Project on Automatic Learning (Phase 3)

Zhuo Liu (0983311)

March 18, 2014

1 Introduction

In this report, we will apply supported vector machine (SVM) to build classifiers on the large and small data sets, and test their performances. Since both data sets are of multiple classes, we will use "1-vs-1" approach to build the classifiers, i.e., suppose there are k classes, we need to use SVM to build $\frac{k(k-1)}{2}$ binary separators, and using vote strategy to classify a new observation. The performance of a classifier is then evaluated by number of supported vectors that appear in any $\frac{k(k-1)}{2}$ separators, mean and standard deviation of 10 accuracies by using 10-fold cross validation. Here, accuracies are over all the classes, not only two classes. In this report, we will try three different kernels for SVM: linear, polynomial and Gaussian, i.e.

$$Linear\ kernel:\ k(x,y) =< x, y >_{\mathbb{R}^p} \tag{1}$$

$$Polynomial\ kernel:\ k(x,y) = (1+ < x, y >_{\mathbb{R}^p})^r \tag{2}$$

$$Gaussian\ kernel:\ k(x,y) = exp(-\frac{||x-y||^2}{2\sigma^2}) \tag{3}$$

In short, SVM is to find the separator with the form:

$$g(x) =< u, x > +b \tag{4}$$

where $<,>$ can be the regular inner product in $\mathbb{R}^p$ or the inner product in a Hilbert space defined by kernel function. u, b satisfy the minimization problem:

$$\min\{\frac{1}{2}||u||^2 + c\sum \xi_i\} \tag{5}$$

with constraints

$$\begin{cases} \xi_i \geq 0 \\ \xi_i - (1 - y_i g(x_i)) \geq 0 \end{cases} \tag{6}$$

Therefore, besides choosing proper parameters in kernel function, we need also to choose optimal parameters c in (5). After determining the optimal parameters for each classifiers, we will draw the graph of observations where all the supported vectors are marked, and show the numbers of supported vectors, means and standard deviations of all $\frac{k(k-1)}{2}$ separators, and give the histograms of non-zero α_i in the best and worst separators among those $\frac{k(k-1)}{2}$ separators.

In phase 2, we have already build linear classifier and non-linear classifier on principle components space and kernel principle components space (for both polynomial kernel and Gaussian kernel), so we can also compare those results to the result obtained by SVM.

2 Database: the Large One

We test each classifiers by 10-fold cross validation, i.e., randomly and equally sepearate data set into 10 subsets, choose 1 as test set, and union of the rest 9 to be training set, then build classifier based on training set and get the accuracy on the test set. We can do this 10 times, and the mean of 10 accuracies will be used to determining the optimal parameters.

2.1 Linear Kernel

The total number of observations in the data set is 1593. The following table shows the accuracies and the number of supported vectors for different choices of c in (5):

c	0.001	0.003	0.005	0.008	0.01	0.1	1
Accuracy	92.40427%	93.65976%	94.09918%	94.09918%	94.03641%	93.59699%	93.59699%
Number of SVs	1218	1003	930	885	872	878	878

So $c = 0.008$ is the optimal parameter. The following two tables shows the perfomance of the classifier by choosing $c = 0.008$. (The first table contains the information about means and standard deviations of 10 accuracies by 10-fold cross validation. Since separator over class i and j is equivalent to separator over class j and i, so either means or standard deviations need only half of the 10 by 10 matrix. In order to save space, we combine these two information into one table, so the upper triangular part of the first table shows the means of accuracies, and lower triangular part shows the stand deviations. Similarly, the second table shows the number of supported vectors by using SVM over class i and class j. All the tables in the rest of the report showing these information has the same scheme.)

class	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'
'0'	NA	99.69%	99.38%	100.0%	99.38%	99.38%	99.07%	99.69%	99.37%	99.06%
'1'	0.98821	NA	97.51%	98.44%	96.28%	99.38%	100.0%	98.13%	99.37%	97.19%
'2'	1.31762	2.46503	NA	98.74%	98.13%	99.06%	99.38%	99.68%	98.41%	99.05%
'3'	0.00000	2.20246	1.64012	NA	100.0%	97.17%	100.0%	98.42%	98.73%	94.64%
'4'	1.31762	3.48381	2.18502	0.00000	NA	99.69%	98.45%	97.18%	99.68%	98.74%
'5'	1.31762	1.31767	1.50952	1.77615	0.98821	NA	98.44%	99.37%	99.68%	99.37%
'6'	1.50952	0.00000	1.31762	0.00000	1.62742	2.20971	NA	99.37%	99.37%	100.0%
'7'	0.98821	2.18502	0.98821	2.21784	2.30757	1.33908	1.33908	NA	98.40%	97.78%
'8'	1.31762	1.33908	2.27328	1.65304	0.98821	1.02009	1.36012	1.67930	NA	96.81%
'9'	1.50952	3.10759	1.52600	4.46340	2.19484	1.31762	0.00000	2.61277	3.01814	NA

