Wife H	F	Neg	45	M	Texas	Texas	Texas	None	
25				Frances L.	Daughter	F	Neg	14	S	Texas	Texas	Texas	None	
26				McIntosh, Francline	Sister	F	Neg	28	S	Texas	Texas	Texas	Musician	Home
27	1018	136	144	Atkins, Leroy	Head	M	Neg	27	M	Arizona	Louisiana	Louisiana	Dealer	Scrap rubber
28				Henrietta	Wife H	F	Neg	24	M	Louisiana	Louisiana	Louisiana	Maid	Hotel
29	1018½	137	145	Slayton, Ella	Head H	F	Neg	37	M	Oklahoma	Oklahoma	Texas	Maid	Hotel
30				Curry, Luella	Daughter	F	Neg	19	M	Oklahoma	United States	Oklahoma	Servant	Private family
31				Turner, Alice	Daughter	F	Neg	17	S	Oklahoma	United States	Oklahoma	Servant	Private family
32	1028	138	146	Pickett, Forest E.	Head	M	Neg	38	M	Louisiana	Louisiana	Louisiana	Laborer	City street
33				Grace	Wife	F	Neg	36	M	Texas	Massachusetts	Texas	Beauty Culturist	Shop Beauty
34				Selina	Mother H	F	Neg	58	Wd	Louisiana	Louisiana	Louisiana	None	
35				Smith, Irvin C.	Lodger	M	Neg	29	M	Louisiana	Louisiana	Louisiana	Janitor	Public Bldg.
36	1028½	139	147	Washington, Freeman	Head	M	Neg	39	M	Texas	Texas	Texas	Porter Pullman	Railway Trains
37				Ardelia	Wife H	F	Neg	36	M	Texas	Texas	Texas	None	
38	1030	140	148	Majors, August	Head	M	Neg	28	M	California	Texas	Missouri	Truck driver	City
39				Edythe	Wife H	F	Neg	26	M	Tennessee	Tennessee	Tennessee	None	
40				Helen	Daughter	F	Neg	7	S	California	California	Tennessee	None	
41				Robert	Son	M	Neg	4	S	California	California	Tennessee	None	
42	1038	141	149	Collins, Irene	Head H	F	Neg	25	D	Virginia	Virginia	Virginia	Maid	Private family
43				Theodore	Son	M	Neg	7	S	California	Alabama	Virginia	None	
44				Majors, Alfred	Lodger	M	Neg	31	D	California	Kansas	Kansas	Laborer	City street
45	1046	142	150	McClane, Cornelius	Head	M	Neg	50	M	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Goals

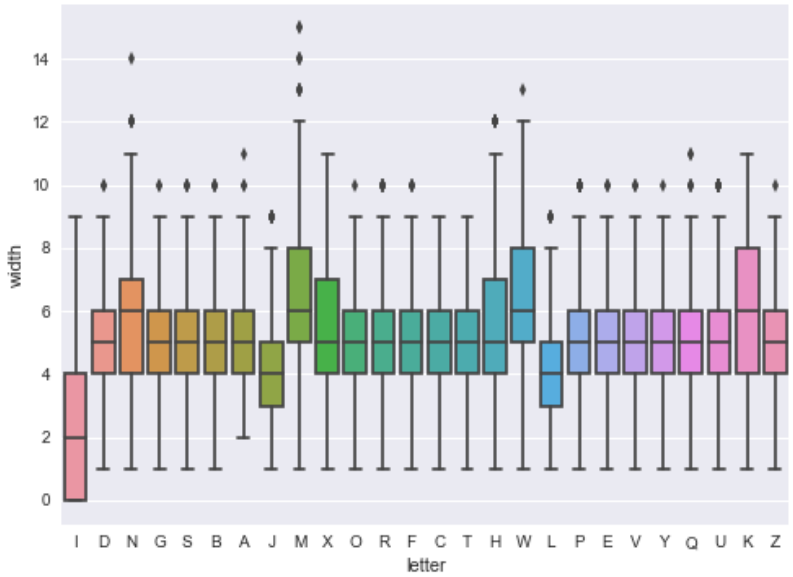
Goals of the Analysis:

