Asmt 5: Regression

Gopal Menon
Turn in through Canvas by 2:45pm:
Wednesday, April 12

1 Singular Value Decomposition (20 points)

First we will compute the SVD of the matrix A we have loaded

$$[U, S, V] = svd(A)$$

Then take the top k components of A for values of $k = 1$ through $k = 10$ using

$$\begin{aligned} Uk &= U(:, 1:k) \\ Sk &= S(1:k, 1:k) \\ Vk &= V(:, 1:k) \\ Ak &= Uk * Sk * Vk' \end{aligned}$$

A: (10 points): Compute and report the L_2 norm of the difference between A and Ak for each value of k using

$$norm(A - Ak, 2)$$

Table 1: L_2 norm of $A - Ak$ for each value of k

k	L_2 Norm
1	40.483
2	26.717
3	25.000
4	22.192
5	17.675
6	15.813
7	13.351
8	12.188
9	9.1206
10	9.0000

B (5 points): Find the smallest value k so that the L_2 norm of $A - Ak$ is less than 10% that of A; k might or might not be larger than 10.

The L_2 norm of A is 120.19 and 10% of that is 12.019. From table 1, we can see that the smallest value of k such that the L_2 norm of $A - Ak$ is less than 10% that of A is 9.

C (5 points): Treat the matrix as 1125 points in 30 dimensions. Plot the points in 2 dimensions in the way that minimizes the sum of residuals squared.

The first two right singular vectors were used and all 1125 points in 30 dimensions were projected on them to get 1125 points in 2 dimensions. Since the first two singular vectors represent eigen vectors, this projection will result in the least sum of residuals squared. The plot is shown below in figure 1.

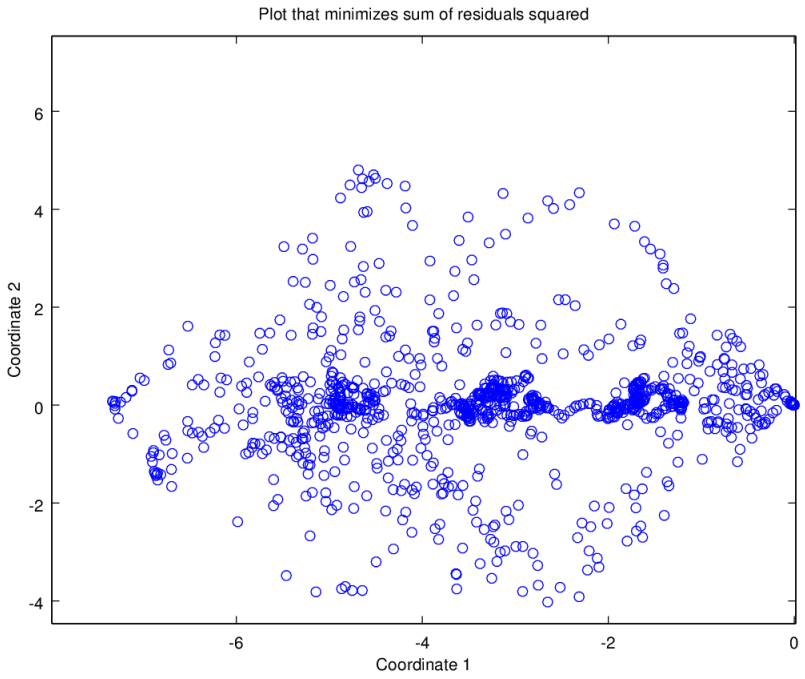

Figure 1: Points plotted in 2D to minimize sum of residuals squared

2 Frequent Directions and Random Projections (40 points)

A (20 points):

- How large does l need to be for the above error to be at most $\frac{\|A\|_F^2}{10}$?

$$\frac{\|A\|_F^2}{10} = 1903.7$$

Table 2: Error for values of l

l	Error
3	2289.9
4	1427.9
5	965.01
6	703.24
7	494.53
8	350.53
9	247.37
10	176.91

From table 2 we can see that with $l = 4$, the error is at most $\frac{\|A\|_F^2}{10}$.

- How does this compare to the theoretical bound (e.g. for $k = 0$).

The theoretical bound is given by $\frac{\|A-A_k\|_F^2}{l-k}$. When $k = 0$, the bound becomes $\frac{\|A\|_F^2}{l}$. This bound evaluates to 4759.3 when $l = 4$.

- How large does l need to be for the above error to be at most $\frac{\|A-A_k\|_F^2}{10}$ for $k = 2$?

