Machine Learning: Problem Set 4

Jingpeng Hong

March 8, 2022

Question 1

c

The 50 most frequently used words in the corpus in word cloud are shown in Figure 1

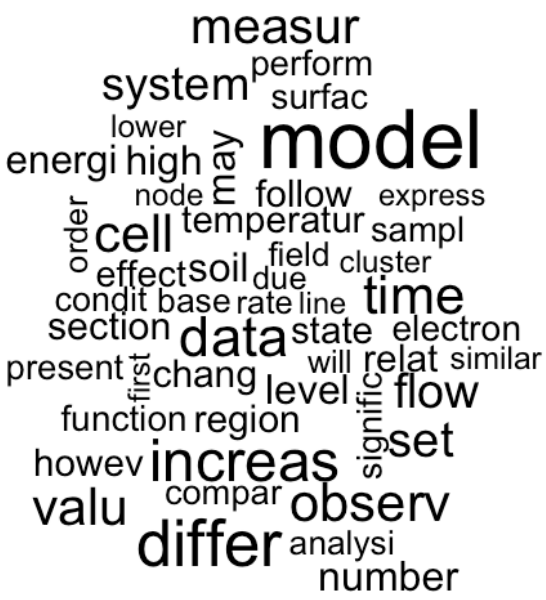

Figure 1: Word Cloud

d

When $k = 2$, results are shown in Table 1. The first topic is like Physics, with specific issues related to Particle Physics /Nuclear Physics, because we have words "electron", "energy", "flow", "temperature", etc. The second topic is like Biology/Ecology because we have words "cell", "soil", "cluster", etc.

Table 1: Topic Model $k = 2$

Topic 1		Topic 2	
term	beta	term	beta
model	0.00749	cell	0.00635
observ	0.00518	soil	0.00493
energi	0.00486	differ	0.00410
flow	0.00457	state	0.00406
measur	0.00447	data	0.00392
valu	0.00438	time	0.00380
differ	0.00432	increas	0.00371
electron	0.00431	cluster	0.00361
temperatur	0.00423	node	0.00360
increas	0.00380	system	0.00347

When $k = 3$, results are shown in Table 2. The Topic 1 is similar to the first topic in $k = 2$, potentially related to Particle Physics /Nuclear Physics; The Topic 2 is ambiguous but similar to Computer Science with words "algorithm", "system", etc; The Topic 3 is similar to the second Biology/Ecology topic in $k = 2$ with additional words "gene", "plant", etc.

Table 2: Topic Model $k = 3$

Topic 1		Topic 2		Topic 3	
term	beta	term	beta	term	beta
model	0.00829	set	0.00866	cell	0.0116
observ	0.00571	state	0.00664	soil	0.00896
energi	0.00543	node	0.00625	express	0.00519
increas	0.00519	cluster	0.00611	plant	0.00517
valu	0.00492	algorithm	0.00608	gene	0.00505
measur	0.00486	function	0.00553	figur	0.00459
electron	0.00482	time	0.00552	differ	0.00422
flow	0.00481	number	0.00514	level	0.00395
temperatur	0.00480	let	0.00505	data	0.00373
surfac	0.00440	system	0.00460	differenti	0.00350

When $k = 5$, results are shown in Table 3. The Topic 1 and Topic 2 are separated from the Topic 3 in $k = 3$, while the Topic 1 is more like Pathology/Medicine with words "patient", "disease", "neuron",

and the Topic 2 is more like Ecology/Environmental Science with words "sediment", "carbon", etc. The Topic 3 and Topic 4 are separated from the Topic 2 in $k = 3$, while the Topic 3 is more like Mathematics with words "set", "theorem", "function", and the Topic 4 is more like Engineering with words "turbine". The Topic 5 is similar to the Topic 1 in $k = 3$, indicating the Nuclear Physics/ Particle Physics.

