Data mining: Homework 1

Vishvas Vasuki

September 21, 2009

1 1

1.1 Notation

Given a training set $\left\{(x_i, y_i)_{i=1}^N\right\}$. Let x be the vector of (x_i); and let y be the corresponding (y_i) vetor. Construct the $N \times 2$ matrix X = [1 x]. Take $w = (w_0, w_1)$.

Below, we use the symbols $\bar{x}, \bar{y}, \sigma_{xy}, \sigma_{xx}$ as defined in the question.

1.2 a

Now, we want to solve the least squares problem: $Xw \approx y$.

Forming the normal equations, we get:

$$\begin{aligned}
X^T X w &= X^T y \\
\begin{pmatrix} N & \sum x_i \\ \sum x_i & x^T x \end{pmatrix} w &= \begin{pmatrix} \sum y_i \\ x^T y \end{pmatrix} \\
\begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \sum x_i^2/N \end{pmatrix} w &= \begin{pmatrix} \bar{y} \\ \sum x_i y_i/N \end{pmatrix} \\
\begin{pmatrix} 1 & \bar{x} \\ 0 & \sum x_i^2/N - \bar{x}^2 \end{pmatrix} w &= \begin{pmatrix} \bar{y} \\ \sum x_i y_i/N - \bar{x}\bar{y} \end{pmatrix}
\end{aligned}$$

We have carried out Gaussian elimination above. Solving these equations for w, after some algebra, we find:

$$\begin{aligned}
w_1 &= \frac{\sum x_i y_i/N - \bar{x}\bar{y}}{\sum x_i^2/N - \bar{x}^2} \\
&= \frac{N^{-1}\sum(x_i - \bar{x})(y_i - \bar{y})}{N^{-1}\sum(x_i - \bar{x})^2} \\
&= \frac{\sigma_{xy}}{\sigma_{xx}} \\
w_0 &= \bar{y} - w_1 \bar{x}
\end{aligned}$$

1.3 b

In general, if $x_i \in R^d$: If $x_{i,j}$ is the jth component of the vector x_i, and if $\bar{x_j} = \frac{\sum_{i=1}^{N} x_{i,j}}{N}$: $w_0 = \bar{y} - \sum_{i=1}^{d} w_i \bar{x_j}$.

1.3.1 Proof

Consider the $N*(d+1)$ matrix $X = [1\ x_i^T]$, so that $X_{i,1} = 1, X_{i,j} = x_{i,j-1} \forall j > 1$. Then, forming the normal equations: $N^{-1}X^T X w = N^{-1} X^T y$.

Examine the first row of $N_{-1}X^T X$: you have $[1\ \bar{x_1}..\bar{x_d}]$, and the first element in $N^{-1}X^T y$ is $\bar{y}$.

2 2

The proof is invalid as $\sum_{i=0}^{\infty} \beta^i A^i$ does not converge $\forall \beta \geq 0$ and multiplyication by ∞ is not well defined. But this will not happen if $\beta < \frac{1}{nt}$ where $\max_{i,j} |A_{i,j}| = t$.

2.1 Proof

Note that A, being an adjascency matrix, satisfies: $A_{i,j} \geq 0$.

Then take $a = \beta n t$. The series $\sum_{i=0}^{\infty} a^i$ is a geometric series which converges to $\frac{1}{1-a}$.

Now consider a matrix $T \in R^{n \times n}$: $T_{i,j} = t$. Now, $\sum_{i=0}^{\infty} T^i$ converges as $T_{j,k}^i = a^{i-1} \forall i > 2$.

Now, $A_{i,j} \leq T_{i,j} = t$. So, the series corresponding to the (i,j)th element of the matrix sum: $\sum_k \beta^k A_{i,j}^k$ is bounded, and this series is non-decreasing. Hence, the series, being non-decreasing and bounded, is convergent.

