Homework 1

Instructor: Inderjit Dhillon Date Due: September 22, 2009

Keywords: *Linear Algebra, Linear Regression*

1. (5 points) Suppose we are given a training set $\{(x_i, y_i)\}_{i=1}^N$.

   (a) (3 points) Let $x_i \in \mathbb{R}$ be the ith data point and $y_i \in \mathbb{R}$ be the target value associated with x_i. Use a linear function of the form $w_0 + w_1 x$ to fit the data. Show that $w_0 = \bar{y} - w_1\bar{x}$ and $w_1 = \sigma_{xy}/\sigma_{xx}$ by using the normal equations, where $\bar{x} = \frac{1}{N}\sum_{i=1}^N x_i$, $\bar{y} = \frac{1}{N}\sum_{i=1}^N y_i$, $\sigma_{xy} = \frac{1}{N}\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$, and $\sigma_{xx} = \frac{1}{N}\sum_{i=1}^N (x_i - \bar{x})^2$.

   (b) (2 points) Suppose now that $x_i \in \mathbb{R}^d$, where $d > 1$, and $y_i \in \mathbb{R}$. What is w_0 in this general case?

2. (3 points) Given an adjacency matrix $A \in \mathbb{R}^{n\times n}$ for an undirected graph, the Katz measure, which is commonly used in social network analysis, is defined as

$$K = \sum_{i=1}^{\infty} \beta^i A^i, \tag{1}$$

   where $\beta \geq 0$ is a scalar parameter, also called the *damping factor*.

   Now we claim that $K = (I - \beta A)^{-1} - I$, for all $\beta \geq 0$.

   Proof: We verify that $I + K = (I - \beta A)^{-1}$.

$$\begin{aligned}(I-\beta A)(I+K) &= I - \beta A + \sum_{i=1}^{\infty}\beta^i A^i - \sum_{i=1}^{\infty}\beta^{i+1}A^{i+1}\\ &= I + \sum_{i=1}^{\infty}\beta^i A^i - (\beta A + \sum_{i=1}^{\infty}\beta^{i+1}A^{i+1})\\ &= I + \sum_{i=1}^{\infty}\beta^i A^i - \sum_{i=1}^{\infty}\beta^i A^i = I.\end{aligned}$$

   Therefore, $K = (I - \beta A)^{-1} - I$.

   Is the above proof correct? If not, is there any condition that β should satisfy?

3. (6 points) In this linear regression problem, you are given two data sets, each containing 150 instances with 120 instances for training and 30 instances for testing. Instances are given as $(\boldsymbol{x}_i, y_i)$, for $i = 1, \ldots, 150$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the ith data point and $y_i \in \mathbb{R}$ is the target value associated with $\boldsymbol{x}_i$. (You can obtain the data sets at `http://www.cs.utexas.edu/~wtang/cs391d/reg.tar.gz`).

   (a) (3 points) For Dataset 1, use the normal equations and SVD to solve the least squares problem, respectively. What are the solutions $\boldsymbol{w}$ you obtained? Do the two methods give similar solutions? What is the training error and testing error? The error can be measured by *Root Mean Squared Error* (RMSE), which is defined as

$$RMSE = \sqrt{\sum_{i=1}^{N} (f(\boldsymbol{x}_i) - y_i)^2 / N}, \tag{2}$$

where $f(\boldsymbol{x}_i)$ is the predicted value.

(b) (3 points) Repeat the above steps for Dataset 2. What do you observe? Explain your answers, especially how you use the SVD.

4. (6 points) Suppose $\boldsymbol{x} = [x_1, x_2, x_3]^T$ with the three attributes: $x_1 = \textit{height}$, $x_2 = \textit{weight}$ and $x_3 = \textit{age}$.

   (a) (2 points) Let Bob measure *height* in inches, *weight* in pounds and *age* in months. On the other hand, let Alice measure *height* in centimeters, *weight* in kilograms and *age* in days. Both Bob and Alice solve the least squares problem: $\min_w \|Xw - y\|_2^2$ in order to predict y, the number of pizza slices to be eaten by $\boldsymbol{x}$. How are the coefficient vectors $\boldsymbol{w}$ obtained by Bob and Alice related to each other? Use the normal equations to explain your answers.

   (b) (2 points) Suppose Bob and Alice both solve a ridge regression problem: $\min_w \|X\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{w}\|_2^2$. What is the relationship between their coefficient vectors?

   (c) (2 points) Suppose the training set $\{(\boldsymbol{x}_i, y_i)\}$ is changed to $\{(\boldsymbol{x}_i, \bar{y}_i)\}$ where $\bar{y}_i = y_i + 1$. How will the coefficient vectors change in the two cases (least squares and ridge regression)? You can take $\boldsymbol{x}_i$ to have measurements made by Bob or Alice.