Table 3: Topic Model $k = 5$

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
term	beta	term	beta	term	beta	term	beta	term	beta
cell	0.0195	soil	0.0123	let	0.0114	model	0.00781	model	0.00997
figur	0.00744	increas	0.00738	set	0.0105	cluster	0.00744	energi	0.00808
express	0.00728	plant	0.00697	state	0.0102	algorithm	0.00734	observ	0.00784
differenti	0.00588	site	0.00576	oper	0.00799	turbin	0.00690	electron	0.00720
cultur	0.00575	sampl	0.00509	follow	0.00771	time	0.00642	temperatur	0.00665
gene	0.00546	sediment	0.00468	order	0.00756	perform	0.00621	measur	0.00613
patient	0.00486	chang	0.00452	theorem	0.00735	differ	0.00617	flow	0.00568
diseas	0.00484	speci	0.00443	modul	0.00652	number	0.00582	particl	0.00535
human	0.00475	data	0.00422	function	0.00650	set	0.00580	surfac	0.00525
neuron	0.00438	carbon	0.00420	element	0.00633	flow	0.00536	differ	0.00469

When $k = 8$ and $k = 10$, we don't show the results here but can see them from the R code. In my opinion, $k = 5$ yields the most meaningful coherence for each topic. For $k = 2$ and $k = 3$, we can find some topics cannot be separated efficiently and mix in one topic; For $k = 8$ and $k = 10$, some topics are split into tow or three very similar topics, which are difficult to distinguish. Hence, $k = 5$ can be the most appropriate choice for k.

e

From the Table below, we can find that the best parameters are $k = 10$ and $\alpha = 0.5$.

k	2				5			
α	0.01	0.1	0.5	1	0.01	0.1	0.5	1
perplexity	5748.743	5779.455	5672.763	5696.255	4362.107	4364.209	4358.953	4408.915
k	8				10			
α	0.01	0.1	0.5	1	0.01	0.1	0.5	1
perplexity	3844.230	3872.365	3870.964	3831.448	3739.974	3701.829	3701.015	3713.257

Table 4: k, α, perplexity

The model estimated from $k = 10$ and $\alpha = 0.5$ is shown below. We can find some topics like Pathology/Medicine (Topic 3), Geology (Topic 6), Mathematics (Topic 7), Computer Science (Topic 8), Atmospheric Science (Topic 9). However, some topics are quite similar and cannot be distinguished, like Topic 5 and Topic 6, Topic 9 and Topic 10.

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
state	0.0181	cluster	0.0195	cell	0.0259	turbin	0.0143	soil	0.0111
node	0.0111	program	0.00819	figur	0.00989	flow	0.0108	particl	0.00922
time	0.0110	set	0.00774	express	0.00847	wind	0.0104	measur	0.00727
algorithm	0.0103	differ	0.00740	cultur	0.00812	speed	0.00910	sampl	0.00671
execut	0.00970	depend	0.00726	differenti	0.00730	veloc	0.00810	surfac	0.00585
modul	0.00881	data	0.00707	gene	0.00646	power	0.00725	materi	0.00568
system	0.00835	slice	0.00673	human	0.00630	model	0.00714	valu	0.00552
set	0.00709	coher	0.00659	mutat	0.00570	increas	0.00671	temperatur	0.00523
implement	0.00634	featur	0.00623	neuron	0.00557	beach	0.00611	film	0.00520
comput	0.00616	function	0.00600	patient	0.00550	plant	0.00604	polym	0.00511
Topic 6		Topic 7		Topic 8		Topic 9		Topic 10	
sediment	0.00865	let	0.0150	network	0.0150	electron	0.0175	model	0.0211
soil	0.00801	set	0.0113	model	0.0124	energi	0.0144	temperatur	0.0120
site	0.00746	theorem	0.00942	flow	0.00832	observ	0.0114	atmospher	0.0113
chang	0.00672	oper	0.00870	node	0.00823	region	0.00897	observ	0.00717
increas	0.00643	order	0.00870	trace	0.00791	field	0.00834	flux	0.00639
sampl	0.00576	lemma	0.00838	algorithm	0.00782	magnetospher	0.00728	transit	0.00615
carbon	0.00552	follow	0.00817	connect	0.00691	ion	0.00645	pressur	0.00607
organ	0.00483	element	0.00757	differ	0.00682	flow	0.00607	rate	0.00576
core	0.00461	function	0.00733	attack	0.00673	direct	0.00552	profil	0.00558
record	0.00445	proof	0.00721	number	0.00670	thermospher	0.00545	depth	0.00555