class	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'
'0'	NA	47	41	52	71	68	81	47	63	48
'1'	NA	NA	75	66	111	50	63	88	98	68
'2'	NA	NA	NA	65	55	64	59	59	89	71
'3'	NA	NA	NA	NA	53	119	61	83	105	89
'4'	NA	NA	NA	NA	NA	58	90	69	49	57
'5'	NA	NA	NA	NA	NA	NA	93	87	90	81
'6'	NA	NA	NA	NA	NA	NA	NA	70	98	47
'7'	NA	NA	NA	NA	NA	NA	NA	NA	91	67
'8'	NA	NA	NA	NA	NA	NA	NA	NA	NA	82
'9'	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Figure 1 shows the observations in PC space when choosing optimal parameter $c = 0.008$ (all the supported vectors are marked by "cross" sign).

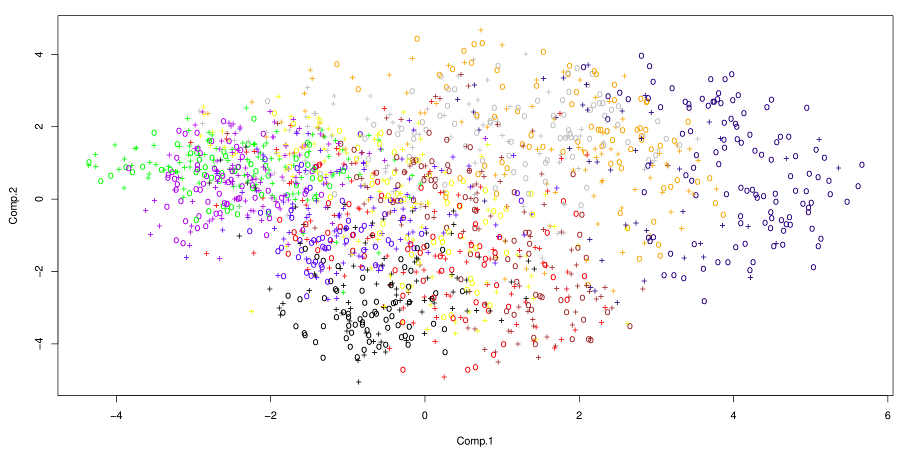


Figure 1: *Projection of data into first two PC space. Here, "cross" represent the supported vector by using linear kernel, and the colors represent different classes: navy–0, green–1, blue–2, black–3, grey–4, brown–5, orange–6, purple–7, yellow–8, red–9.*

Among the 45 SVM separators, separator over class '6' and class '9' has best perfomance, and separator over class '3' and class '9' has worst performance.

Figure 2 show the histograms of non-zero α_i of these two separators.

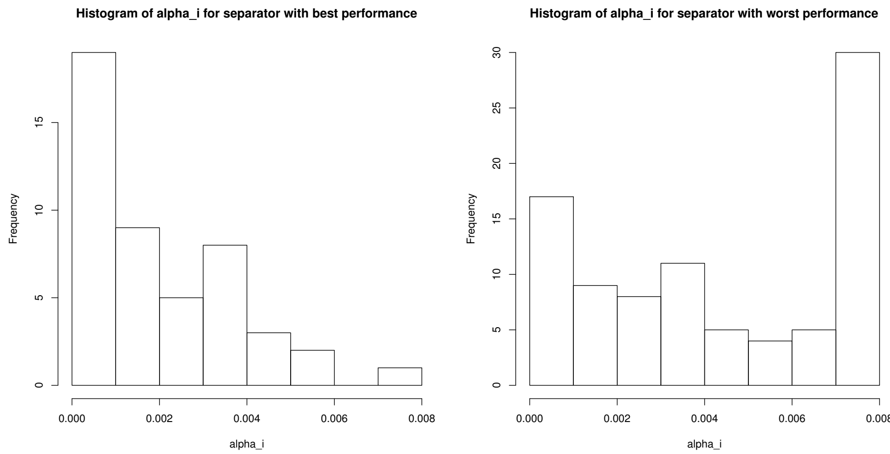


Figure 2: *Histograms of α_i for SVM separators with best and worst performances by using linear kernel on large data set with optimal parameter. Left: best, separator over class '6' and class '9', # of SVs is 47. Right: worst, separator over class '3' and '9', # of SVs is 89.*

2.2 Polynomial Kernel

The following table shows the accuracies by choosing different c in (5) and r in (2):

r \ c	10^{-5}	2×10^{-5}	3×10^{-5}	4×10^{-5}	5×10^{-5}	10^{-4}
2	95.16635%	95.79410%	95.79410%	95.73123%	95.73123%	95.73123%
3	95.35468%	95.35468%	95.35468%	95.35468%	95.35468%	95.35468%
4	88.57502%	88.57502%	88.57502%	88.57502%	88.57502%	88.57502%
5	70.80979%	70.80979%	70.80979%	70.80979%	70.80979%	70.80979%