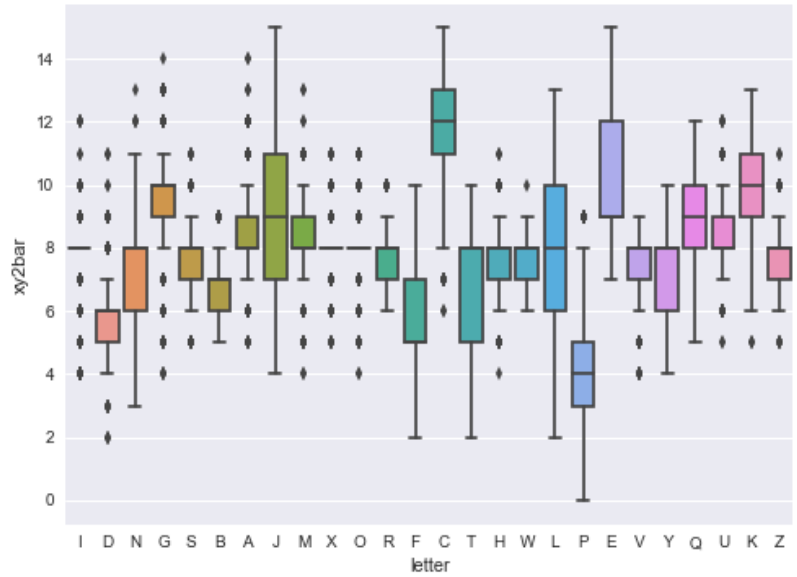
- Use the data of letter characteristics to create a model that can accurately classify handwritten letters
- Automate document transcription by using the model to predict letters given their characteristics