Question 2

a

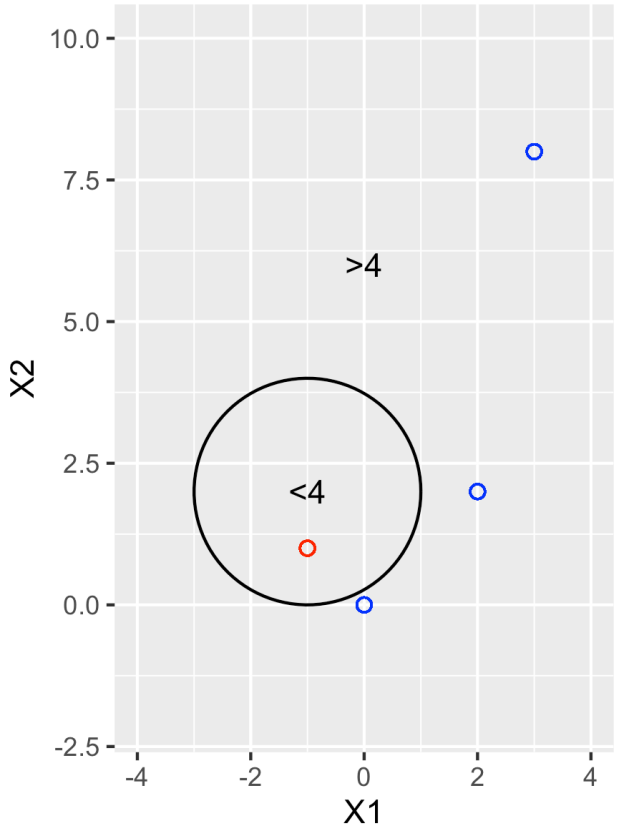

Figure 2: Non-linear decision boundary, $(1 + X_1)^2 + (2-X_2)^2 = 4$

b

As shown in Figure 2, for points inside the circle (and on the boundary) we have $(1+X_1)^2+(2-X_2)^2 \leq 4$; for points outside the circle we have $(1 + X_1)^2 + (2-X_2)^2 > 4$

c

As shown in Figure 2, Blue Class: (0, 0), (2, 2), (3,8); Red Class: (-1, 1).

d

We can rewrite the non-linear boundary as:

$$1 + 2X_1 - 4X_2 + X_1^2 + X_2^2 = 0 \tag{1}$$

This boundary is linear in terms of X_1, X_1^2, X_2, and X_2^2.

Question 3

Linear

In Figure 3 we show the linear decision boundary.

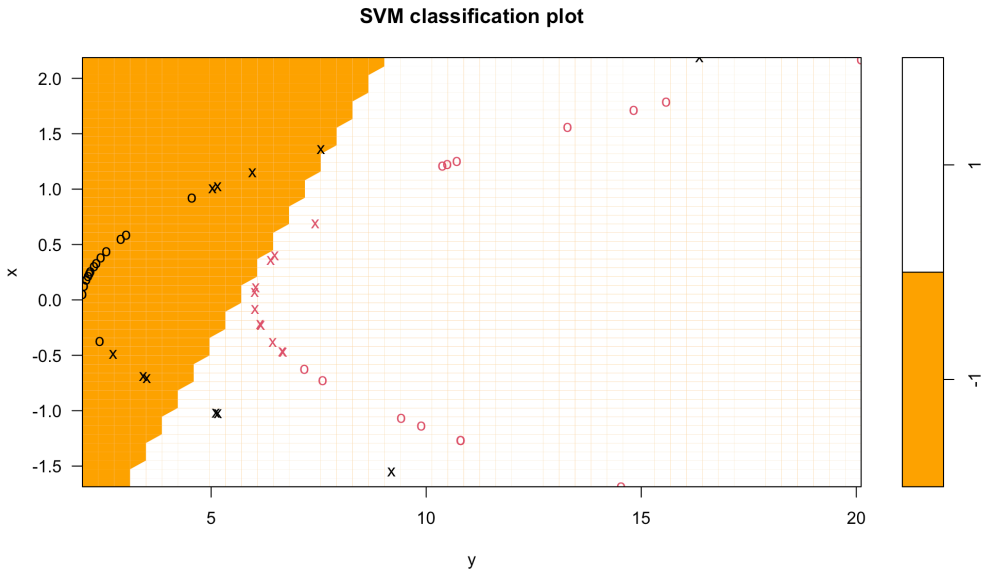

Figure 3: linear support vector classifier

Training		Truth	
		-1	1
Predict	-1	21	0
	1	4	25

Test		Truth	
		-1	1
Predict	-1	18	0
	1	7	25

Table 5: Training & Test Error Rates, linear support vector classifier

From Table 5, we can calculate

- training error rate = $4/50 = 8\%$
- test error rate = $7/50 = 14\%$

Polynomial

In Figure 4 we show the decision boundary with a polynomial kernel.

