CS 391D Data Mining: A Mathematical Perspective Fall 2009

Solutions to Homework 1

Lecturer: Inderjit Dhillon Date Due: September 22, 2009

Keywords: *Linear Algebra, Linear Regression*

1. (5 points)

   (a) (3 points) Let $\boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$, $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$, $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$, and $\mathbf{1} \in \mathbb{R}^N$ with all elements equal to one.

   The normal equations can be written as:

$$\begin{bmatrix} \mathbf{1}^T \\ \boldsymbol{x}^T \end{bmatrix} \begin{bmatrix} \mathbf{1} & \boldsymbol{x} \end{bmatrix} \boldsymbol{w} = \begin{bmatrix} \mathbf{1}^T \\ \boldsymbol{x}^T \end{bmatrix} \boldsymbol{y} \Rightarrow \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{N}\boldsymbol{x}^T\boldsymbol{x} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \frac{1}{N}\boldsymbol{x}^T\boldsymbol{y} \end{bmatrix}. \tag{1}$$

   From above, we can get

$$w_0 = \bar{y} - w_1\bar{x}, \tag{2}$$

   and

$$\begin{aligned} w_1 &= \frac{\frac{1}{N}\sum_{i=1}^N x_i y_i - \bar{x}\bar{y}}{\frac{1}{N}\sum_{i=1}^N x_i^2 - \bar{x}^2} = \frac{\frac{1}{N}\sum_{i=1}^N x_i y_i - (\frac{1}{N}\sum_{i=1}^N x_i)\bar{y} - \bar{x}(\frac{1}{N}\sum_{i=1}^N y_i) + \bar{x}\bar{y}}{\frac{1}{N}\sum_{i=1}^N x_i^2 - (\frac{1}{N}\sum_{i=1}^N x_i)\bar{x} - \bar{x}(\frac{1}{N}\sum_{i=1}^N x_i) + \bar{x}^2} \\ &= \frac{\frac{1}{N}\sum_{i=1}^N (x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y})}{\frac{1}{N}\sum_{i=1}^N (x_i^2 - x_i\bar{x} - \bar{x}x_i + \bar{x}^2)} = \frac{\sigma_{xy}}{\sigma xx} \end{aligned} \tag{3}$$

   (b) (2 points) Similarly, let $\boldsymbol{w}' = \begin{bmatrix} w_0 \\ \boldsymbol{w} \end{bmatrix}$, where $\boldsymbol{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$, and $X = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_N^T \end{bmatrix} \in \mathbb{R}^{N\times d}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$.

   The normal equations can be written as:

$$\begin{bmatrix} \mathbf{1}^T \\ X^T \end{bmatrix} \begin{bmatrix} \mathbf{1} & X \end{bmatrix} \begin{bmatrix} w_0 \\ \boldsymbol{w} \end{bmatrix} = \begin{bmatrix} \mathbf{1}^T \\ X^T \end{bmatrix} \boldsymbol{y} \Rightarrow \begin{bmatrix} 1 & \bar{\boldsymbol{x}}^T \\ \bar{\boldsymbol{x}} & \frac{1}{N}X^TX \end{bmatrix} \begin{bmatrix} w_0 \\ \boldsymbol{w} \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \frac{1}{N}X^T\boldsymbol{y}. \end{bmatrix} \tag{4}$$

   From above, we can get

$$w_0 = \bar{y} - \bar{\boldsymbol{x}}^T\boldsymbol{w}, \tag{5}$$

   and $\boldsymbol{w}$ can be solved from

$$(X^TX - N\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T)\boldsymbol{w} = (X^T - \bar{\boldsymbol{x}}\mathbf{1}^T)\boldsymbol{y}. \tag{6}$$

2. (4 points) The proof is not correct since $\sum_{i=1}^{\infty} \beta^i A^i$ might diverge.

   Suppose $A \in \mathbb{R}^{N\times N}$, since A is symmetric (for undirected graph), the eigenvalue decomposition of A will be $A = U\Lambda U^T$, where $U^TU = I$ and Λ is the diagonal matrix with the eigenvalues $\{\lambda_j\}_{j=1}^N$ of A on the diagonal.

   Therefore,

$$\sum_{i=1}^{\infty} \beta^i A^i = U(\sum_{i=1}^{\infty} \beta^i \Lambda^i)U^T. \tag{7}$$

Let λ' denote the eigenvalue with the largest absolute value. In order to ensure $\sum_{i=1}^{\infty} \beta^i A^i$ converge, $|\beta\lambda'| < 1$ must be satisfied, which implies $\beta < 1/|\lambda'|$.