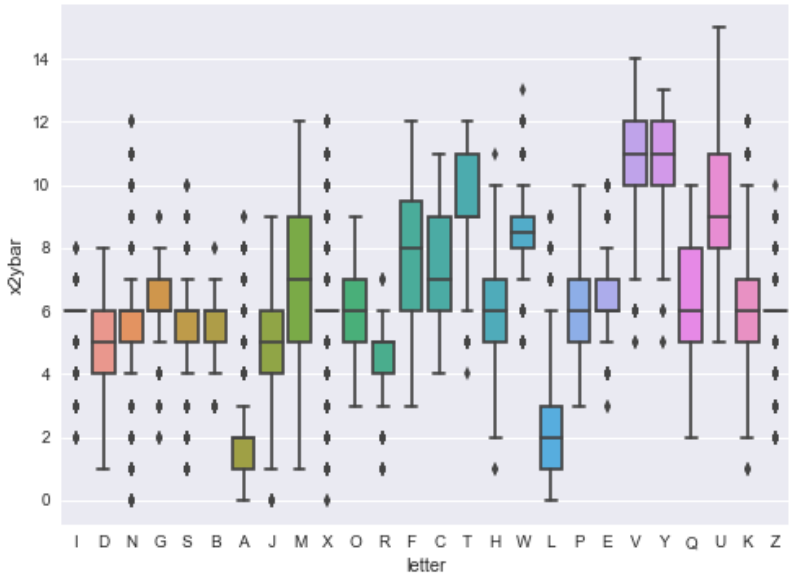
Data Description

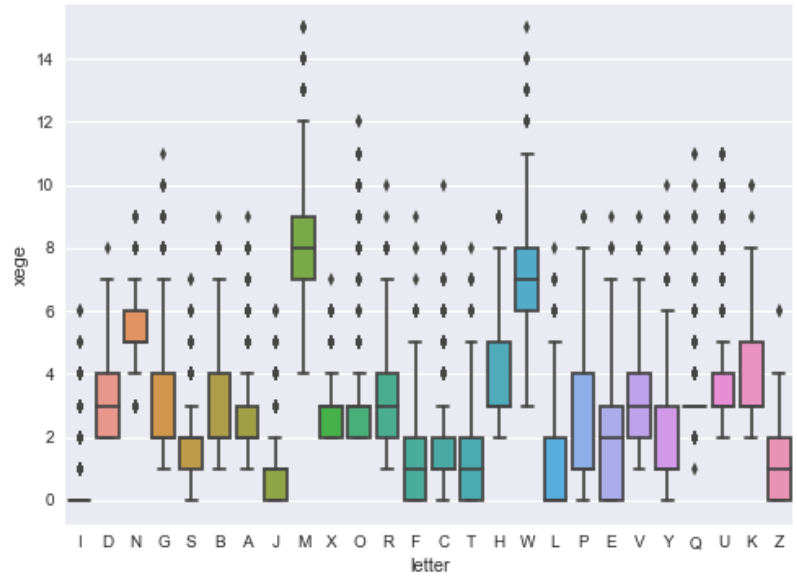
	Variable Name	Description
1	letter	capital letter (26 values from A to Z)
2	xbox	horizontal position of box (integer)
3	ybox	vertical position of box (integer)
4	width	width of box (integer)
5	high	height of box (integer)
6	pix	total # of pixels (integer)
7	xbar	mean x of pixels in box (integer)
8	ybar	mean y of pixels in box (integer)
9	x2bar	mean x variance (integer)
10	y2bar	mean y variance (integer)
11	xybar	mean x y correlation (integer)
12	x2ybr	mean of x * x * y (integer)
13	xy2br	mean of x * y * y (integer)
14	xege	mean edge count left to right (integer)
15	xegvy	correlation of x-ege with y (integer)
16	yege	mean edge count bottom to top (integer)
17	yegvx	correlation of y-ege with x (integer)

Data Description









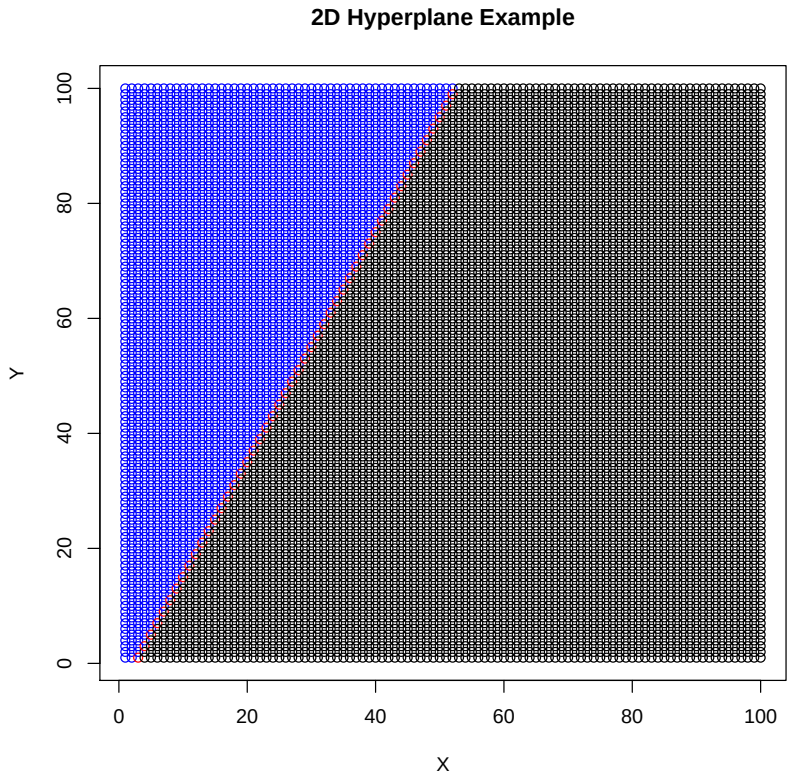
Model Selection

We explored using Support Vector Machines and Random Forests, but found that Support Vector Machines seemed to perform better. Some advantages of SVMs are:

- They perform well in high-dimensioal space (we can use all the variables
- They are computationally efficient
- They are flexible according to the parameters that you choose to define them (e.g. kernel functions)

Intro To Support Vector Machines

Support vector machines are based on the idea of dividing hyperplanes.