Next table shows the number of supported vectors for different c and r:

r \ c	10^{-5}	2×10^{-5}	3×10^{-5}	4×10^{-5}	5×10^{-5}	10^{-4}
2	1304	1303	1306	1306	1306	1306
3	1382	1382	1382	1382	1382	1382
4	1477	1477	1477	1477	1477	1477
5	1526	1526	1526	1526	1526	1526

Therefore, we choose $c = 2\times10^{-5}$ and $r = 2$ as our optimal parameters. The following two tables shows the perfomance of the classifier by choosing optimal parameters. Figure 3 shows the supported vectors among all vectors. We can observe from these two tables that among the 45 SVM separators, separator over class '3' and class '6' has best perfomance, and separator over class '5' and class '9' has worst performance. Figure 4 shows the histogram of non-zero α_i for these two separators.

class	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'
'0'	NA	99.38%	99.75%	99.69%	98.14%	98.13%	98.76%	99.37%	99.37%	97.81%
'1'	1.31762	NA	97.51%	99.38%	94.12%	99.69%	99.38%	97.81%	97.48%	97.81%
'2'	2.18502	3.84148	NA	98.11%	98.13%	99.69%	100.0%	97.48%	95.86%	98.42%
'3'	0.98821	1.31762	2.67826	NA	99.38%	97.48%	100.0%	96.85%	97.13%	93.06%
'4'	3.01904	4.48198	2.63523	1.31762	NA	99.69%	97.20%	98.12%	99.05%	97.81%
'5'	2.18502	0.98821	0.98821	2.50261	0.98821	NA	99.06%	99.05%	98.09%	90.22%
'6'	2.18502	1.29784	0.00000	0.00000	2.73520	1.50952	NA	99.37%	99.68%	98.43%
'7'	1.97642	4.17967	3.22981	2.55191	1.62271	1.52600	1.31762	NA	97.76%	98.10%
'8'	1.31762	4.11608	3.00827	2.37615	1.52600	2.24305	1.02009	2.62088	NA	91.37%
'9'	2.10921	3.31047	1.66868	4.12560	2.57273	6.34766	1.64702	2.19821	5.07747	NA

class	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'
'0'	NA	118	241	119	149	151	151	129	133	255
'1'	NA	NA	249	256	231	243	150	258	177	246
'2'	NA	NA	NA	288	257	254	236	284	269	273
'3'	NA	NA	NA	NA	265	229	146	186	189	290
'4'	NA	NA	NA	NA	NA	261	192	269	262	274
'5'	NA	NA	NA	NA	NA	NA	193	204	192	294
'6'	NA	NA	NA	NA	NA	NA	NA	158	179	269
'7'	NA	NA	NA	NA	NA	NA	NA	NA	179	271
'8'	NA	NA	NA	NA	NA	NA	NA	NA	NA	296
'9'	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

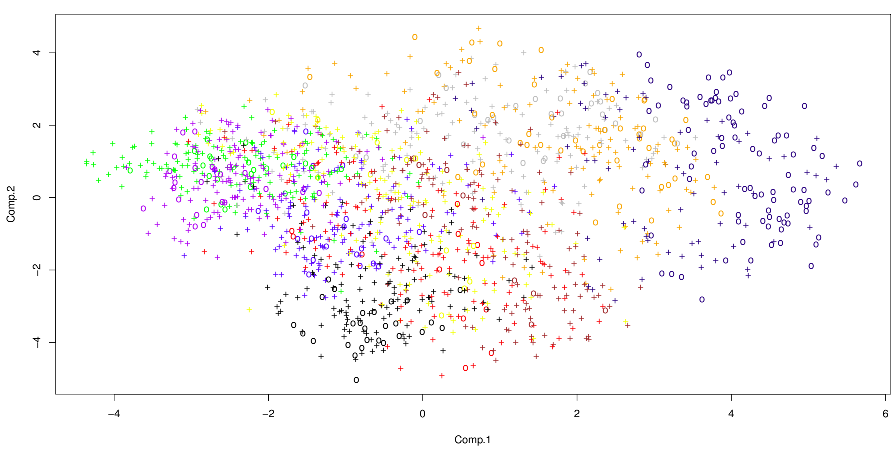


Figure 3: *Projection of data into first two PC space. Here, "cross" represent the supported vector by using polynomial kernel, and the colors represent different classes: navy–0, green–1, blue–2, black–3, grey–4, brown–5, orange–6, purple–7, yellow–8, red–9.*

