Sharif University of Technology
Electrical Engineering Department

Machine Learning
HW 2

Amir Hossein Yari
99102507

June 8, 2023

Contents

1. Lasso Regression

1. Whitening the dataset causes the features to be independent

$$
\begin{aligned}
J_\lambda &= \frac{1}{2}||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 + \lambda||\mathbf{w}||_1 \\
&= \frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda(|\mathbf{w}_1| + |\mathbf{w}_2| + \cdots + |\mathbf{w}_d|) \\
&= \frac{1}{2}\mathbf{y}^T\mathbf{y} - \frac{1}{2}\mathbf{y}^T\mathbf{X}\mathbf{w} - \frac{1}{2}\mathbf{w}^T\mathbf{X}^T\mathbf{y} + \frac{1}{2}\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda(|\mathbf{w}_1| + |\mathbf{w}_2| + \cdots + |\mathbf{w}_d|) \\
&= \frac{1}{2}\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} + \frac{1}{2}\mathbf{w}^T\mathbf{w} + \lambda(|\mathbf{w}_1| + |\mathbf{w}_2| + \cdots + |\mathbf{w}_d|) \\
&= \frac{1}{2}||\mathbf{y}||^2 - \mathbf{y}^T\sum_{i=1}^{d} w_i X_i + \frac{1}{2}\sum_{i=1}^{d} w_i^2 + \lambda(\sum_{i=1}^{d}|w_i|) \\
&= \frac{1}{2}||\mathbf{y}||^2 + \sum_{i=1}^{d}\frac{1}{2}w_i^2 + \lambda|w_i| - \mathbf{y}^T w_i X_i \\
&\Rightarrow g(y) = \frac{1}{2}||\mathbf{y}||^2 \quad , \quad f(X_i, \mathbf{y}, w_i, \lambda) = -\mathbf{y}^T w_i X_i + \frac{1}{2}w_i^2 + \lambda|w_i|
\end{aligned}
$$

2. If $w_i \geqslant 0$

$$
\begin{aligned}
&\Rightarrow J_\lambda = \frac{1}{2}||\mathbf{y}||^2 + \sum_{i=1}^{d}\frac{1}{2}w_i^2 + \lambda w_i - \mathbf{y}^T w_i X_i \\
&w_i^* = \underset{w_i}{\operatorname{argmin}}\ J_\lambda(w_i) = \underset{w_i}{\operatorname{argmin}}\ \frac{1}{2}||\mathbf{y}||^2 - \mathbf{y}^T w_i X_i + \frac{1}{2}w_i^2 + \lambda w_i \\
&\frac{\partial J_\lambda(w_i)}{\partial w_i} = -\mathbf{y}^T X_i + w_i + \lambda = 0 \Rightarrow w_i = \mathbf{y}^T X_i - \lambda
\end{aligned}
$$

3. If $w_i < 0$

$$\Rightarrow J_\lambda = \frac{1}{2}||\mathbf{y}||^2 + \sum_{i=1}^{d} \frac{1}{2} w_i^2 - \lambda w_i - \mathbf{y}^T w_i X_i$$

$$w_i^* = \underset{w_i}{\text{argmin}}\ J_\lambda(w_i) = \underset{w_i}{\text{argmin}}\ \frac{1}{2}||\mathbf{y}||^2 - \mathbf{y}^T w_i X_i + \frac{1}{2} w_i^2 - \lambda w_i$$

$$\frac{\partial J_\lambda(w_i)}{\partial w_i} = -\mathbf{y}^T X_i + w_i - \lambda = 0 \Rightarrow w_i = \mathbf{y}^T X_i + \lambda$$

4. w_i would equal to zero?

To find the conditions under which w_i would equal to zero in Lasso regression, we can substitute w_i into the equation:

$$w_i = \mathbf{y}^T X_i - \lambda \text{sign}(w_i)$$
$$w_i + \lambda \text{sign}(w_i) = \mathbf{y}^T X_i$$

Now, we consider two cases:
If $w_i > 0 \Rightarrow$
In this case, we have $\text{sign}(w_i) = 1$, so we can simplify the equation as:

$$w_i + \lambda = \mathbf{y}^T X_i$$

If $w_i = 0$, then we have:

$$\lambda = \mathbf{y}^T X_i$$

If $w_i < 0 \Rightarrow$
In this case, we have $\text{sign}(w_i) = -1$, so we can simplify the equation as:

$$w_i - \lambda = \mathbf{y}^T X_i$$

If $w_i = 0$, then we have:

$$\lambda = -\mathbf{y}^T X_i$$

In general:

$$\lambda = sign(w_i)\mathbf{y}^T X_i$$

If the above relationship is established, the condition will be applied.

