MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering and Computer Science
6.036—Introduction to Machine Learning
Spring Semester 2016

Assignment 3 SOLUTIONS

Issued: Friday, March 4th, 9:00 AM

Collaborative Filtering

(a)

$$X = UV^T = \begin{bmatrix} 24 & 6 & 30 \\ 8 & 2 & 10 \\ 12 & 3 & 15 \\ 12 & 3 & 15 \\ 20 & 5 & 25 \end{bmatrix}$$

(b) Let D be the set of index of observation.

$$\begin{aligned} J_{square} &= \sum_{i,j \in D} (Y_{ij} - X_{ij})^2/2 \\ &= [(5-24)^2 + (7-30)^2 + (2-2)^2 + (1-3)^2 + (4-15)^2 \\ &+ (4-12)^2 + (3-5)^2 + (6-25)^2]/2 \\ &= 722 \end{aligned}$$

$$\begin{aligned} J_{reg} &= \frac{\lambda}{2}\|U\|^2 + \frac{\lambda}{2}\|V\|^2 \\ &= \frac{\lambda}{2}\sum_{i=1}^{n}\|U_i\|^2 + \frac{\lambda}{2}\sum_{i=1}^{n}\|V_i\|^2 \\ &= 62.5 \end{aligned}$$

(c) With V fixed as $[4, 1, 5]^T$, we can represent prediction X as:

$$X = UV^T = \begin{bmatrix} 4U_1 & 1U_1 & 5U_1 \\ 4U_2 & 1U_2 & 5U_2 \\ 4U_3 & 1U_3 & 5U_3 \\ 4U_4 & 1U_4 & 5U_4 \\ 4U_5 & 1U_5 & 5U_5 \end{bmatrix}$$

Let D be the set of index of observation, the new estimate $U^{(1)}$ should be:

$$\begin{aligned} U^{(1)} &= \arg\min_U J(U) \\ &= \arg\min_U \sum_{i,j \in D} (Y_{ij} - (UV)_{ij})^2/2 + \sum_{i=1}^{5}\frac{\lambda}{2}\|U\|^2 \\ &= \arg\min_U [(5-4U_1)^2 + (7-5U_1)^2 + (2-1U_2)^2 + (1-1U_3)^2 + (4-5U_3)^2 + \\ &(4-4U_4)^2 + (3-1U_5)^2 + (6-5U_5)^2]/2 + \sum_{i=1}^{5}\frac{\lambda}{2}\|U\|^2 \end{aligned}$$

To minimize this loss, we differentiate it with respect to U_1 and equate it to zero.

$$\begin{aligned}\frac{\partial J(U)}{\partial U_1} &= 2 \times (5 - 4U_1) \times (-4) + 2 \times (7 - 5U_1) \times (-5) + \lambda \times 2U_1 \\ &= 83 \times U_1 - 110 = 0\end{aligned}$$

Therefore $U_1^{(1)} = \frac{55}{42}$

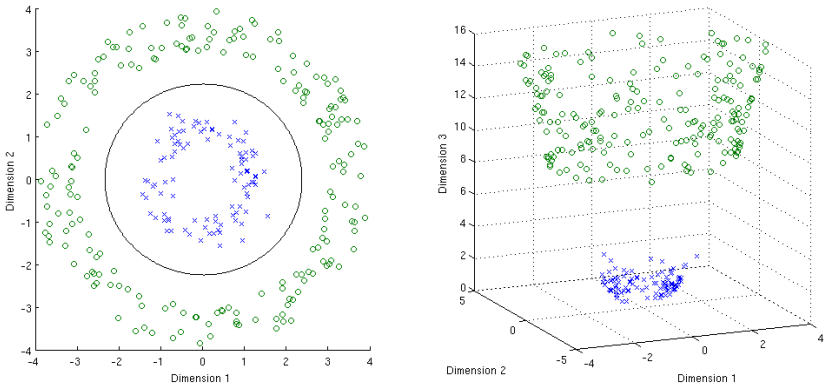
Similarly we have

$$\begin{aligned}U_2^{(1)} &= 1 \\ U_3^{(1)} &= \frac{7}{9} \\ U_4^{(1)} &= \frac{16}{17} \\ U_5^{(1)} &= \frac{11}{9}\end{aligned}$$