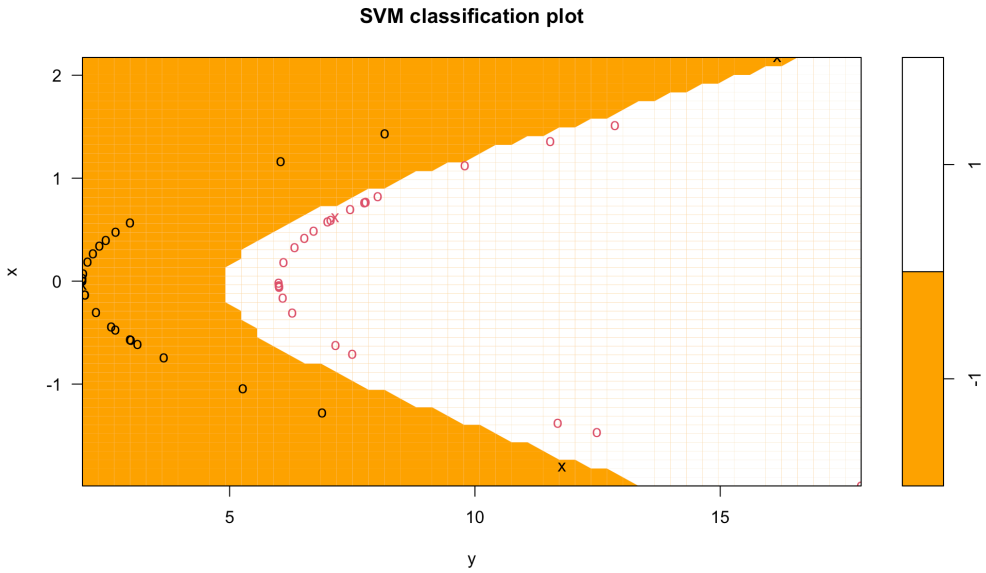

Figure 4: Polynomial Kernel

Training		Truth	
		-1	1
Predict	-1	25	0
	1	0	25

Test		Truth	
		-1	1
Predict	-1	24	0
	1	1	25

Table 6: Training & Test Error Rates, polynomial kernel

From Table 6, we can calculate

- training error rate = 0
- test error rate = $1/50 = 2\%$

Radial

In Figure 5 we show the decision boundary with a radial kernel.

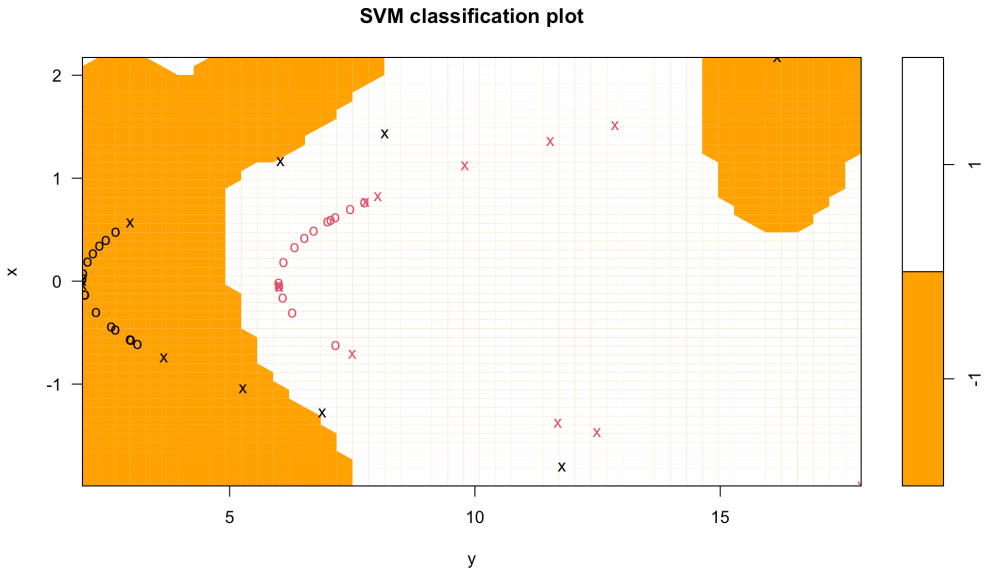

Figure 5: Radial Kernel

Training		Truth	
		-1	1
Predict	-1	23	0
	1	2	25

Test		Truth	
		-1	1
Predict	-1	19	0
	1	6	25

Table 7: Training & Test Error Rates, radial kernel

From Table 7, we can calculate

- training error rate = $2/50 = 4\%$
- test error rate = $6/50 = 12\%$

Comparison

From the training error rates we can see Linear > Radial > Polynomial, therefore, the SVM with a polynomial kernel (degree $= 3$) or a radial kernel will outperform the linear support vector classifier. We can also find the polynomial kernel performs best on the test data.