5. Ridge regression vs Lasso Regression

In Ridge regression, the regularization term in the loss function is given by $\frac{1}{2}\lambda||w||_2^2$. The effect of this term is to shrink the weights towards zero, but it does not lead to exact sparsity as in Lasso regression.
To determine when a weight w_i in Ridge regression will be zero, we can set the derivative of the loss function with respect to w_i to zero and solve for w_i.

$$\Rightarrow J_\lambda = \frac{1}{2}||\mathbf{y}||^2 + \sum_{i=1}^{d} \frac{1}{2}w_i^2 + \frac{1}{2}\lambda w_i^2 - \mathbf{y}^T w_i X_i$$

$$\frac{\partial J_\lambda(w_i)}{\partial w_i} = -\mathbf{y}^T X_i + w_i + \lambda w_i = 0 \Rightarrow w_i = \frac{\mathbf{y}^T X_i}{1+\lambda}$$

From this equation, we see that w_i will only be zero if $\mathbf{y}^T X_i = 0$, which means that the corresponding feature i is uncorrelated with the response variable $\mathbf{y}$. In other words, if a feature is not correlated with the response variable, its corresponding weight will be zero in Ridge regression.
The main difference between Ridge regression and Lasso regression is that Lasso regression has a sparsity-inducing regularization term (L1 norm) while Ridge regression has a smoothness-inducing regularization term (L2 norm). This leads to a difference in the behavior of the two methods in terms of feature selection. In Lasso regression, the sparsity-inducing term can lead to exact sparsity, meaning that some weights are exactly equal to zero, resulting in feature selection. In contrast, Ridge regression does not typically lead to exact sparsity, but instead shrinks the weights towards zero, resulting in less severe feature selection.

2. Bayesian Analysis of Exponential Distribution

1. MLE of Exponential Distribution

$$L(\theta|X) = \prod_{i=1}^{N} \theta e^{-\theta X_i}$$

$$logL(\theta|X) = log(\prod_{i=1}^{N} \theta e^{-\theta X_i}) = \sum_{i=1}^{N} log(\theta) - \theta X_i$$

$$\frac{\partial\ logL(\lambda|X)}{\partial\,\theta} = 0 \Rightarrow \frac{N}{\theta} - \sum_{i=1}^{N} X_i = 0 \Rightarrow \hat{\theta} = \frac{N}{\sum_{i=1}^{N} X_i} \quad \checkmark$$

2. MLE of Three Data

According to previous part we have:

$$\hat{\theta} = \frac{N}{\sum_{i=1}^{N} X_i}$$

So:

$$\hat{\theta} = \frac{3}{5+6+4} = \frac{3}{15} = 0.2$$

3. Choose Prior Parameter

$$E[\Theta] = \int_0^\infty \theta\ \lambda e^{-\lambda\theta} d\theta = [-\theta e^{-\lambda\theta}]|_0^\infty + \int_0^\infty e^{-\lambda\theta} d\theta = [0-0] + [-\frac{1}{\lambda} e^{-\lambda\theta}]|_0^\infty$$

$$= 0 + [0 - (-\frac{1}{\lambda})] = \frac{1}{\lambda}$$

So $\lambda = 3$ for $E[\Theta] = \frac{1}{3}$.

4. Posterior Distribution

$prior: \;\; p_\Theta(\theta) = \lambda \; e^{-\lambda\theta}$

$likelihood \; function: \;\; p_{D|\Theta}(D|\theta) = \theta^N e^{-\theta \sum_{i=1}^N X_i}$

$marginal \; likelihood: \;\; p(D) = \int p_{D|\Theta}(D|\theta) \; p_\Theta(\theta) \; d\theta = \int \theta^N e^{-\theta \sum_{i=1}^N X_i} \; \lambda \; e^{-\lambda\theta} \; d\theta$

$= \lambda \int \theta^N e^{-\theta(\lambda + \sum_{i=1}^N X_i)} \; d\theta$

This is the kernel of a gamma distribution with shape parameter $N+1$ and rate parameter $\lambda + \sum_{i=1}^N X_i$. Therefore, we have:

$$p_{\Theta|D}(\theta|D;\lambda) = Gamma(\theta|N+1, \lambda + \sum_{i=1}^N X_i)$$