3. (6 points)

   (a) (3 points) The normal equations are $\hat{X}^T \hat{X} \boldsymbol{w} = \hat{X}^T \boldsymbol{y}$, where $\hat{X} = \begin{bmatrix} \mathbf{1} & X \end{bmatrix}$. So the coefficient vector $\boldsymbol{w}$ can be solved in Matlab as: $\boldsymbol{w} = \hat{X}^T \hat{X} \backslash \hat{X}^T \boldsymbol{y}$ by using the Matlab "\" operator. The resulting coefficient vector is

```
w_normal =
     9.380296842426794e-01
    -2.197506989029683e-01
    -1.092679523646183e+00
     2.722846226418876e-01.
```

   The RMSE on the training/testing set:

```
TrainErr = 1.566269622399142e-01
TestErr =  1.725918643729227e-01.
```

   By using SVD, we first compute the Singular Value Decomposition of matrix $\hat{X}$:

```
[U,S,V] = svd(X_hat, 0);
```

   where the singular values are

```
diag(S) =
     8.056021983474565e+00
     1.769863323706678e+00
     1.199186168087752e+00
     4.150522541340332e-01.
```

   Therefore, the coefficient vector $\boldsymbol{w} = VS^{-1}U^T\boldsymbol{y}$, which is

```
w_svd =
     9.380296842426810e-01
    -2.197506989029685e-01
    -1.092679523646186e+00
     2.722846226418890e-01.
```

   The RMSE on the training/testing set:

```
TrainErr = 1.566269622399142e-01
TestErr =  1.725918643729227e-01.
```

   (b) (3 points) Similarly, by solving the normal equations $\hat{X}^T \hat{X} \boldsymbol{w} = \hat{X}^T \boldsymbol{y}$, we get

```
w_normal =
     8.543650856508991e-01
     2.007685445231479e+06
    -3.143921286530009e+06
     2.007685865462310e+06
    -4.015371281250000e+06
     3.143920312500000e+06.
```

   The RMSE on the training/testing set:

```
TrainErr = 1.638103624995622e-01
TestErr =  1.962736211669317e-01.
```

In case of solving by SVD, the singular values are

```
diag(S) =
     9.507897664011070e+00
     1.883697351038450e+00
     1.330370950897240e+00
     5.158720698357471e-01
     6.505639266509868e-08
     5.081070284793642e-08.
```

Note that the last two singular values are quite small, which implies that the matrix $\hat{X}$ is close to being rank deficient. If we keep all singular values when solving the coefficient vector $\boldsymbol{w}$, we will get some values in $\boldsymbol{w}$ with very large magnitude.

```
w_svd =
     9.098900750486882e-01
     8.481804011589671e+05
    -6.162979188546608e+05
     8.481808681791546e+05
    -1.696361215122565e+06
     6.162968577463540e+05.
```

The RMSE on the training/testing set:

```
TrainErr = 1.545990226733030e-01
TestErr =  1.782932794388689e-01.
```

The correct way of using SVD is to drop the singular values which are close to zero. In this case, the resulting coefficient vector does not have any large values:

```
w_svd =
     9.380296974205432e-01
    -2.285063480290051e-01
    -5.463397867121318e-01
     2.635289825394068e-01
     1.751129730921534e-02
    -5.463397608584619e-01.
```

The RMSE on the training/testing set:

```
TrainErr = 1.566269604102106e-01
TestErr =  1.725918644908762e-01
```

4. (6 points)

   (a) (2 points) Suppose X_A is measured by Alice and X_B is measured by Bob. Let $\hat{X}_A$ denote $\begin{bmatrix}\mathbf{1} & X_A\end{bmatrix}$, $\hat{X}_B$ denote $\begin{bmatrix}\mathbf{1} & X_B\end{bmatrix}$, and $\hat{D}$ denote $\begin{bmatrix}1 & 0\\ 0 & D\end{bmatrix}$, where D is a diagonal matrix with diagonal entries describing the difference between measurements. The relationship between these two measures can be characterized by $\hat{X}_B = \hat{X}_A\hat{D}$, By the normal equations, the coefficient vector obtained by Alice is

$$\boldsymbol{w}_A = (\hat{X}_A^T \hat{X}_A)^{-1} \hat{X}_A^T \boldsymbol{y}, \tag{8}$$

and the coefficient vector obtained by Bob is

$$\boldsymbol{w}_B = (\hat{X}_B^T \hat{X}_B)^{-1} \hat{X}_B^T \boldsymbol{y} = (\hat{D} \hat{X}_A^T \hat{X}_A \hat{D})^{-1} \hat{D} \hat{X}_A^T \boldsymbol{y} = \hat{D}^{-1} (\hat{X}_A^T \hat{X}_A)^{-1} \hat{X}_A^T \boldsymbol{y}, \tag{9}$$

which implies $\boldsymbol{w}_A = \hat{D} \boldsymbol{w}_B$.

(b) (2 points) Similarly, if Bob and Alice both solve the ridge regression problem, then by the normal equations, the coefficient vector obtained by Alice is

$$\boldsymbol{w}_A = (\hat{X}_A^T \hat{X}_A + \lambda I)^{-1} \hat{X}_A^T \boldsymbol{y}, \tag{10}$$

and the coefficient vector obtained by Bob is

$$\boldsymbol{w}_B = (\hat{X}_B^T \hat{X}_B + \lambda I)^{-1} \hat{X}_B^T \boldsymbol{y} = (\hat{D} \hat{X}_A^T \hat{X}_A \hat{D} + \lambda I)^{-1} \hat{D} \hat{X}_A^T \boldsymbol{y} = \hat{D}^{-1} (\hat{X}_A^T \hat{X}_A + \lambda \hat{D}^{-2})^{-1} \hat{X}_A^T \boldsymbol{y}. \tag{11}$$

Comparing (10) with (11), there is no explicit relationship between their coefficient vectors.

