Asmt 6: Regression

Turn in through Canvas by 2:45pm:
Wednesday, March 27
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/A6/X.dat`
- `http://www.cs.utah.edu/~jeffp/teaching/cs5140/A6/Y.dat`
- `http://www.cs.utah.edu/~jeffp/teaching/cs5140/A6/M.dat`
- `http://www.cs.utah.edu/~jeffp/teaching/cs5140/A6/W.dat`

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`

For python, you can for instance use the following approach:

```
def read(path):
    reval = []
    with open(path, encoding='utf8') as f:
        for line in f:
            if len(line.strip()) == 0:
                continue
            s = line.strip().split()
            vector = [float(v) for v in s]
            reval.append(vector)
    return np.array(reval)
```

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 Linear Regression (75 points)

We will find coefficients `A` to estimate `X*A` $\approx$ `Y`, using the provided datasets `X` and `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^2*eye(15))*X'*Y`

A (30 points): Solve for the coefficients `A` (or `As`) using Least Squares and Ridge Regression with $s = \{1, 5, 10, 15, 20, 25, 30\}$ (i.e. s will take on one of those 7 values each time you try, say obtaining `A05` for $s = 5$). For each set of coefficients, report the error in the estimate $\hat{Y}$ of Y as `norm(Y - X*A,2)`.

B (30 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 `A` using, say, `X1 and Y1` and then measure `norm(Y(67:100) - X(67:100,:)*A,2)`.

C (15 points): Which approach works best (averaging the results from the three subsets): Least Squares, or for which value of s using Ridge Regression?

2 Orthogonal Matching Pursuit (25 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 as **Algorithm 18.2.1**) to recover the non-zero entries from `S`. Record the order in which you find each entry and the residual vector after each step.

BONUS: L_1 ball geometry (2 points)

In class we examined pictures of L_1-balls in $\mathbb{R}^2$, but in higher dimensions their geometry is more wild. Its boundary has "facets" of k-dimensions for each $k \in \{0, \ldots, d\}$, these are linear subspaces of of rank k.

As a function of d and k, derive how many different facets there are of k-dimensions when on the boundary of L_1-balls in $\mathbb{R}^d$. Explain how you derive your answer.