Kernels

(a) We can rewrite the kernel as $K(x, q) = (x^T q + 1)^2 = \left(1 + \sum_{i=1}^{2} x_i q_i\right)^2 = (x_1 q_1 + x_2 q_2 + 1)^2$. Expanding and combining terms gives $x_1^2 q_1^2 + x_2^2 q_2^2 + 2x_1 x_2 q_1 q_2 + 2x_1 q_1 + 2x_2 q_2 + 1$. We could then rewrite this expression as $\phi(x)^T \phi(q)$ where $\phi(x) = [x_1^2, x_2^2, \sqrt{2}x_1 x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1]$. The $\sqrt{2}x_1 x_2$ dimension represents the product of features in the original feature space. Notice that this term is only nonzero when both x_1 and x_2 are nonzero. If our original features were entries in a bag of words vector, this feature would capture the presence of two different words simultanously occuring in the same piece of text.

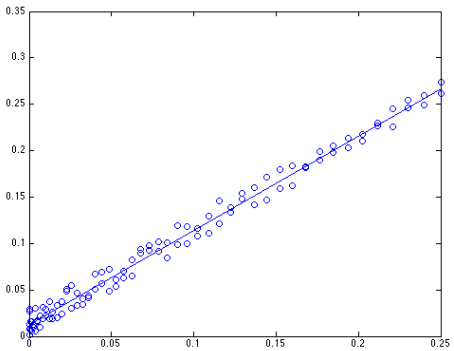
(b) With $x = [x_1; x_2]$, one mapping which could satisfy the mapping is $\phi(x)_3 = x_1^2 + x_2^2$. The decision boundary is shown below:



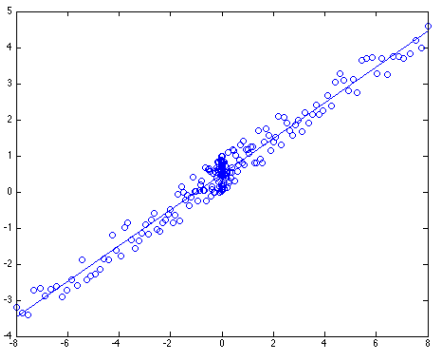
Linear Regression and Regularization

(a) In both (i.) and (ii.) the data seem to follow a non-linear pattern so a linear regression would not be a good model for the data. However, we can certainly use a non-linear transformation $\phi(X)$ to transform X such that a linear regression would be a good model for $(\phi(x^{(i)}), y^{(i)})$. In (i.), the data seem to roughly follow $Y = X^2$, so a quadratic transformation would likely produce a feature vector

ϕ that would result in $(\phi(x^{(i)}), y^{(i)})$ that would be well-fit by a linear regression. In this case, we must also note that the x-values are shifted by 0.5 (since the parabola is symmetric by centered at 0.5), thus giving us a $\phi(X) = (X - 0.5)^2$. Similarly, in (ii.), the data seem to roughly follow $Y = X^3$, so $\phi(X) = X^3$ would produce a feature vector ϕ that would result in $(\phi(x^{(i)}), y^{(i)})$ that would be well-fit by a linear regression. Below are plots of $(\phi(x^{(i)}), y^{(i)})$, along with the corresponding linear regression, for the above choices of ϕ in the cases of both (i.) and (ii.).



i.



ii.

(b) The ridge regression objective function is given by

$$L(\theta_0, \theta) = \sum_{t=1}^{n} (y^{(t)} - \theta x^{(t)} - \theta_0)^2 + \lambda \theta^2$$

i The gradient is a two-dimensional vector $\nabla L = \left[\frac{\partial L}{\partial \theta_0}, \frac{\partial L}{\partial \theta}\right]$ where

$$\frac{\partial L}{\partial \theta_0} = -2 \sum_{t=1}^{n} (y^{(t)} - \theta x^{(t)} - \theta_0)$$

$$\frac{\partial L}{\partial \theta} = 2\lambda\theta - 2 \sum_{t=1}^{n} (y^{(t)} - \theta x^{(t)} - \theta_0) x^{(t)}$$