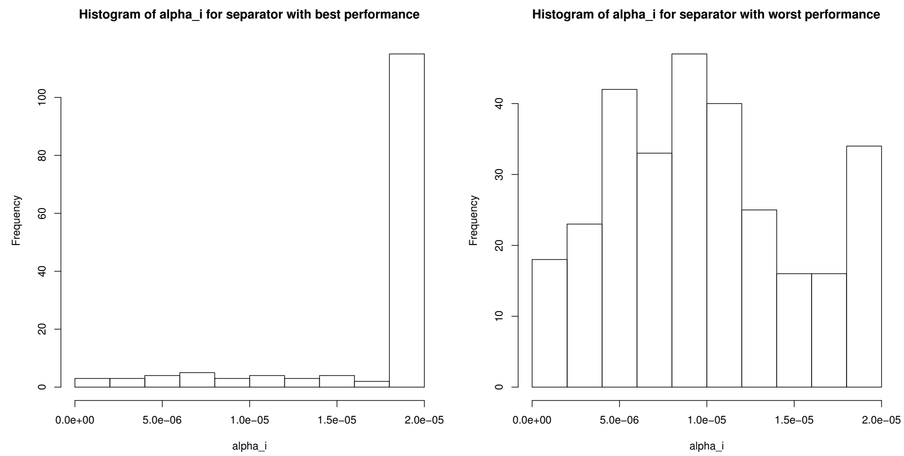


Figure 4: *Histograms of α_i for SVM separators with best and worst performances by using polynomial kernel on large data set with optimal parameter. Left: best, separator over class '3' and class '6', # of SVs is 146. Right: worst, separator over class '5' and '9', # of SVs is 294.*

2.3 Gaussian Kernel

At last, we will see the SVM by using Gaussian kernel. The parameters are c and σ, but instead of σ, we will use $\sigma^* = \frac{1}{2\sigma^2}$ in the following accuracy table:

σ^* \ c	0.2	0.5	1	10	20	50	100
0.0001	12.24105%	71.12367%	85.05964%	93.40866%	93.97363%	94.28751%	93.72254%
0.001	88.44947%	92.71814%	93.78531%	95.22913%	95.22913%	95.22913%	95.22913%
0.003	91.21155%	94.28751%	95.16635%	95.79410%	95.79410%	95.79410%	95.79410%
0.005	90.83490%	94.22473%	95.54300%	95.66855%	95.66855%	95.66855%	95.66855%
0.007	88.32392%	93.78531%	95.41745%	95.60578%	95.60578%	95.60578%	95.60578%
0.01	64.53233%	91.71375%	94.03641%	94.03641%	94.03641%	94.03641%	94.03641%
0.1	6.465788%	6.465788%	6.654112%	6.779661%	6.779661%	6.779661%	6.779661%

The following table shows the number of SVs by different σ^* and c:

σ^* \ c	0.2	0.5	1	10	20	50	100
0.0001	1593	1586	1537	1089	976	883	886
0.001	1486	1307	1173	1030	1031	1031	1031
0.003	1420	1281	1225	1224	1224	1224	1224
0.005	1454	1361	1347	1355	1355	1355	1355
0.007	1516	1443	1444	1446	1446	1446	1446
0.01	1572	1523	1521	1520	1520	1520	1520
0.1	1593	1593	1593	1593	1593	1593	1593

Therefore, we choose $\sigma^* = 0.003$ and $c = 10$ as our optimal parameters. The following two tables shows the perfomance of the classifier by choosing optimal parameters. Figure 5 shows the supported vectors among all vectors. We can observe from these two tables that among the 45 SVM separators, separator over class '0' and class '3' has best perfomance, and separator over class '3' and class '9' has worst performance. Figure 6 shows the histograms of non-zero α_i of these two separators.

class	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'
'0'	NA	99.69%	99.28%	100.0%	99.38%	99.38%	99.07%	99.69%	99.37%	98.75%
'1'	0.98821	NA	99.07%	99.07%	97.52%	99.69%	100.0%	97.50%	99.37%	98.13%
'2'	1.31762	1.50952	NA	99.06%	99.38%	99.69%	99.69%	100.0%	98.41%	99.68%
'3'	0.00000	1.49449	1.54223	NA	100.0%	96.86%	100.0%	98.74%	99.36%	96.53%
'4'	1.31762	3.20350	1.97642	0.00000	NA	100.0%	99.38%	98.75%	100.0%	99.06%
'5'	1.31762	0.98821	0.98821	3.31584	0.00000	NA	98.44%	99.68%	99.68%	98.74%
'6'	1.50952	0.00000	0.98821	0.00000	1.31762	2.20971	NA	99.37%	99.37%	100.0%
'7'	0.98821	1.97642	0.00000	2.19484	1.62702	1.02009	1.33908	NA	98.72%	98.42%
'8'	1.31762	1.33908	2.27328	1.31762	0.00000	1.02009	1.36012	1.64012	NA	98.72%
'9'	1.62702	2.18502	0.98821	3.10663	1.52600	1.62702	0.00000	2.26549	1.65304	NA

class	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'
'0'	NA	41	99	47	55	47	72	47	57	111
'1'	NA	NA	153	149	89	129	60	176	73	135
'2'	NA	NA	NA	157	125	142	139	161	173	147
'3'	NA	NA	NA	NA	118	85	50	68	82	170
'4'	NA	NA	NA	NA	NA	125	69	144	120	121
'5'	NA	NA	NA	NA	NA	NA	70	72	73	173
'6'	NA	NA	NA	NA	NA	NA	NA	63	76	117
'7'	NA	NA	NA	NA	NA	NA	NA	NA	74	129
'8'	NA	NA	NA	NA	NA	NA	NA	NA	NA	185
'9'	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