5. Find Expected Value of Posterior Distribution

We know that expected value of gamma distribution with shape parameter α and rate parameter β is $\frac{\alpha}{\beta}$. So:

$$E[\Theta|D,\lambda] = \frac{N+1}{\lambda + \sum_{i=1}^N X_i}$$

3. Naive Bayes

1. Find $p_{Y|X_1,X_2}(y|x_1 = 0, x_2 = 0)$

$$p_{Y|X_1,X_2}(y|x_1 = 0, x_2 = 0) = \frac{p_{X_1,X_2|Y}(x_1 = 0, x_2 = 0|y)p_Y(y)}{p_{X_1,X_2}(x_1 = 0, x_2 = 0)}$$

$$p_{X_1,X_2|Y}(x_1 = 0, x_2 = 0|y) = p_{X_1|Y}(x_1 = 0|y)p_{X_2|Y}(x_2 = 0|y)$$

$$= \theta_c^{x_1}(1 - \theta_c^{(1-x_1)}) \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(x_2-\mu_c)^2}{2\sigma_c^2}} = [0.121, 0.199, 0.121]$$

$$p_Y(y) = \boldsymbol{\pi} = [0.5, 0.25, 0.25]$$

$$p_{X_1,X_2}(x_1 = 0, x_2 = 0) = p_{X_1}(x_1 = 0)p_{X_2}(x_2 = 0) = 0.5 \times 0.121 + 0.25 \times 0.199$$

$$+ 0.25 \times 0.121 = 0.1405$$

$$\Rightarrow p_{Y|X_1,X_2}(y|x_1 = 0, x_2 = 0) = [\frac{0.5 \times 0.121}{0.1405}, \frac{0.25 \times 0.199}{0.1405}, \frac{0.25 \times 0.121}{0.1405}]$$

$$= [0.431, 0.354, 0.215]$$

2. Find $p_{Y|X_1}(y|x_1 = 0)$

$$p_{Y|X_1}(y|x_1 = 0) = \frac{p_{X_1|Y}(x_1 = 0|y)p_Y(y)}{p_{X_1}(x_1 = 0)}$$

$$p_{X_1|Y}(x_1 = 0|y) = \theta_c^{x_1}(1 - \theta_c^{(1-x_1)}) = [0.5, 0.5, 0.5]$$

$$p_Y(y) = \boldsymbol{\pi} = [0.5, 0.25, 0.25]$$

$$p_{X_1}(x_1 = 0) = 0.5 \times 0.5 + 0.25 \times 0.5 + 0.25 \times 0.5 = 0.5$$

$$\Rightarrow p_{Y|X_1}(y|x_1 = 0) = [\frac{0.5 \times 0.5}{0.5}, \frac{0.25 \times 0.5}{0.5}, \frac{0.25 \times 0.5}{0.5}] = [0.5, 0.25, 0.25]$$

3. Find $p_{Y|X_2}(y|x_2 = 0)$

$$p_{Y|X_2}(y|x_2 = 0) = \frac{p_{X_2|Y}(x_2 = 0|y)p_Y(y)}{p_{X_2}(x_2 = 0)}$$

$$p_{X_2|Y}(x_2 = 0|y) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(x_2-\mu_c)^2}{2\sigma_c^2}} = [0.242, 0.398, 0.242]$$

$$p_Y(y) = \boldsymbol{\pi} = [0.5, 0.25, 0.25]$$

$$p_{X_2}(x_2 = 0) = 0.5 \times 0.242 + 0.25 \times 0.398 + 0.25 \times 0.242 = 0.281$$

$$\Rightarrow p_{Y|X_2}(y|x_2 = 0) = [\frac{0.5 \times 0.242}{0.281}, \frac{0.25 \times 0.398}{0.281}, \frac{0.25 \times 0.242}{0.281}] = [0.431, 0.354, 0.215]$$

4. Justify the pattern

According to the obtained results, we have:

$$p_{Y|X_1,X_2}(y|x_1 = 0, x_2 = 0) = p_{Y|X_2}(y|x_2 = 0)$$

Due to the equality of $p_{X_1|Y}(x_1 = 0|y)$ for all classes, X_1 was ineffective in this problem.
In other word, the equality of likelihood for each class it will not affect on the posterior distribution.