- Red points (the hyperplane) are where $0 = 50 + 10y - 20x$
- Blue is where $0 < 50 + 10y - 20x$
- Black is where $0 > 50 + 10y - 20x$

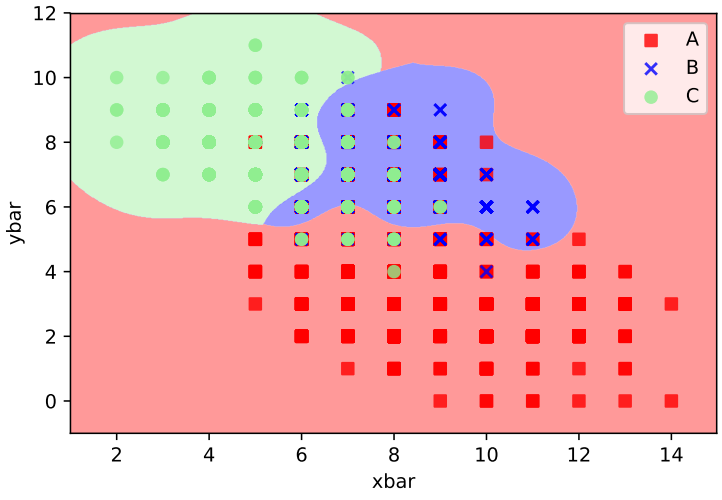
In our case, instead of 2 dimensions, we will define a plane in 16-dimensional hyperspace.

Choosing a Hyperplane

Calculate the hyperplane that separates the two classes perfectly and maximizes the minimum distance between the points and the line (this distance is called the margin).
What if we can't get perfect separation?

- Introduce slack variables
- Observations with slack greater than 0 are allowed to lie close to or on the wrong side of the plane, but we penalize each slack variable based on its distance from the margin of our SVM.
- We use a tuning parameter to set the biggest penalty we're willing to accept.

Extending to non-linear planes



- SVMs won't work if we limit ourselves to linear planes
- Basis functions allow us to remap our explanatory variables so that the result is linearly separable
- Common basis functions include polynomial, radial, and neural nets
- Small values of the tuning parameter will increase the non-linearity of the hyperplanes.

Extending to multiple classes

SVM planes split one class from another, but what about when we have 26 classes?