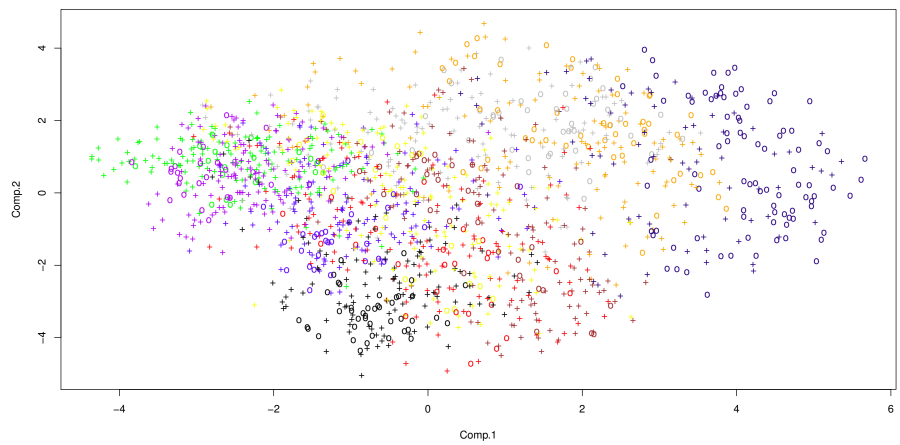
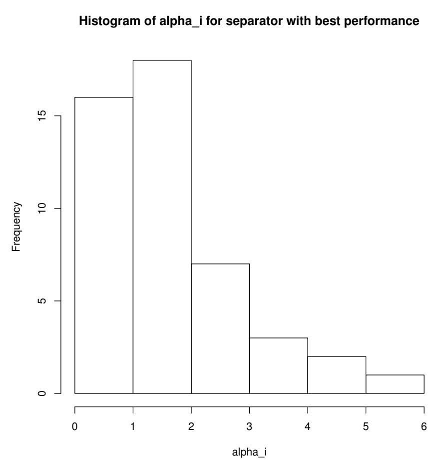


Figure 5: *Projection of data into first two PC space. Here, "cross" represent the supported vector by using Gaussian kernel, and the colors represent different classes: navy–0, green–1, blue–2, black–3, grey–4, brown–5, orange–6, purple–7, yellow–8, red–9.*



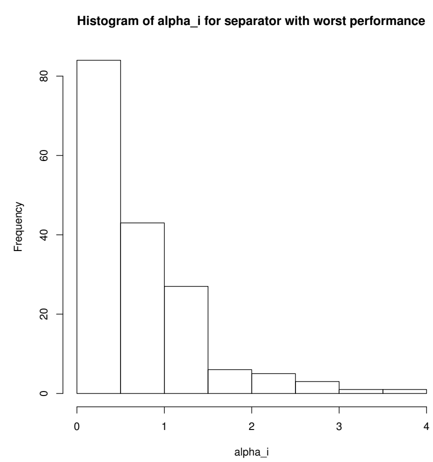


Figure 6: *Histograms of α_i for SVM separators with best and worst performances by using Gaussian kernel on large data set with optimal parameter. Left: best, separator over class '0' and class '3', # of SVs is 47. Right: worst, separator over class '3' and '9', # of SVs is 170.*

3 Database: the Small One

3.1 Linear Kernel

By different experiments, we choose $c = 10$ as our optimal parameter, the following table shows the accuracies for different choices of c:

c	0.5	1	5	10	20	50
Accuracy	94.67532%	95.28139%	95.84416%	96.01732%	95.88745%	95.84416%
Number of SVs	509	457	386	361	344	332

We choose $c = 10$ as optimal parameter. The following two tables shows the perfomance of the classifier by choosing optimal parameters. Figure 7 shows the supported vectors among all vectors. We can observe from these two tables that among the 21 SVM separators, separator over class 'foliage' and class 'sky' has best perfomance, and separator over class 'foliage' and class 'window' has worst performance. Figure 8 shows the histograms of non-zero α_i for these two separators.

class	brickface	cement	foliage	grass	path	sky	window
brickface	NA	99.55%	99.70%	100.0%	100.0%	100.0%	99.09%
cement	1.43740	NA	98.18%	99.84%	100.0%	100.0%	94.55%
foliage	0.63884	1.56484	NA	99.70%	100.0%	100.0%	91.21%
grass	0.00000	0.47913	0.63884	NA	99.70%	100.0%	99.85%
path	0.00000	0.00000	0.00000	0.63884	NA	100.0%	100.0%
sky	0.00000	0.00000	0.00000	0.00000	0.00000	NA	100.0%
window	1.05940	2.69150	4.21347	0.47913	0.00000	0.00000	NA

class	brickface	cement	foliage	grass	path	sky	window
brickface	NA	14	14	10	8	9	17
cement	NA	NA	29	11	14	9	78
foliage	NA	NA	NA	10	10	6	163
grass	NA	NA	NA	NA	10	11	8
path	NA	NA	NA	NA	NA	7	11
sky	NA	NA	NA	NA	NA	NA	8
window	NA	NA	NA	NA	NA	NA	NA

