Assignment 5 – Regression *

Alex Clemmer

April 4, 2012

1 Singular Value Decomposition

k	$\|\|M - Mk\|\|_2$
1	9.168090
2	8.537161
3	7.714555
4	5.650823
5	5.273193
6	5.073055
7	4.686644
8	4.639646
9	4.422348
10	4.303812

Experimentally, the "sweet spot" where $||M - Mk||_2 \approx 0.1 \cdot ||M||_2$ happens in the range of 68-71. As you can see they all occur reasonably close to the 10% threshold:

k	$\|\|M\|\|_2$	$\|\|M - Mk\|\|_2$	$\frac{\|\|M - Mk\|\|_2}{\|\|M\|\|_2}$
68	11.065524	1.210348	0.109380
69	11.065524	1.162212	0.105030
70	11.065524	1.132387	0.102335
71	11.065524	1.072011	0.096878

2 Column Sampling

A: We will see the distinctive downward trend, which is what we expect:

*CS 6955 Data Mining; Spring 2012 Instructor: Jeff M. Phillips, University of Utah

t	Type 1 error
1	9.190782
2	9.183018
3	8.577500
4	8.565876
5	8.524149
6	7.726940
7	7.725883
8	7.725882
9	7.725854
10	7.725518
11	7.724697
12	7.724691
13	7.582239
14	7.541526
15	7.533645
16	7.532958
17	7.532932
18	7.532928
19	7.508738
20	7.479237
21	7.463557
22	7.444950
23	7.444944
24	7.443253
25	7.443088
26	7.443010
27	7.390979
28	7.361148
29	7.357355
30	7.356612

t	Type 2 error
1	9.190782
2	9.186633
3	9.186633
4	9.186606
5	9.186605
6	9.186597
7	7.725978
8	7.053509
9	7.053187
10	7.045771
11	7.038306
12	7.029954
13	6.709284
14	6.505643
15	5.655842
16	5.637149
17	5.636823
18	5.634222
19	5.632514
20	5.013471
21	5.012855
22	5.011494
23	5.011477
24	5.009456
25	5.003089
26	5.002817
27	5.002699
28	5.002577
29	4.880626
30	4.774794

B: Empirically it is shown that for Type 1, the t value equivalent to the SVD error at $k = 5$ occurs between 52 and 53:

t	Type 1 error
52	7.211431
53	4.025061

For Type 2, we can see (in the table from part A) that the equivalent t value occurs between 19 and 20.

C: This will depend slightly on which t you pick; for Type 1 error, I picked $t = 53$ (since it's closer to SVD error for $k = 5$), and for Type 2 error I picked $t = 20$ (again, since it's closer).

To find the estimated number of 1's in for Type 1 error $t = 53$, we just sum up each of the top t columns in $P * M$ – we get 524. For reference, the actual number of non-zero entries in M for these t columns is (surprisingly) 524.

To get the estimated number of 1's for Type 2 error $t = 20$, we do the same thing. This time we get 200.78. The actual number of 1's in these t columns is 228, so this method is a little bit off.

For the top $k = 5$ columns of Uk (the so-called $U5$ matrix), the number of nonzero entries is 27. So, as we can see, the SVD requires dramatically less data than regression (no surprise there).

3 Linear Regression

A: First, the case of *least squares regression*, the error is given by $||Y - XA||_2 = 2.646467$. The case of *ridge regression* is slightly more complicated, since it depends on some coefficient s:

s	$\|\|Y - XA\|\|_2$
0.1	2.647344
0.3	2.653911
0.5	2.665972
1.0	2.714202
2.0	2.855114

B: First, the case of *least squares*:

Subset of X used	Subset of Y used	cross-validated error
$X(1:8,:)$	$Y(1:8)$	3.8681
$X(3:10,:)$	$Y(3:10)$	3.1435
$[X(1:4,:); X(7:10,:)]$	$[Y(1:4); Y(7:10)]$	5.3077

Now the (more complicated) problem of *ridge regression*. For the case where our subset of X is given by $X1 = X(1:8,:)$, and our subset of Y is given by $Y1 = Y(1:8)$:

s	cross-validated error
0.100000	3.637044
0.300000	3.246969
0.500000	2.935846
1.000000	2.408336
2.000000	1.970086

For the case where our subset of X is given by $X2 = X(3:10,:)$, and our subset of Y is given by $Y2 = Y(3:10)$:

s	cross-validated error
0.100000	3.108616
0.300000	3.041127
0.500000	2.976517
1.000000	2.826331
2.000000	2.566478

Finally, for the case where our subset of X is given by $X3 = [X(1:4,:); X(7:10,:)]$, and our subset of Y is given by $Y3 = [Y(1:4); Y(7:10)]$:

s	cross-validated error
0.100000	5.422124
0.300000	5.604590
0.500000	5.745986
1.000000	6.000867
2.000000	6.325721