Question 4

b

From the results below we find that "cost=1" results in the lowest cross-validation error rate. We can see that when "cost" is large, the margins will be narrow. The classifier will fit the training data well, which may have low bias but high variance. On the contrary, when "cost" is small, the margins will be wide, which may have low variance but high bias. "cost=1" can control the bias-variance trade-off best among these values. Now we use the best model obtained through cross-validation to make predictions.

	cost	error	dispersion
1	1e-03	0.09692308	0.04792216
2	1e-02	0.07647436	0.03607259
3	1e-01	0.04583333	0.02886633
4	1e+00	0.01269231	0.01338091
5	5e+00	0.01782051	0.01229997
6	1e+01	0.02038462	0.01074682
7	1e+02	0.03314103	0.02709221

Table 8: Cross-validation

c

Using SVMs with radial kernels, we can find that (cost=10, gamma=0.01) will have the smallest error (Table 9).

	cost	gamma	error	dispersion
1	0.1	0.01	0.08660256	0.03616159
2	1.0	0.01	0.07115385	0.04047552
3	10.0	0.01	0.02532051	0.02649099
4	0.1	0.10	0.07628205	0.03718317
5	1.0	0.10	0.05583333	0.04059655
6	10.0	0.10	0.04070513	0.02441000
7	0.1	1.00	0.57403846	0.03308207
8	1.0	1.00	0.06608974	0.04319459
9	10.0	1.00	0.05583333	0.04235779

Table 9: Cross-validation, radial kernel

Using SVMs with polynomial kernels, we can find that (cost=10, degree=2) will have the smallest error (Table 10).

	cost	degree	error	dispersion
1	0.1	2	0.5791026	0.03412280
2	1.0	2	0.5791026	0.03412280
3	10.0	2	0.5586538	0.04837302
4	0.1	3	0.5791026	0.03412280
5	1.0	3	0.5791026	0.03412280
6	10.0	3	0.5791026	0.03412280
7	0.1	4	0.5791026	0.03412280
8	1.0	4	0.5791026	0.03412280
9	10.0	4	0.5791026	0.03412280

Table 10: Cross-validation, polynomial kernel

d

Create plots displaying pairs of variables at a time.

