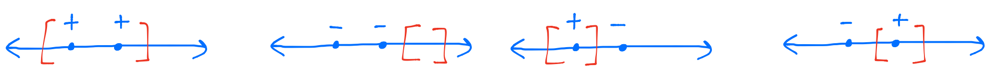
ML HW # 4

Joshua Shapiro
11/20/16

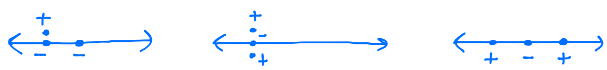
1) 1) Let the input space be the real line, and let H be the hypothesis class of intervals. That is, each hypothesis h is associated with a closed interval [a,b], for some constants a ≤ b, and h(x) is +1 if and only if x lies within this interval. What is the VC dimension of H? Prove that your answer is correct.

For 1 point: (proving H can shatter)

For 2 points: (proving H can shatter)



For 3 points: (proving H cannot shatter)



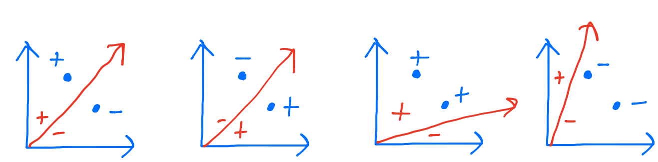
For any sets of points there exists some labeling such that for all h in H, h cannot achieve the labeling. Therefore, VC dimension is 2.

1) 2) Let the input space be the real line, and let H be the hypothesis class of unions of k intervals. That is, each hypothesis h is associated with k closed intervals and h(x) is +1 if and only if x lies in the union of these intervals. What is the VC dimension of H? Justify.

If k is allowed to equal infinity, the VC dimension of H will be infinity. This is because for any n points, there exists a point set S of size n that for all possible labelings of S there exists some h in H which achieves that labeling. It would be impossible to draw each example for every value of n, so my reasoning is as follows: For n points, choose a point set S where no points are overlapping (for example, the 3rd example of 3 points above (not the 1st or second example of 3 points)). Then, for all labels, simply draw bounds around the + points making a new interval and leave the - points outside of all intervals. K will need to be infinite, but it will always be possible to find enough intervals to label the points properly.

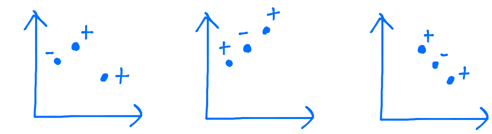
1) 3) Let the input space be R^2, and let H consist of all homogeneous linear separators (i.e., linear separators which pass through the origin). Show that H has a VC dimension of 2.

First prove H can shatter 2 points:



There exists a point set S of size 2 that for all possible labelings of S there exists some h in H which achieves that labeling.

Then prove H cannot shatter 3 points:



For any set of 3 points there exists a labeling such that no h in H can achieve that labeling.

Therefore, VC dimension = 2

3) A) Randomly Simulated Data

1) LASSO

a) Values of λ vs non-zero coefficients of w:

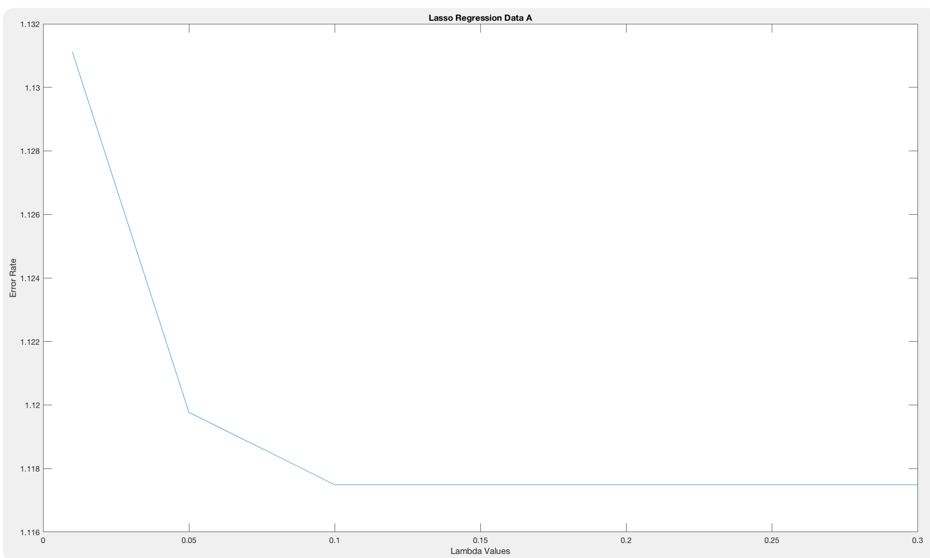
λ:	0.01	0.05	0.1	0.2	0.3
# non-zero:	13	2	0	0	0

b) Value of λ with minimum number of 0s:

0.1, 0.2, 0.3

c) Test error with respect to λ

λ:	0.01	0.05	0.1	0.2	0.3
test error:	1.1311	1.1198	1.1175	1.1175	1.1175



d) Value of λ that yields lowest test error:

0.1, 0.2, 0.3 All λ have 0 non-zero coefficients for their w.

e) Can you get all entries to be 0 for dataset A?

Yes, when λ is 0.1or higher.

2) RIDGE

a) Values of λ vs non-zero coefficients of w:

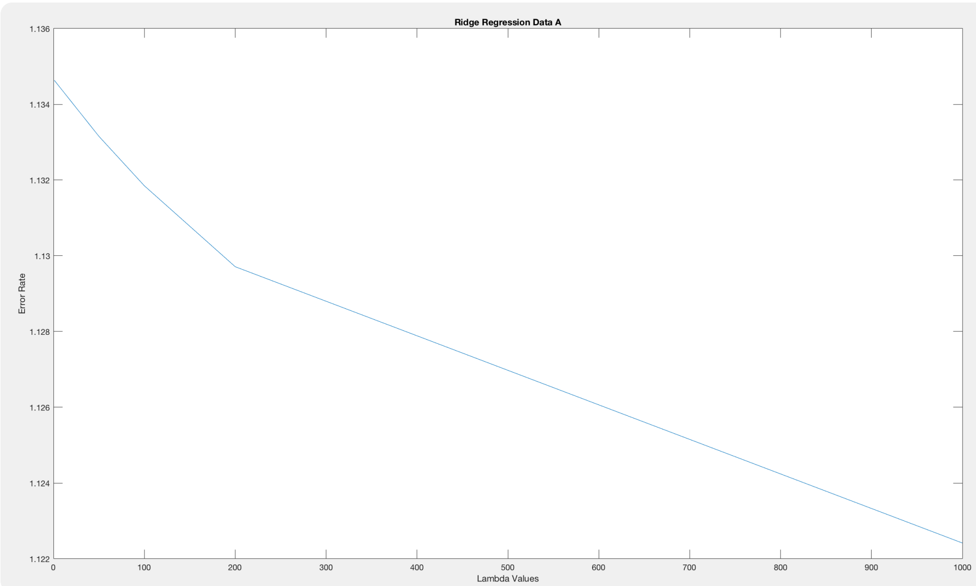
λ:	1	50	100	200	1000
# non-zero:	20	20	20	20	20

b) Value of λ with minimum number of 0s:

1, 50, 100, 200, 1000

c) Test error with respect to λ

λ:	1	50	100	200	1000
test error:	1.1346	1.1332	1.1318	1.1297	1.1224



d) Value of λ that yields lowest test error:

1000. This λ has the same number of non-zero coefficients as other λs.

e) Can you get all entries to be 0 for dataset A?

No.

3) B) Cloud Data

1) LASSO

a) Values of λ vs non-zero coefficients of w:

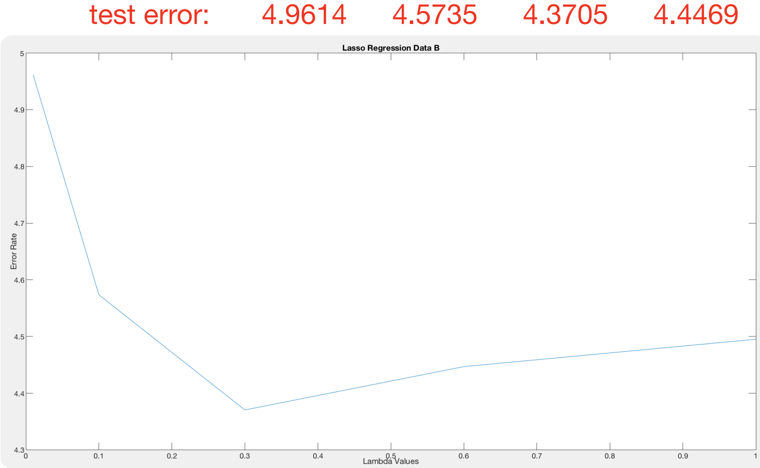
λ:	0.01	0.1	0.3	0.6	1
# non-zero:	4	3	2	2	0

b) Value of λ with minimum number of 0s:

1

c) Test error with respect to λ

λ:	0.01	0.1	0.3	0.6	1
test error:	4.9614	4.5735	4.3705	4.4469	4.4952



d) Value of λ that yields lowest test error:

0.3 , not minimum number of non-zero coefficients

e) Can you get all entries to be 0?

Yes, when λ is 1

2) RIDGE

a) Values of λ vs non-zero coefficients of w:

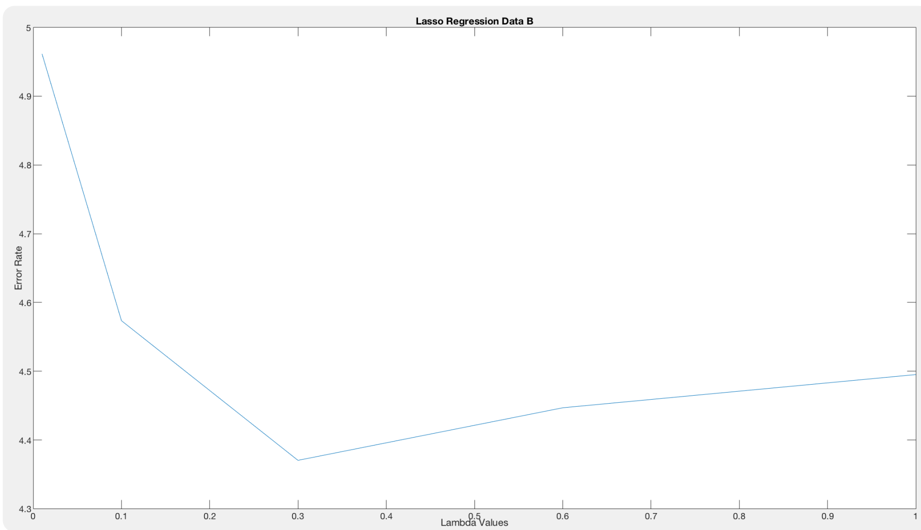
λ:	1	20	40	60	80
# non-zero:	9	9	9	9	9

b) Value of λ with minimum number of 0s:

1, 20, 40, 60, 80

c) Test error with respect to λ

λ:	1	20	40	60	80
test error:	91.1534	29.4198	47.6048	63.7569	77.1634



d) Value of λ that yields lowest test error:

20. This λ has the same number of non-zero coefficients as other λs.

e) Can you get all entries to be 0 for dataset A?

No.

3) Overall results & wd/wb table:

	LASSO				RIDGE			
	wb#	wbTestErr	wd#	wdTestErr	wb#	wbTestErr	wd#	wdTestErr
DATA A	0	1.1175	0	1.1175	20	N/A	20	1.1224
DATA B	0	4.4952	2	4.3705	9	N/A	9	29.4198

Note: wbTestErr is N/A for Ridge, as every lambda yields the same number of non-zero coefficients.

It seems that ridge regression performs worse than lasso regression for both datasets. For Dataset A (randomly generated data) increasing lambda minimized test error in both lasso and ridge regression. For Dataset B (cloud data) increasing lambda did not necessarily minimize test error. For lasso regression, minimized test error happened when lambda was 0.3, and there were 2 non-zero coefficients. For ridge regression, minimized test error happened when lambda was 20.

4) i) parameter turning (values filled in is test error -- not in %)

KERNEL	C = 0.01	C = 0.1	C = 1	C = 20	C = 50
Linear	0.2370	0.0420	0.0148	0.0099	0.0074
Polynomial	0.0395	0.0074	0.0025	0.0025	0.0025
Gaussian Radial Basis Function	0.2370	0.0420	0.0148	0.0099	0.0074

4) iii) 4-fold cross validation:

Average error for each learning algorithm using best C value from table above:

Linear w/ C = 50: 1.28%
Polynomial w/ C = 50: 0.43%
RBF w/ C = 50: 0.48%