  For $k = 2$, $\frac{\|A-A_k\|_F^2}{10} = 295.18$. From table 2, we can see that when $l = 9$, the error is at most $\frac{\|A-A_k\|_F^2}{10}$ for $k = 2$.

B (20 points): Estimate how large should l be in order to achieve $max_{\|x\|=1} \left| \|Ax\|^2 - \|Bx\|^2 \right| \leq \frac{\|A\|_F^2}{10}$. To estimate the relationship between l and the error in this randomized algorithm, you will need to run multiple trials. Be sure to describe how you used these multiple trials, and discuss how many you ran and why you thought this was enough trials to run to get a good estimate.

In the random projections algorithm, we want a mapping $\mu : \mathbb{R}^d \to \mathbb{R}^l$ with $l \ll d$ where all $\mathbb{R}^d$ is compressed into $\mathbb{R}^l$. And we want all distances preserved so that for all points $a, a' \in A$

$$(1-\varepsilon)\|a-a'\| \leq \|\mu(a)-\mu(a')\| \leq (1+\varepsilon)\|a-a'\|$$

The *Johnson-Lindenstrauss Lemma* says that if the reduced number of dimensions $l = O\left(\frac{1}{\varepsilon^2}\log\left(\frac{n}{\delta}\right)\right)$, then for all $a, a' \in A$ the above bound is satisfied with a probability of at least $1-\delta$.

If $\varepsilon = 0.10$ and $\delta = 0.05$, then the number of rows we need to choose for the above bound to be satisfied with a probability of at least 0.95 is

$$\begin{aligned} l &= O\left(\frac{1}{\varepsilon^2}\log\left(\frac{n}{\delta}\right)\right) \\ &= O\left(\frac{1}{0.10^2}\log\left(\frac{1125}{0.05}\right)\right) \\ &= O\left(100\log\left(22500\right)\right) \\ &= O\left(435.22\right) \end{aligned}$$

Table 3: Errors for number of rows l in the random projection

Number of rows l	Error $= norm(A' * A - B' * B, 2)$
430	1135.41
431	1175.94
432	1177.35
433	1511.90
434	610.41
435	696.70
436	1401.60
437	1315.81
438	287.31
439	621.36
440	435.97

The errors for various values of l around 435 are shown in table 3. In all the cases it is seen to be less than $\frac{\|A\|_F^2}{10}$, which was already computed to be 1903.7.

3 Linear Regression (40 points)

A (20 points): Solve for the coefficients C (or Cs) using Least Squares and Ridge Regression with $s = 0.1, 0.3, 0.5, 1.0, 2.0$ (i.e. s will take on one of those 5 values each time you try, say obtaining $C05$ for $s = 0.5$). For each set of coefficients, report the error in the estimate $\hat{Y}$ of Y as $norm(Y - X * C, 2)$.

Table 4: Least Squares Regression Error

Error
25.823822

Table 5: Ridge Regression Error

s	Error
0.1	25.823824
0.3	25.823943
0.5	25.824676
1.0	25.833907
2.0	25.919640

B (20 points): Create three row-subsets of X and Y

- $X1 = X(1 : 66, :)$ and $Y1 = Y(1 : 66)$
- $X2 = X(34 : 100, :)$ and $Y2 = Y(34 : 100)$
- $X3 = [X(1 : 33, :); X(67 : 100, :)]$ and $Y3 = [Y(1 : 33); Y(67 : 100)]$

Repeat the above procedure on these subsets and *cross-validate* the solution on the remainder of X and Y. Specifically, learn the coefficients C using, say, $X1$ and $Y1$ and then measure $norm(Y(67 : 100) - X(67 : 100, :) * C, 2)$.

Table 6: Least Squares Regression Error

Training Subsets	Error
$X1, Y1$	14.787131
$X2, Y2$	14.525954
$X3, Y3$	23.194199
Average	17.502428

Table 7: Ridge Regression Error

Training Subsets	$Error_{s=0.1}$	$Error_{s=0.3}$	$Error_{s=0.5}$	$Error_{s=1.0}$	$Error_{s=2.0}$
$X1, Y1$	14.785569	14.773048	14.747842	14.627345	14.167225
$X2, Y2$	14.525810	14.525195	14.526353	14.555734	14.809322
$X3, Y3$	23.195206	23.203230	23.219111	23.292336	23.610916
Average	17.502195	17.500491	17.497769	17.491805	17.529154

Which approach works best (averaging the results from the three subsets): Least Squares, or for which value of s using Ridge Regression?

The approach that worked best was for Ridge Regression with $s = 1.0$, which had the lowest average error.