Linear

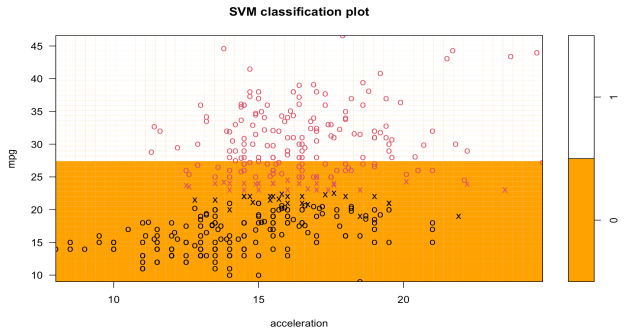

Figure 6: gas mileage ~ acceleration

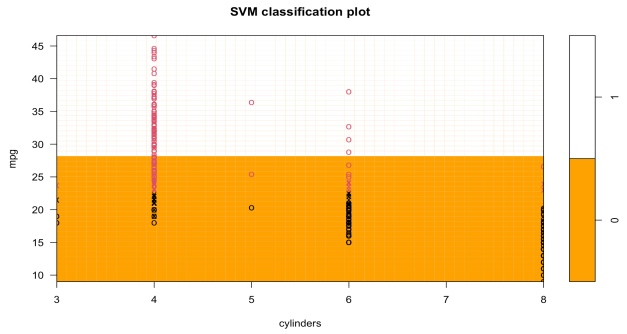

Figure 7: gas mileage ~ cylinders

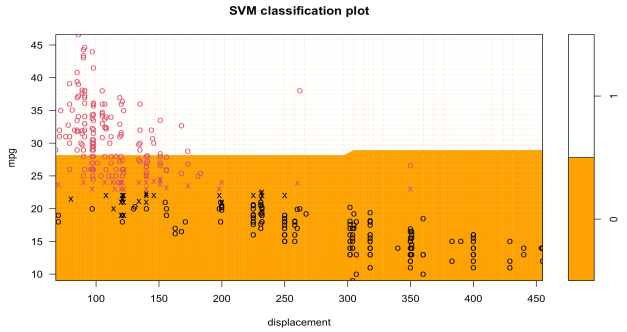

Figure 8: gas mileage ~ displacement

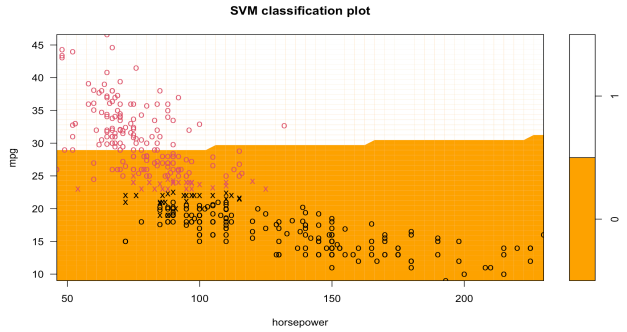

Figure 9: gas mileage ~ horsepower

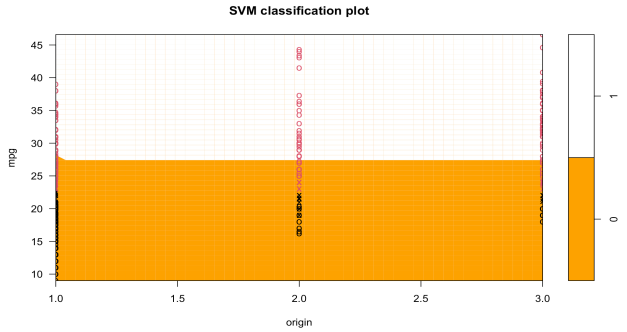

Figure 10: gas mileage ~ origin

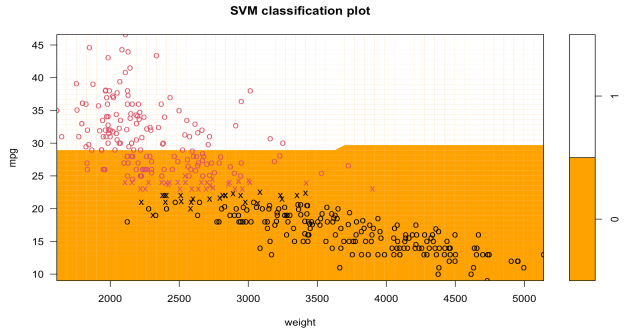

Figure 11: gas mileage ~ weight

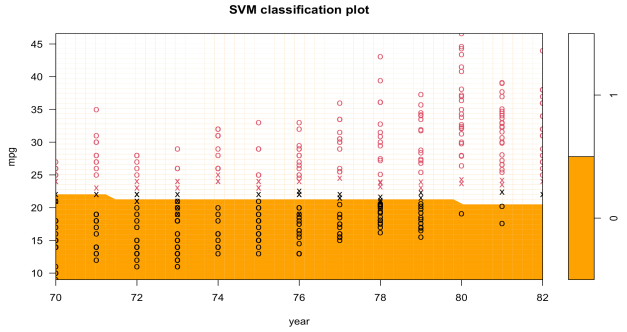

Figure 12: gas mileage ~ year