3 3

The code used:

```
load dataset1
Xtrain = [ones(30,1) Xtrain]
Xtest = [ones(120,1) Xtest]

A = Xtrain'*Xtrain
b = Xtrain'*Ytrain

w = A\b
sqError = (Ytrain -Xtrain*w)'*(Ytrain -Xtrain*w)
sqrt(sqError/30)
sqError = (Ytest -Xtest*w)'*(Ytest -Xtest*w)
sqrt(sqError/120)
```

```
[U S V] = svd(Xtrain)


w = V*inv(S'*S)*V'*Xtrain'*Ytrain
sqError = (Ytrain -Xtrain*w)'*(Ytrain -Xtrain*w)
sqrt(sqError/30)
sqError = (Ytest -Xtest*w)'*(Ytest -Xtest*w)
sqrt(sqError/120)

load dataset2
Xtrain = [ones(30,1) Xtrain]
Xtest = [ones(120,1) Xtest]

A = Xtrain'*Xtrain
b = Xtrain'*Ytrain

w = A\b
sqError = (Ytrain -Xtrain*w)'*(Ytrain -Xtrain*w)
sqrt(sqError/30)
sqError = (Ytest -Xtest*w)'*(Ytest -Xtest*w)
sqrt(sqError/120)

[U S V] = svd(Xtrain)
T=(S(1:4,1:4)*S(1:4,1:4))
b = inv(T)*V(:,1:4)'*Xtrain'*Ytrain

w = V(:,1:4)'\b
sqError = (Ytrain -Xtrain*w)'*(Ytrain -Xtrain*w)
sqrt(sqError/30)
sqError = (Ytest -Xtest*w)'*(Ytest -Xtest*w)
sqrt(sqError/120)
```

3.1 a

RMS Error by solving Normal equations: Training error: 0.1566. Test error: 0.1726.

RMS Error using SVD: Training error: 0.1566. Test error: 0.1726.

3.2 b

RMS Error by solving Normal equations: Training error: 0.1576. Test error: 0.1822.

RMS Error using SVD: Training error: 0.1566. Test error: 0.1726.

Please see the attached code to see how I use the SVD: I drop the columns of V corresponding to $\sigma_i \approx 0$.

Using SVD, as expected, yields more accurate results than using LU.

The second data-set contains no extra information compared to the first data-set. So, SVD's performance on the two datasets is identical.

4 4

4.1 Notation

$x = (x_1, x_2, x_3)$ represents height, weight and age of data-point x.

Bob uses the units: inches, pounds, months. Alice uses the units centimeters, kilograms, days.

4.1.1 Assumption about w and the linear model

We assume that $w \in R^{3+1}$. w is indexed from 0 to 3. The linear model is $y \approx w_0 + \sum_{i=1}^{3} w_i x_i$.

4.1.2 The diagonal matrix D

Let x_A, x_B be the observations of the same data point, as measured by Alice and Bob. Then, $x_A^T = x_B^T D'$ where D' is a diagonal matrix expressing the factors which relate the units used by Alice to the units used by Bob, like cm/in etc..

Arrange the observations of various data points $\{x\}$ by Alice and Bob as rows in the matreces A and B, but ensure that the first column of both these matrices is 1. Note that $A = BD$, where $D = \begin{pmatrix} 1 & 0 \\ 0 & D' \end{pmatrix}$.

4.2 a

The normal equations are $A^T A w = A^T y$ and $B^T B z = B^T y$ for Alice and Bob. The former can be rewritten as $D^T B^T B D w = D^T B^T y$ or $B^T B D w = B^T y$. Thus, we see that $Dw = z$ or $w = D^{-1} z$.

This tells us the relationship between z and w, the solutions of Bob and Alice to the least squares problem.

4.3 b

The regularization part of the objective in the ridge regression problem should ideally not include w_0. We assume that this is the case with the question, and that $\lambda \|w\|_2$ term considers only the $w_1..w_3$ terms. For this reason, we rewrite the objective as: $\min_w \|Xw\| + w^T I' w$, where $I' = \begin{pmatrix} 0 & 0 \\ 0 & I_3 \end{pmatrix}$.

$(A^T A + \lambda I')w = A^T y$ and $(B^T B + \lambda I')z = B^T y$ express the solution to the ridge regression problem. The former can be rewritten as : $(D^T B^T B D + \lambda I')w = D^T B^T y$.

So, $(D^T B^T B D + \lambda I')w = D^T (B^T B + \lambda I')z$. This tells us the relationship between z and w, the solutions of Bob and Alice to the ridge regression problem.

4.4 c

Suppose the label vector in the training set is changed to $\bar{y} = y + 1$. Let the modified least squares solution be w'. As we noted in the first problem's solution, $w'_0 = \frac{\sum \bar{y_i}}{N} - \sum_{i=1}^{d} w_i \bar{x_j}$. $w'_0 = 1 + \frac{\sum y_i}{N} - \sum_{i=1}^{d} w_i \bar{x_j} = 1 + w_0$.

However, $\forall i > 0, w'_i = w_i$. This equality does not change when ridge regression is applied, as in the regularizer, w_0 is omitted.