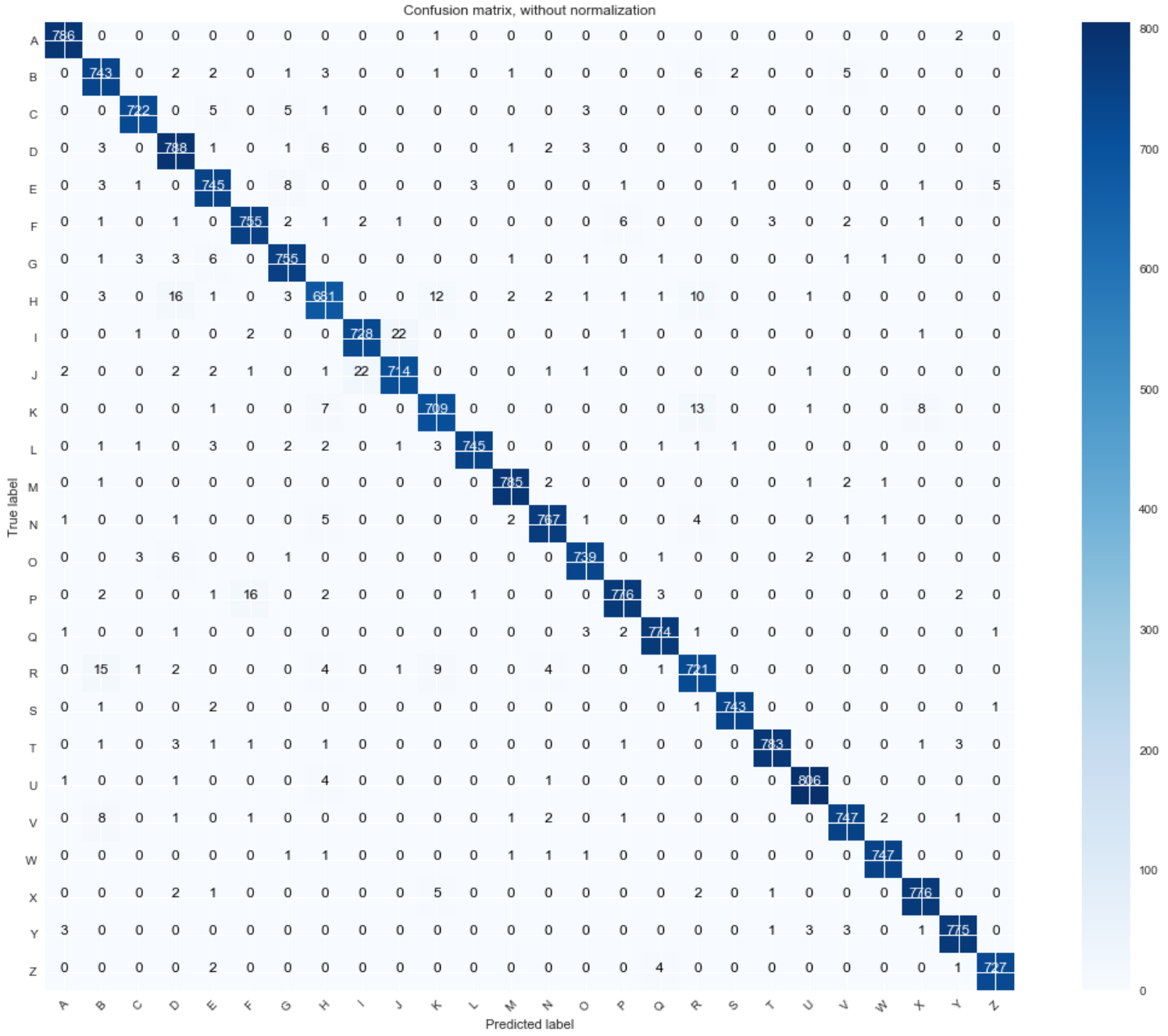
- Build a SVM using A and B data only, then A vs. C, etc
- This builds $\binom{k}{2}$ SVMs, 325 in our case.
- Tally the 325 votes, majority becomes final prediction

Parameter tuning

We performed a grid search and looked at overall accuracy for various parameter possibilities to choose:

- Kernel (we chose radial)
- Tuning parameter (11 worked best)
- Gamma–Precision of radial Gaussian distribution (setting it to auto, 1/# features, worked best).

Model Performance



Overall Accuracy

Method	Out of Sample Accuracy
Random Forest	96.70%
Support Vector Machine	97.77%

Table 1: SVMs slightly outperform Random Forests with 1000 trees and max 3 features

Letter Ranking

Accuracy	Letter		
0.9987	A	0.9801	O
0.9947	S	0.9776	D
0.9926	U	0.9766	E
0.9920	W	0.9764	V
0.9911	Y	0.9745	N
0.9905	Z	0.9726	B
0.9886	X	0.9652	F
0.9885	Q	0.9639	J
0.9874	M	0.9621	K
0.9862	T	0.9617	R
0.9842	L	0.9616	I
0.9823	C	0.9614	P
0.9819	G	0.9319	H

Table 2: Ranking of which letters are easiest to predict

Variable Importance

SVM variable importance is tough, but a rough idea can be given by looking at how far accuracy drops when the model is built and tested without that variable.

Top Five Most Important	Top Five Least Important
1. yege	1. xbox
2. x2bar	2. width
3. y2bar	3. xbar
4. xybar	4. high
5. x2ybar	5. pix

Conclusions

Goals:

- We accomplished our goals by creating a model that can accurately predict a letter given its characteristics

Shortcomings:

- Since we used an SVM, the results aren't interpretable, but that's ok since we are more interested in prediction
- Since we are working in a high-dimensional space, there isn't a good way to visualize the SVM and its decision boundaries

Future Work

- Results could be improved be ensembling several different approaches.
- Added features would probably help SVM (does well in high dimensions)
- At the extreme, neural nets could be used on pixel by pixel information