Letter Classification

Matt Oehler and Jackson Curtis

Stat 536

April 13, 2018

Problem Background

0	tents Caledo tentry Loss remains or other divident of econy	angeles	-		Los angelia 10.62	Rock No	207.	PIFTRENT	MENT OF COMMER I CENSUS OF 1 POPULATIO	THE UNITE	D STATE	BS: 193		a April 8th	Equatorism Di	frict No	7	Sheet?
	PLACE OF ASSOCI	10/2		BOME DATA	77250001, DENGEROOF	-	non, Beologiethes)	PLACE OF SURE	and a dissert to beauty which the		enante)	CIFIAND		/			meth	, Eur
	L Day See to	clearly person whose place of about an April 1, 1930, was in this family	RELATION AND AND AND AND AND AND AND AND AND AN				Flore of birth of each pers		or her parson. If bench	LUNGUISE) OF	CODE				THE REAL PROPERTY.	cos -	Shear each everyment withdeapp to websplice	
H	Topic Co.	Dier emme feit, fem für den mate mit mitter man, dass Dielet einer panne Diesen April, 702. Dass st. Mitte betrause April, 202	this person to the head of the family	Control of the Contro	Color or rece Color or rece Enrich Enrich Education Color Co	\$20 61	Plans of high of each per the Solind States, give which highlates is non Formh from Canada do Pizzon	nitrated. (See Instruction agint), and Irish Pres State FATER	1000023	Languageapokes in botto belone coming to the Dained States	For edity cas extr. In pas with in these schools Sin Garry by	Tour of least plant falls	Photos alle in	OCCUPATION THE PROPERTY OF STREET, STR	DOCULET Infacts on because, as and the sold day production oligans, politic related tio.	Dange of such	Tor copies or property of the copies or copies	To 134
H	962 (27 13	Grung Edward 26		1 1 1 10	n n n n n		- 19		10	11	A 3 C		14	ti		D 127	te p	00 21
2	192 127 13	Young, Edward 24.	Head	R 30	M Ag 45 M 29	No lice	askanose	United State			24			Laborer	Construction	7/11/10	924	No
1	TT	- areelia	Wife H. Deughter		F Nog 14 5	No year	arkennes		Writed State	-	14	-			Drughetry	7/10 4	iger	
4	910 121 13	Bishop, annie	Head H	0 2510	F 11 62 M 30	I. L.	P	Russia	arkoness	77	75 0 11	40.01	Spe	Rose	, ,			
5		Fried Janual	dodger	1	F W 62 M 29 M W 73 M 23	the fire	Contain h	Part 1	Guldentie	Helson Come		1910 6		There	-	-	-	-
6		Elneck, Belle	Acustic	1 11	F W 32 W	16 100	Hannet store :	Tuetrie.	Russis.	anguer	25 14 6	7770 83		Selemen	1.116	J2301 .	-	No.
7	-	Thorris	Brandese		M 60 13 5	ou be	Ten White	austria	Consylvania		4 11 1	-	Vau		sign since	25/11/1	Con	
8	621 189 13	Wiles, Lette	Head I	P 30	F She 65 180	16 16	Lindersee	anited States	Jennessee		41 1				Briosh front	8716 11		
10	-	Herry	gen .	1 11	M 400 35 5	No 1400	Dirainia.	Virginia	Tennessee		141 1		344	Janiter	market		line	No
11	440 Line 15	Bell, Milvin	Lught	0 20	F 168 42 S	No Eyes	Fignesses	Vicamies	Jennesses		11		teu	Lernent	Rivet land	1210 W	ac.	
	1304/120/17	- Willie True	Wile H	200	F the 30 M 23	No the	and Certine	north Cerolina			26		94	Laborer	Tin factory	7135 4	94	16
12	186 /3/ /3	1 Mostyonery, Jane	Hees	0 5700 R	Maleg HI M 21	No Ges	Celifornia	Georgia	Surgia	-	0	-	44	Zeze	' '	-		
16		- Virginia	Wile H	2705 11	F. No. 39 M 17	11 9	devisione	United States	United States			-	tru		Flour nitt		yes	No.
16		Whater Pris	South	1 11				Luisiana	Legge		71 1	-	ges		Private food		igo	
10	110A 130 14	1 tirkultrick, Robert	High	P. 36	M No. 53 M 15	M. See	backer 6	Kentucky	Kentucky		at	-	740	Enterteiner auto Trechonia	Right Club	20/1 44	40	16
	-	Buch C	Wile H	1 11	F Neg 52 M 16	No Sug	Kentucks	Kentucky	Kentucki		54 1		64	XPAL.	range	1612 18	yes	100
18	29032 132 10	1 Brown, aleck E.	Alex	P 35	M Nec 63 M 26	No Toca	Thisewil	Kentuckie	Kentuck		11 1		Lus		Shut Coreline	737/14	44 45	16
1	4-1-1-	- Resie	Wife H		F 1/2 54 M 16	No see	Thiseouri	Jeanware	Jennessee		12		Bace	Ceterer	Club house	199 4	16 45	1
20	1000 134 14		Heed	0 9000 P	M Me 22 5 F W 65 WL	No go	Gansas	Thisesusi	Thisourie		20		140	Perter	les stère	7190 4	40	1/0
72	S Town	Cempbell, De C.	Lodger	0 1000 15	M W 64 5	1/ 12	New York	Here york	newyork		4		140				/	
23	Veres 135 14	3 Curry, Senie C.	Ned.	0 6000	M No. 47 M 18	100 1940	Paine	Theine	maine		75	-	100	Dispire the	Maryalismit	off W	900	1/4
21		- Tule 2	Wife H	N LCOS				Dirginia -	Dirginia	-	0 1	-	240	Janeter	Bellio Bldg.	01/16 H	gre .	
25		- Francis S	Schulter		F Neg 28 S	Stra Stra	Luces	Jesu	Segue	-	10	-	94	Thomas.		-		
25		The intest, Francisco	Lister		F Neg 28 5	No bes	Leces	Jens	Line	-	12			Musician	J	15/1 o	41	
27	1018 136 14	atkin Leroy	Head	R 25 R	M Neg 29 M 21	No Beer	arian.	Louisiane	Louisiera		93 1		Tue	Bealer	Stare rubby		Ska	11
25	6	S Slayton, Elle	Wife H		F Neg 24 M 11	16 40	Louisiera	Louisiene	Louisiene		93		Luce	marie		15714		100
23	1 16183174 14	S Slayton, Elle	Helen	8 20 R	F Weg 37 M W.	10 40	Ofletone	Eklehema	Lynn		34		1	mid	Letel	7211 W		
31	1-1-	Juney Luellas	Saytte	H - H -	F Wag 19 M 45	No rue	Aletone	arited State	Helefone		74		lya		Private famil	1216 W	See	
22	1128 138 14	Richett, Farest E.	Red	0 9000 R	F MED 17 S M New 31 M W	No lyes	Medone	axited States	Oklaheme	-	74	-	44	dervent	Birthfamil	1216 10		
23		- Grace	idile	2000	F 100 36 M 2	at the	Lees	Lauregene	dequeiras		9	-	1500	Labour	City struck		ges	per wood
		- deline	matter A		F Ave ST UK	Ala Sec	Louisiana	Deschunte	Leges		-5	-	140	Resity Cathin	Stopkent	80960	Yes	
25		Smith Irvine C.	Loder		M No 21 4 21	No Buch	Luisiana	aberrana	Louisiana		15	-	Gu	More	011. 01.	700	-	
26	कार्याम (र		Here	P 16	M Mg 39 M 21	16 40	Ledas	Jess	Legas	-	17	-	Tu	Partir Billow	Public Feldy	DATE W	yes	16
57	4	artelia	efile H		F No 36 M 1	A6 120	Jevas	Jesus	Leves		37	1	füra		curry	1711 W	344	150
28	2 1638 HO H		Nead	R 25 R	M Acc 28 M 20	Ab de	California	teres	Priesuri		11		Tue	True driver	O.L.	4X93 M	tere	5540 kgml
22	1	- Elelyn	dife "	-	F 110 24 M 11		Tenherse	Dennusce	Tennessee		11		ac	None	7	11112 11	7	7 80
	1	Robert	den	-	F Neg 1 S	Geo Ges	Celifernia	Celifernie	Syrance.		1)		10	None				
42	1017 11 11	Collins Irene		P 30 P	MAKE 47 S	No	California	California	Syraessee	-	0	-	-	Tione	-			
	SCHOOL ST.	- Theodore		20 8	F Me LY D	sh go	Celifornia	wegizia	Virginia		11		lace	Thaid	Rivet James	1576 4	400	
43	1 1 1	Majoro, alfred	for		M 160 9 51	ass		Bioliera	dirginia		93			none				