Radial

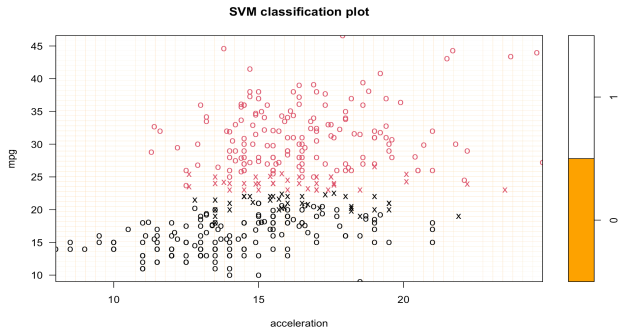

Figure 13: gas mileage ~ acceleration

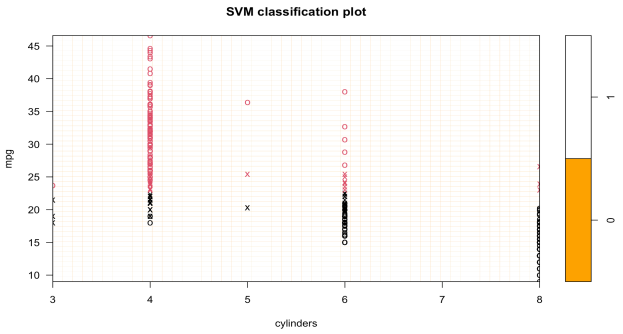

Figure 14: gas mileage ~ cylinders

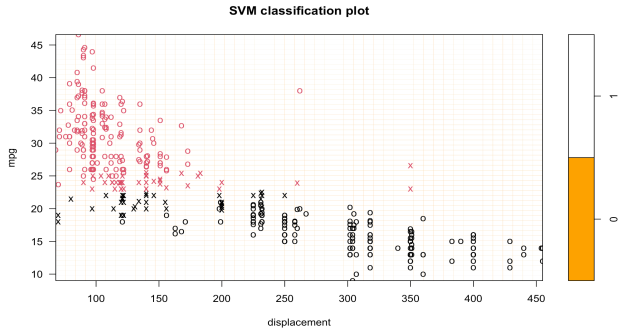

Figure 15: gas mileage ~ displacement

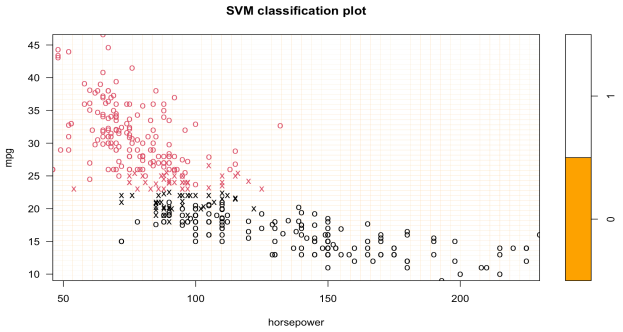

Figure 16: gas mileage ~ horsepower

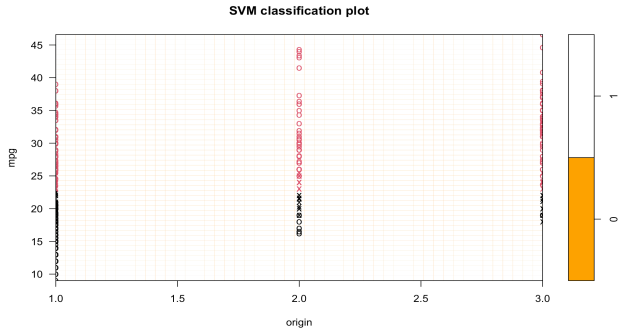

Figure 17: gas mileage ~ origin

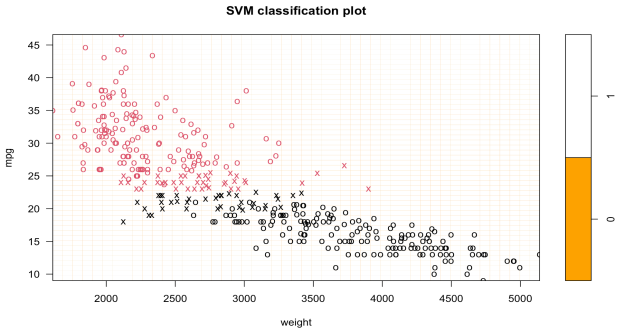

Figure 18: gas mileage ~ weight

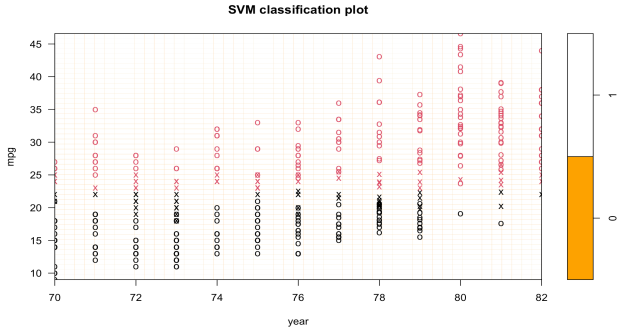

