Asmt 5: Regression

Turn in through Canvas by 5pm:
Wednesday, April 09
20 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 (3 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'
```
Compute and report the L_2 norm of the difference between A and Ak for each value of k using
```
norm(A-Ak,2)
```
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.

2 Frequent Directions (10 points)

Use the stub file `FD.m` to create a function for the Frequent Directions algorithm (**Algorithm 16.2.1**). We will consider running this code on matrix `A`.

A (4 points): We can measure the error $\max_{\|x\|=1} |\|Ax\|^2 - \|Bx\|^2|$ as `norm(A'*A - B'*B, 2)`. How large does `l` 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 `norm(A, 'fro')^2`.

B (6 points): Frequent Directions should also satisfy another bound based on its Frobenious norm. We can compute $A\Pi_{B_k}$ using `Bk = B(1:k,:)` and then calculating `A * Bk' * pinv(Bk * Bk') * Bk`. How large does `l` need to be to achieve

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

for each value $k \in \{1, 2, 3, 4, 5, 6, 7\}$. Answer both by running your algorithm and reporting the theoretical bound provided in the notes.

3 Linear Regression (7 points)

We will find coefficients `A` (not the same as the loaded matrix) to estimate `X*A = Y`. We will compare two approaches *least squares* and *ridge regression*.

Least Squares: Set `A = inverse(X' * X)*X'*Y`

Ridge Regression: Set `As = inverse(X'*X + s*eye(6))*X'*Y`

A (3 points): Solve for the coefficients A (or As) using Least Squares and Ridge Regression with $s = \{0.1, 0.3, 0.5, 1.0, 2.0\}$. For each set of coefficients, report the error in the estimate $\hat{Y}$ of Y as `norm(Y - X*A,2)`.

B (4 points): Create three row- subsets of `X` and `Y`

- `X1 = X(1:8,:)` and `Y1 = Y(1:8)`
- `X2 = X(3:10,:)` and `Y2 = Y(3:10)`
- `X3 = [X(1:4,:); X(7:10,:)]` and `Y3 = [Y(1:4); Y(7:10)]`

Repeat the above procedure on these subsets and *cross-validate* the solution on the remainder of `X` and `Y`. Specifically, learn the coefficients `A` using, say, `X1 and Y1` and then measure `norm(Y(9:10) - X(9:10,:)*A,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 `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 `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.