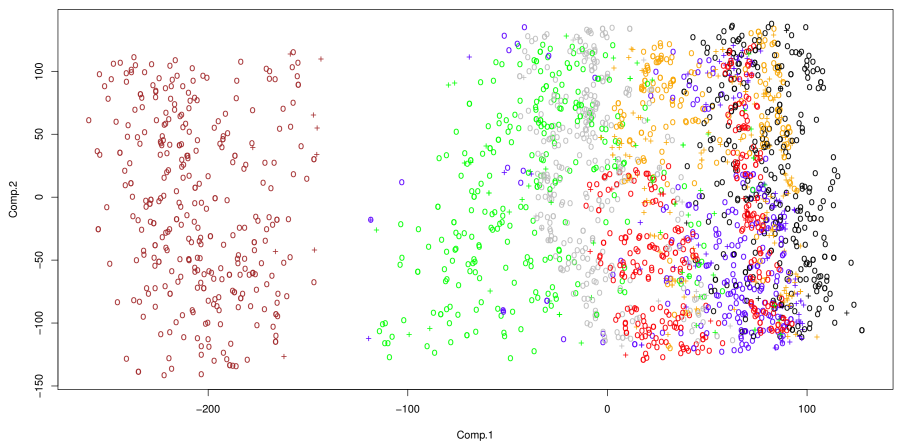


Figure 7: *Projection of data into first two PC space. Here, "cross" represent the supported vector by using linear kernel, and the colors represent different classes: red–brickface, brown–sky, blue–foliage, green–cement, orange–window, grey–path, black–grass.*

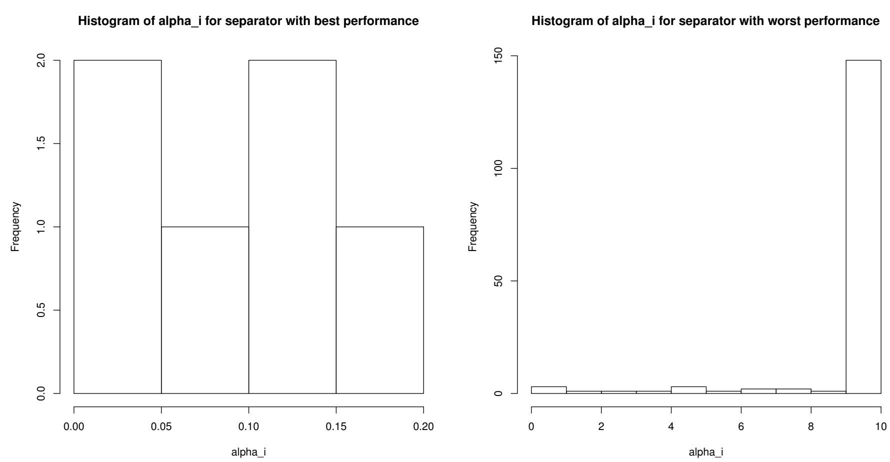


Figure 8: *Histograms of* α_i *for SVM separators with best and worst performances by using linear kernel on small data set with optimal parameter. Left: best, separator over class 'foliage' and class 'sky', # of SVs is 6. Right: worst, separator over class 'foliage' and 'window', # of SVs is 163.*

3.2 Polynomial Kernel

The following table shows the accuracies by choosing different c and r:

r \ c	0.5	0.9	1	1.2	5	10
2	96.79654%	97.22944%	97.18615%	97.31602%	96.83983%	96.79654%
3	96.58009%	96.58009%	96.62338%	96.58009%	96.45022%	96.36364%
4	96.58009%	96.58009%	96.58009%	96.53680%	96.36364%	96.36364%
5	96.40693%	96.40693%	96.40693%	96.40693%	96.40693%	96.40693%

Next table shows the total number of supported vectors for different c and r:

r \ c	0.5	0.9	1	1.2	5	10
2	356	339	342	338	324	330
3	321	321	318	320	314	313
4	351	352	349	354	351	351
5	328	328	328	328	328	328

Therefore, we choose $c = 1.2$ and $r = 2$ as our optimal parameters. The following two tables shows the perfomance of the classifier by choosing optimal parameters. Figure 9 shows the supported vectors among all vectors. We can observe from these two tables that among the 21 SVM separators, separator over class 'brickface' and class 'path' has best perfomance, and separator over class 'foliage' and class 'window' has worst performance. Figure 10 shows the histogram of non-zero α_i for these two separators.