ii To find the θ, θ_0 which minimize L, we note that because this objective function is convex, any point where $\nabla L(\theta_0, \theta) = 0$ is a global minimum. Thus, we set the gradient equal to zero and

solve for θ, θ_0 to find the minimizers:

$$\frac{\partial}{\partial \theta_0} = -2\sum_{t=1}^{n}(y^{(t)} - \theta x^{(t)} - \theta_0) = -2\sum_{t=1}^{n}(y^{(t)} - \theta x^{(t)}) + 2\sum_{t=1}^{n}\theta_0 = 0$$

$$\implies -2n\theta_0 = -2\sum_{t=1}^{n}(y^{(t)} - \theta x^{(t)}) \implies \theta_0 = \frac{1}{n}\sum_{t=1}^{n}(y^{(t)} - \theta x^{(t)})$$

$$\frac{\partial}{\partial \theta} = 2\lambda\theta - 2\sum_{t=1}^{n}(y^{(t)} - \theta x^{(t)} - \theta_0)x^{(t)}$$

$$= 2\lambda\theta - 2\sum_{t=1}^{n}\left(y^{(t)} - \theta x^{(t)} - \left[\frac{1}{n}\sum_{s=1}^{n}(y^{(s)} - \theta x^{(s)})\right]\right) \cdot x^{(t)} = 0$$

$$\implies \lambda\theta - \sum_{t=1}^{n} x^{(t)}y^{(t)} + \theta\sum_{t=1}^{n} {x^{(t)}}^2 + \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}(y^{(s)} - \theta x^{(s)})x^{(t)} = 0$$

$$\implies \lambda\theta - \sum_{t=1}^{n} x^{(t)}y^{(t)} + \theta\sum_{t=1}^{n} {x^{(t)}}^2 + \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}(y^{(s)} - \theta x^{(s)})x^{(t)} = 0$$

$$\implies \lambda\theta - \sum_{t=1}^{n} x^{(t)}y^{(t)} + \theta\sum_{t=1}^{n} {x^{(t)}}^2 + \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n} y^{(s)}x^{(t)} - \frac{1}{n}\theta\sum_{t=1}^{n}\sum_{s=1}^{n} x^{(s)}x^{(t)} = 0$$

$$\implies \widehat{\theta} = \frac{\sum_{t=1}^{n} x^{(t)}y^{(t)} - \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n} y^{(s)}x^{(t)}}{\lambda + \sum_{t=1}^{n} {x^{(t)}}^2 - \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n} x^{(s)}x^{(t)}}$$ is the value of θ which minimizes $L(\theta_0, \theta)$.

Note that if we define $\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x^{(t)}$, then we can also rewrite the above expression in a nicer form:

$$\widehat{\theta} = \frac{\sum_{t=1}^{n}(x^{(t)} - \bar{x})y^{(t)}}{\lambda + \sum_{t=1}^{n} x^{(t)}(x^{(t)} - \bar{x})}$$

Finally, we can plug this value of $\widehat{\theta}$ back into expression $\widehat{\theta}_0 = \frac{1}{n}\sum_{t=1}^{n}(y^{(t)} - \theta x^{(t)})$ to find the corresponding $\widehat{\theta}_0$ which together with $\widehat{\theta}$ minimizes L.

(c) Large λ penalty on θ results in flatter lines. Therefore, (b) and (d) correspond to (2) and (4). Large λ penalty on the offset θ_0 produces lines with reduced y-intercept. Since the offset for (d) must be smaller, (d) matches with (4); (b) matches with (2). (a) and (c) corresponds to (1) and (3). Since the offset of (a) must be smaller, (a) matches with (3), whereas (c) matches with (1).

