

Asmt 5: Regression

Turn in through Canvas by 5pm:

Monday, April 11

100 points

Overview

In this assignment you will explore regression techniques on high-dimensional data.

You will use a few data sets for this assignment:

- <http://www.cs.utah.edu/~jeffp/teaching/cs5140/A5/A.dat>
- <http://www.cs.utah.edu/~jeffp/teaching/cs5140/A5/X.dat>
- <http://www.cs.utah.edu/~jeffp/teaching/cs5140/A5/Y.dat>
- <http://www.cs.utah.edu/~jeffp/teaching/cs5140/A5/M.dat>
- <http://www.cs.utah.edu/~jeffp/teaching/cs5140/A5/W.dat>

and a file stub:

- <http://www.cs.utah.edu/~jeffp/teaching/cs5140/A5/FD.m>

These data sets are in matrix format and can be loaded into MATLAB or OCTAVE. By calling `load filename` (for instance `load X.dat`) it will put in memory the data in the file, for instance in the above example the matrix X . You can then display this matrix by typing

X

As usual, it is highly recommended that you use LaTeX for this assignment. If you do not, you may lose points if your assignment is difficult to read or hard to follow. Find a sample form in this directory: <http://www.cs.utah.edu/~jeffp/teaching/latex/>

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

```
Uk = U(:, 1:k)
```

```
Sk = S(1:k, 1:k)
```

```
Vk = V(:, 1:k)
```

```
Ak = Uk*Sk*Vk'
```

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

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

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

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.

2 Frequent Directions (40 points)

Use the stub file `FD.m` to create a function for the Frequent Directions algorithm (**Algorithm 15.2.1**). We will consider running this code on matrix A .

A (15 points): We can measure the error $\max_{\|x\|=1} |\|Ax\|^2 - \|Bx\|^2|$ as $\text{norm}(A' * A - B' * B, 2)$. How large does λ need to be for the above error to be at most $\|A\|_F^2/10$? How does this compare to the theoretical bound (e.g. for $k = 0$).

Note: you can calculate $\|A\|_F^2$ as $\text{norm}(A, 'fro')^2$.

B (25 points): Frequent Directions should also satisfy another bound based on its Frobenius norm. We can compute $A\Pi_{B_k}$ using $B_k = B(1:k, :)$ and then calculating $A * \text{pinv}(B_k) * B_k$. How large does λ need to be to achieve

$$\|A - A\Pi_{B_k}\|_F^2 \leq 1.1 \cdot \|A - A_k\|_F^2;$$

for each of the 7 values of value k in the set $\{1, 2, 3, 4, 5, 6, 7\}$. Answer both by running your algorithm and reporting the theoretical bound provided in the notes. (e.g., you should report 7 pairs of values, an empirical and theoretical bound for each value k)

3 Linear Regression (40 points)

We will find coefficients C (was a_1, \dots, a_d in notes, but changed to avoid confusion with matrix A in **Q1**) to estimate $X * C \approx Y$, using the provided datasets X and Y . We will compare two approaches *least squares* and *ridge regression*.

Least Squares: $\text{Set } C = \text{inverse}(X' * X) * X' * Y$

Ridge Regression: $\text{Set } C_s = \text{inverse}(X' * X + s^2 * \text{eye}(12)) * X' * Y$

A (20 points): Solve for the coefficients C (or C_s) 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 $C_{0.5}$ for $s = 0.5$). For each set of coefficients, report the error in the estimate \hat{Y} of Y as $\text{norm}(Y - X * C, 2)$.

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

- $X_1 = X(1:66, :)$ and $Y_1 = Y(1:66)$
- $X_2 = X(34:100, :)$ and $Y_2 = Y(34:100)$
- $X_3 = [X(1:33, :); X(67:100, :)]$ and $Y_3 = [Y(1:33); Y(67:100)]$

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, X_1 and Y_1 and then measure $\text{norm}(Y(67:100) - X(67:100, :) * C, 2)$.

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

4 BONUS (3 points)

Consider a linear equation $\bar{W} = M * S$ where M is a measurement matrix filled with random values $\{-1, 0, +1\}$ (although now that they are there, they are no longer random), and \bar{W} is the output of the sparse signal S when measured by M .

Use Orthogonal Matching Pursuit (as described in the notes) to recover the non-zero entries from S . Record the order in which you find each entry and the residual vector after each step.