class	brickface	cement	foliage	grass	path	sky	window
brickface	NA	99.09%	100.0%	100.0%	100.0%	99.85%	99.09%
cement	1.46378	NA	97.88%	99.70%	100.0%	99.55%	95.91%
foliage	0.00000	1.77847	NA	99.85%	100.0%	100.0%	93.94%
grass	0.00000	0.63884	0.47913	NA	99.85%	99.70%	99.70%
path	0.00000	0.00000	0.00000	0.47913	NA	99.85%	100.0%
sky	0.47913	1.02265	0.00000	0.63884	0.47913	NA	99.85%
window	1.05940	2.02651	1.74955	0.63884	0.00000	0.47913	NA

class	brickface	cement	foliage	grass	path	sky	window
brickface	NA	33	34	50	23	42	35
cement	NA	NA	39	39	50	30	69
foliage	NA	NA	NA	31	30	28	93
grass	NA	NA	NA	NA	40	41	38
path	NA	NA	NA	NA	NA	26	38
sky	NA	NA	NA	NA	NA	NA	32
window	NA	NA	NA	NA	NA	NA	NA

3.3 Gaussian Kernel

At last, we will see the SVM by using Gaussian kernel, here, the parameters are c and σ, in order to simplify the notation, we will use $\sigma^* = \frac{1}{2\sigma^2}$ in the following

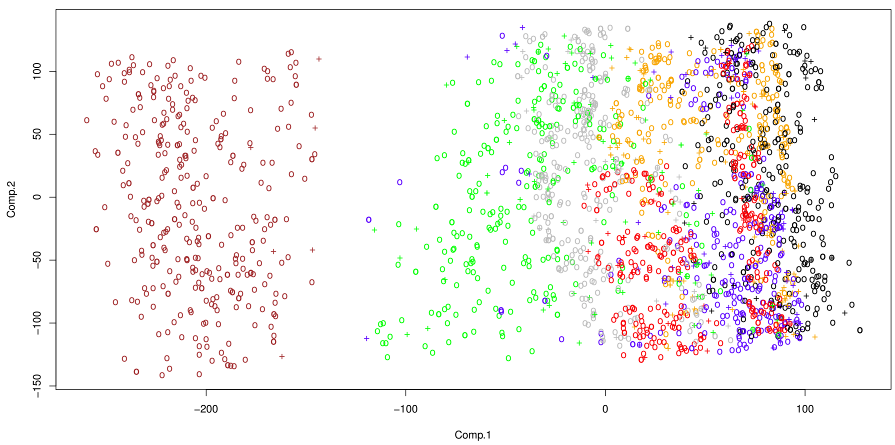


Figure 9: *Projection of data into first two PC space. Here, "cross" represent the supported vector by using polynomial kernel, and the colors represent different classes: red–brickface, brown–sky, blue–foliage, green–cement, orange–window, grey–path, black–grass.*

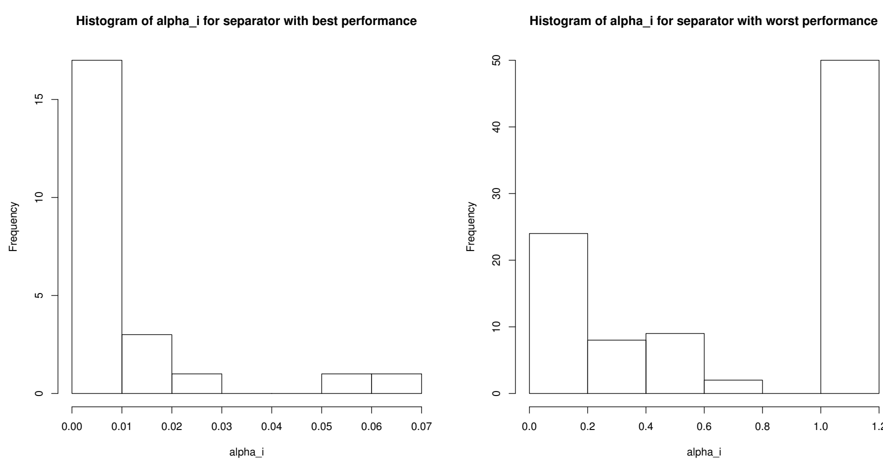


Figure 10: *Histograms of α_i for SVM separators with best and worst performances by using polynomial kernel on small data set with optimal parameter. Left: best, separator over class 'brickface' and class 'path', # of SVs is 23. Right: worst, separator over class 'foliage' and 'window', # of SVs is 93.*

accuracy table:

σ^* \ c	0.5	1	5	10	20	50
0.08	93.50649%	94.45887%	96.36364%	96.79654%	96.79654%	96.96970%
0.1	93.80952%	94.63203%	96.49351%	96.92641%	97.27273%	97.09957%
0.2	93.67965%	94.97835%	96.75325%	96.75325%	96.92641%	96.92641%
0.4	93.37662%	95.45455%	96.40693%	96.40693%	96.32035%	96.36364%
0.6	93.11688%	94.80519%	95.49784%	95.62771%	95.71429%	95.71429%
0.8	92.98701%	94.45887%	95.28139%	95.41126%	95.41126%	95.36797%
1	92.51082%	94.19913%	94.97835%	94.93506%	94.89177%	95.02165%