Note that if we do not include w_0 in the regularizer, the ridge regression solution will be changed to

$$\boldsymbol{w} = \left(\hat{X}^T \hat{X} + \lambda \begin{bmatrix} 0 & \\ & I \end{bmatrix} \right)^{-1} \hat{X}^T \boldsymbol{y}. \tag{12}$$

By following the same arguments above, we get

$$\boldsymbol{w}_A = \left(\hat{X}_A^T \hat{X}_A + \lambda \begin{bmatrix} 0 & \\ & I \end{bmatrix} \right)^{-1} \hat{X}_A^T \boldsymbol{y}, \quad \text{and} \quad \boldsymbol{w}_B = \hat{D}^{-1} \left(\hat{X}_A^T \hat{X}_A + \lambda \begin{bmatrix} 0 & \\ & D^{-2} \end{bmatrix} \right)^{-1} \hat{X}_A^T \boldsymbol{y}. \tag{13}$$

Again, there is no explicit relationship between their coefficient vectors.

(c) (2 points) Let $\boldsymbol{w}$ denote the coefficient vector obtained by using the original target variable $\boldsymbol{y}$, $\boldsymbol{w}'$ denote the coefficient vector obtained by using the new target variable $\boldsymbol{y}' = \boldsymbol{y} + \boldsymbol{1}$, and $\bar{\boldsymbol{x}}$ denote the mean vector of the data $\frac{1}{N} X^T \boldsymbol{1}$.

In the least squares problem, from problem 1(b), we have already solved by the normal equations that

$$w_0 = \frac{1}{N} \boldsymbol{1}^T \boldsymbol{y} - \bar{\boldsymbol{x}}^T \boldsymbol{w}, \quad \text{and} \quad (\frac{1}{N} X^T X - \bar{\boldsymbol{x}} \bar{\boldsymbol{x}}^T) \boldsymbol{w} = \frac{1}{N} (X^T - \bar{\boldsymbol{x}} \boldsymbol{1}^T) \boldsymbol{y} \tag{14}$$

If we replace $\boldsymbol{y}$ with $\boldsymbol{y}' = \boldsymbol{y} + \boldsymbol{1}$, then

$$w_0' = 1 + \frac{1}{N} \boldsymbol{1}^T \boldsymbol{y} - \bar{\boldsymbol{x}}^T \boldsymbol{w}', \tag{15}$$

and

$$(\frac{1}{N} X^T X - \bar{\boldsymbol{x}} \bar{\boldsymbol{x}}^T) \boldsymbol{w}' = \frac{1}{N} (X^T - \bar{\boldsymbol{x}} \boldsymbol{1}^T) \boldsymbol{y} + \frac{1}{N} (X^T \boldsymbol{1} - \bar{\boldsymbol{x}} \boldsymbol{1}^T \boldsymbol{1}) = \frac{1}{N} (X^T - \bar{\boldsymbol{x}} \boldsymbol{1}^T) \boldsymbol{y}. \tag{16}$$

Therefore, in the least squares problem, $w_0' = w_0 + 1$ and $\boldsymbol{w}' = \boldsymbol{w}$.

Similarly, in the ridge regression problem, the normal equations are

$$\begin{bmatrix} N & \boldsymbol{1}^T X \\ X^T \boldsymbol{1} & X^T X + \lambda I \end{bmatrix} \begin{bmatrix} w_0 \\ \boldsymbol{w} \end{bmatrix} = \begin{bmatrix} \boldsymbol{1}^T \\ X^T \end{bmatrix} \boldsymbol{y}, \tag{17}$$

where w_0 is not included in the regularizer.

From above, we can get

$$w_0 = \frac{1}{N}\mathbf{1}^T\boldsymbol{y} - \bar{\boldsymbol{x}}^T\boldsymbol{w}, \quad \text{and} \quad (\frac{1}{N}X^TX + \frac{\lambda}{N}I - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T)\boldsymbol{w} = \frac{1}{N}(X^T - \bar{\boldsymbol{x}}\mathbf{1}^T)\boldsymbol{y}. \tag{18}$$

Following the same arguments above, if we replace $\boldsymbol{y}$ with $\boldsymbol{y}' = \boldsymbol{y} + \mathbf{1}$, we can get $w_0' = w_0 + 1$ and $\boldsymbol{w}' = \boldsymbol{w}$.

Note that if w_0 is included in the regularizer, we will get different solutions of w_0 and $\boldsymbol{w}$ by simply increasing the target variable $\boldsymbol{y}$ by one. This partially explains why we normally do not put w_0 into the regularizer.