Goals

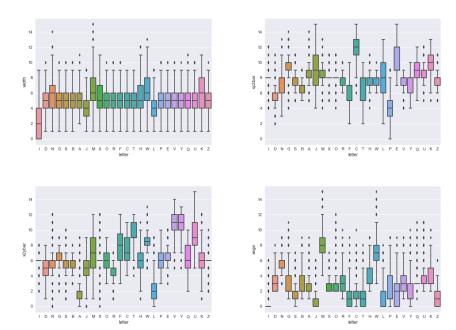
Goals of the Analysis:

- 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



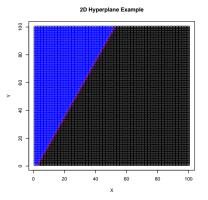
Model Selection

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.

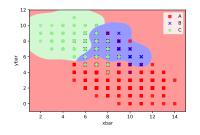
Choosing a Hyperplane

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?

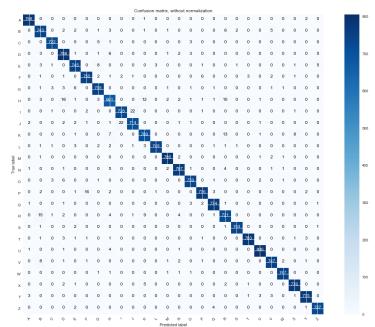
- ▶ Build a SVM using A and B data only, then A vs. C, etc
- ▶ This builds $\binom{k}{2}$ SVMs, 325 in our case.
- ▶ Tally the 325 votes, majority becomes final prediction

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	Α	0.9801	0
0.9947	S	0.9776	D
0.9926	U	0.9766	Ε
0.9920	W	0.9764	V
0.9911	Υ	0.9745	Ν
0.9905	Z	0.9726	В
0.9886	Χ	0.9652	F
0.9885	Q	0.9639	J
0.9874	М	0.9621	K
0.9862	T	0.9617	R
0.9842	L	0.9616	1
0.9823	C	0.9614	Р
0.9819	G	0.9319	Н

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