Therefore, we choose $\sigma^* = 0.1$ and $c = 20$ as our optimal parameters. The following two tables shows the perfomance of the classifier by choosing optimal parameters. Figure 11 shows the supported vectors among all vectors. We can observe from these two tables that among the 21 SVM separators, separator over class 'brickface' and class 'path' has best perfomance, and separator over class 'foliage' and class 'window' has worst performance. Figure 12 shows the histogram of non-zero α_i for these two separators.

class	brickface	cement	foliage	grass	path	sky	window
brickface	NA	99.40%	99.85%	99.85%	100.0%	99.39%	99.85%
cement	1.05940	NA	99.09%	99.24%	100.0%	99.85%	95.45%
foliage	0.47913	1.27769	NA	99.70%	100.0%	99.85%	95.00%
grass	0.47913	1.47246	0.63884	NA	99.70%	99.40%	99.55%
path	0.00000	0.00000	0.00000	0.63884	NA	99.85%	100.0%
sky	1.05940	0.47913	0.47913	1.05940	0.47913	NA	99.70%
window	0.47913	2.85700	2.26429	0.73189	0.00000	0.63884	NA

class	brickface	cement	foliage	grass	path	sky	window
brickface	NA	71	71	95	65	66	73
cement	NA	NA	95	87	112	83	112
foliage	NA	NA	NA	67	65	57	161
grass	NA	NA	NA	NA	85	62	83
path	NA	NA	NA	NA	NA	72	71
sky	NA	NA	NA	NA	NA	NA	58
window	NA	NA	NA	NA	NA	NA	NA

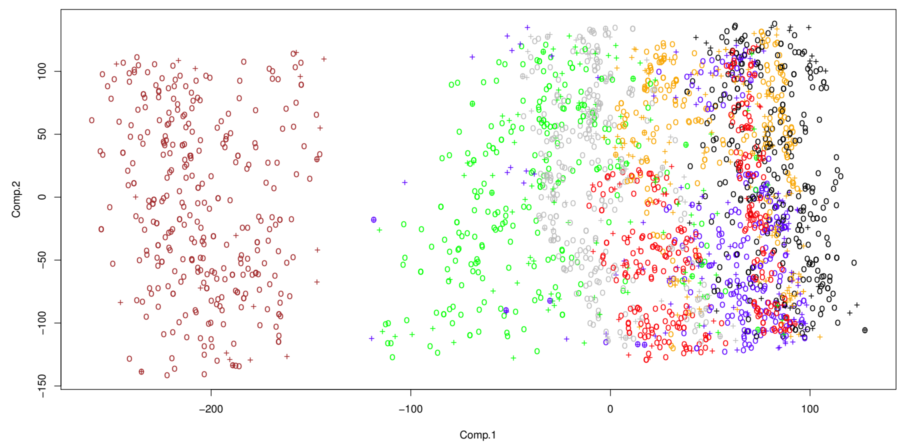


Figure 11: *Projection of data into first two PC space. Here, "cross" represent the supported vector by using Gaussian kernel, and the colors represent different classes: red–brickface, brown–sky, blue–foliage, green–cement, orange–window, grey–path, black–grass.*

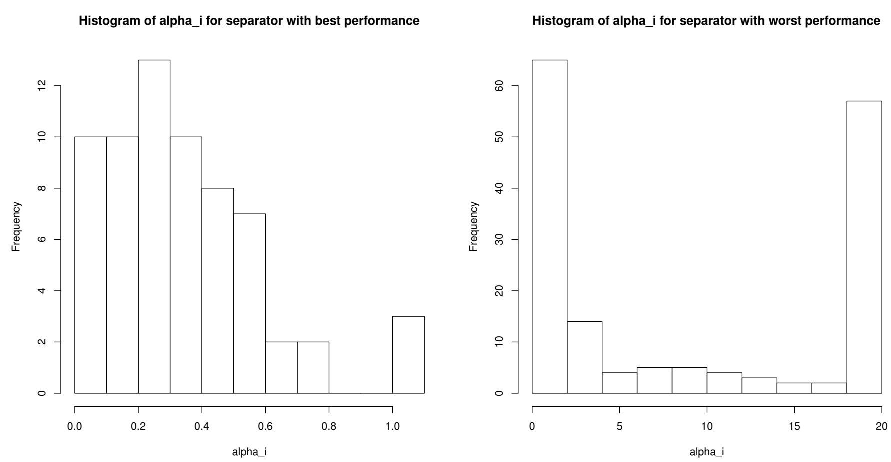


Figure 12: *Histograms of α_i for SVM separators with best and worst performances by using Gaussian kernel on small data set with optimal parameter. Left: best, separator over class 'brickface' and class 'path', # of SVs is 65. Right: worst, separator over class 'foliage' and 'window', # of SVs is 161.*

A Appendix: running times

The following table shows the running time on large data sets (1593 observations, 256 descriptors, 10 classes):

linear	polynomial	Gaussian
5.362499 secs	14.70856 secs	9.907342 secs

And the next table shows the running time on small data sets (2320 observations, 19 descriptors, 7 classes):

linear	polynomial	Gaussian
0.7461352 secs	0.685853 secs	0.8766298 secs