Figure 19: gas mileage ~ year

Polynomial

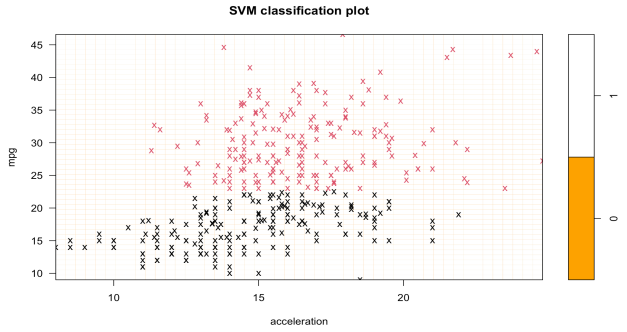

Figure 20: gas mileage ∼ acceleration

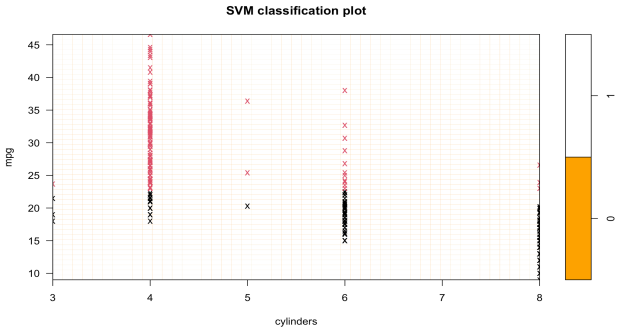

Figure 21: gas mileage ∼ cylinders

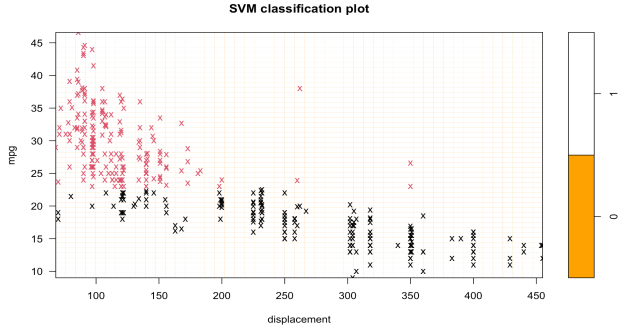

Figure 22: gas mileage ∼ displacement

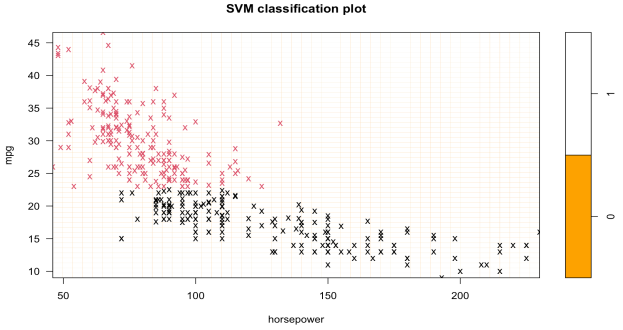

Figure 23: gas mileage ∼ horsepower

Figure 24: gas mileage ∼ origin

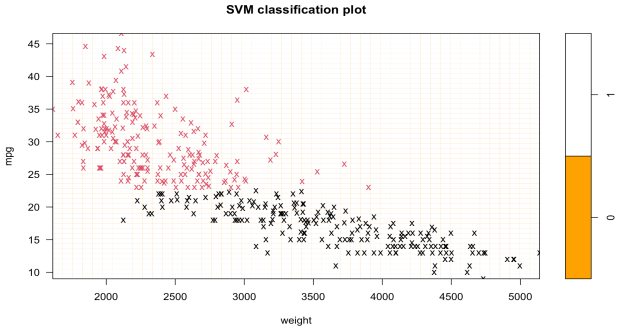

Figure 25: gas mileage ∼ weight

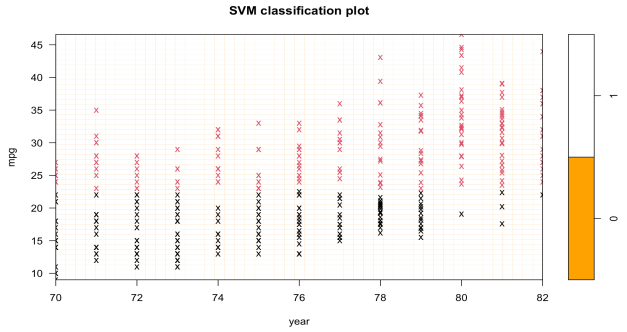

Figure 26: gas mileage ∼ year