OPTIONAL: More About Linear Regression

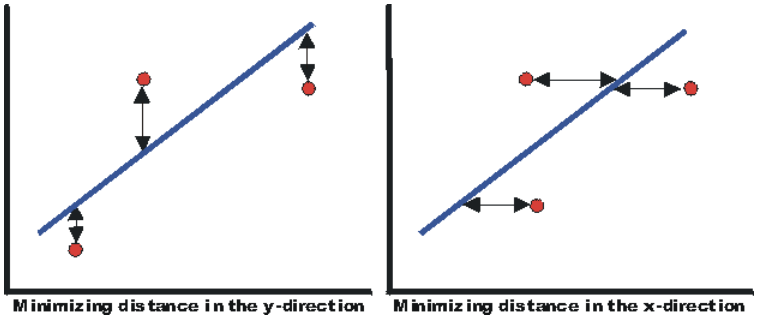
(a) This is not always true. When considering the least squares regression of Y on X given the data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$, we are minimizing the difference between the observed values Y and the fitted values provided by the model $\hat{Y} = ax + b$. That is, we are finding the values of a and b that minimize the error introduced by our model for predicting Y:

$$\min_{a,b}\sum_{i=1}^{n}(ax^{(i)} + b - y^{(i)})^2$$

However, when considering the regression of X on Y given the data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$, we are minimizing the difference between the observed values X and the fitted values provided by the model $\hat{X} = cy + d$, or the error introduced by our model for predicting X:

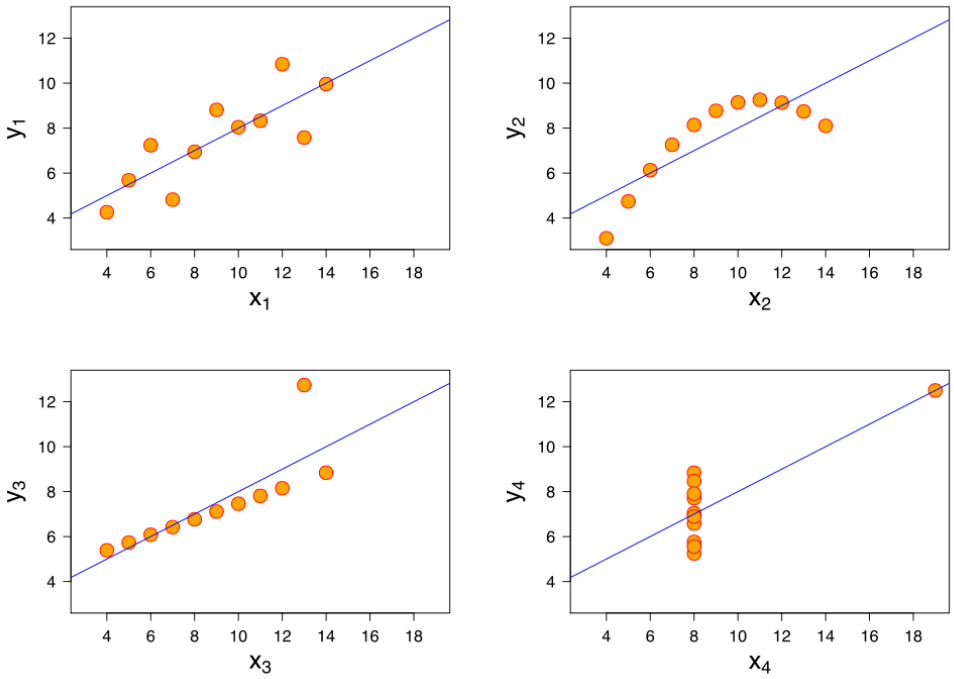
$$\min_{c,d} \sum_{i=1}^{n} (cy^{(i)} + d - x^{(i)})^2$$

Note that in each case we are either minimizing our error for predicting Y or our error for predicting X. In other words, if we are regressing Y on X, then the regression line is fit such that the vertical deviations from the points to the line are minimized. On the other hand, if we are regressing X on Y, then we are minimizing the horizontal deviations from the points to the regression line. For an illustration, see the figure below.



(b) Yes, this is certainly possible. For example, when fitting a linear regression using the least squares approach, the regression line simply minimizes the sum of the squares of the residuals[1], so it is possible to create two distinct sets of points whose linear regression function is the same by ensuring that the sum of the squares of the residuals in each case is minimized by the same line. A famous example of this is Anscombe's quartet (pictured below), which is an example of four such sets of data points that vary considerably but that have nearly identical statistical properties and the same regression line.

[1] A residual is the difference between an observed value and the fitted value provided by the model


y1
x1
y2
x2
y3
x3
y4
x4
4
6
8
10
12
14
16
18