4. Desicion Boundary

1. Find the decision boundary with $\sigma_2^2 = 10^6$

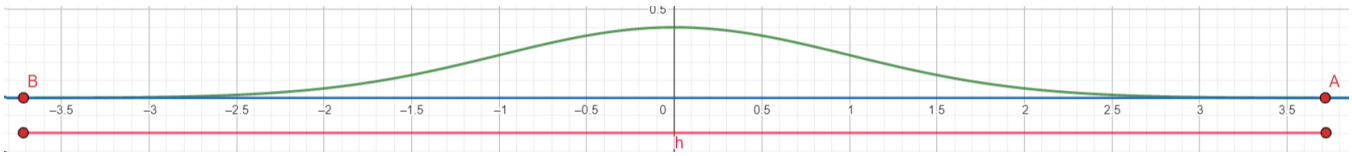
$$p_X(x|Y=1) = p_X(x|Y=2) \Rightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \frac{1}{1000\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2\times 10^6}}$$

$$(2\pi)^{-\frac{1}{2}} e^{-\frac{x^2}{2}} = \frac{(2\pi)^{-\frac{1}{2}}}{1000} e^{-\frac{(x-1)^2\times 10^{-6}}{2}} \Rightarrow (2\pi)^{-\frac{1}{2}} e^{-\frac{x^2}{2}} \times 1000 = \frac{(2\pi)^{-\frac{1}{2}}}{1000} e^{-\frac{(x-1)^2\times 10^{-6}}{2}} \times 1000$$

$$\Rightarrow \frac{500 \times 2^{\frac{1}{2}}}{\pi^{\frac{1}{2}}} e^{-\frac{x^2}{2}} = \frac{1}{2^{\frac{1}{2}}\pi^{\frac{1}{2}}} e^{-\frac{(x-1)^2\times 10^{-6}}{2}} \Rightarrow \frac{500 \times 2^{\frac{1}{2}}}{\pi^{\frac{1}{2}}} e^{-\frac{x^2}{2}} = \frac{1}{2^{\frac{1}{2}}\pi^{\frac{1}{2}}} e^{-\frac{(x-1)^2\times 10^{-6}}{2}}$$

$$\Rightarrow x = \pm 3.71692$$

green curve $= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, blue curve $= \frac{1}{1000\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2\times 10^6}}$, A,B = intersection point



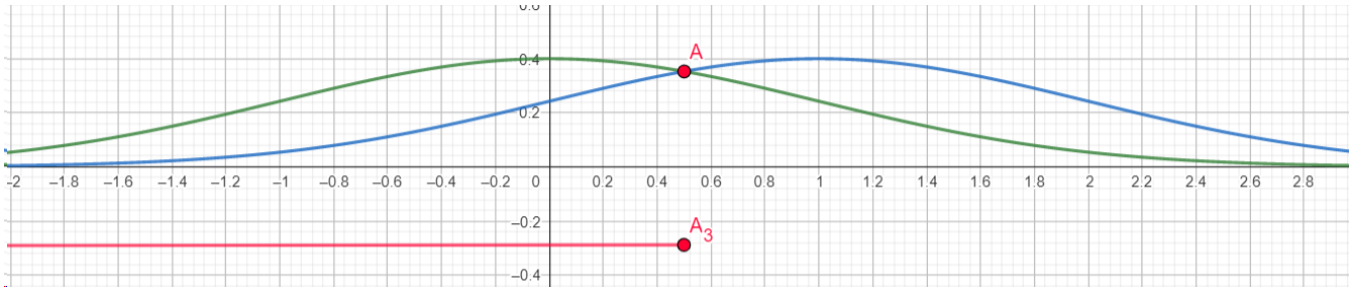
The red line is the x's that satisfy the condition.

2. Find the decision boundary with $\sigma_2^2 = 1$

$$p_X(x|Y=1) = p_X(x|Y=2) \Rightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}}$$

$$\ln\left((2\pi)^{-\frac{1}{2}}\right) - \frac{x^2}{2} = \ln\left((2\pi)^{-\frac{1}{2}}\right) - \frac{(x-1)^2}{2} \Rightarrow x = \frac{1}{2}$$

A = intersection point



The red line is the x's that satisfy the condition.

5. Newton's Method as Solver For Linear Regression Problem

1. Find the Hessian

$$J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{\theta}^T\mathbf{x}^{(i)} - y^{(i)})^2$$

$$\frac{\partial}{\partial\theta_j}\left(\frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{\theta}^T\mathbf{x}^{(i)} - y^{(i)})^2\right) = \sum_{i=1}^{m}(\boldsymbol{\theta}^T\mathbf{x}^{(i)} - y^{(i)})x_j^{(i)}$$

where $x_j^{(i)}$ denotes the j-th element of the input vector $\mathbf{x}^{(i)}$. We can write this in matrix form as:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbf{X}^T(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})$$

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\left(\frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{\theta}^T\mathbf{x}^{(i)} - y^{(i)})^2\right) = \sum_{k=1}^{m} x_i^{(k)}x_j^{(k)}$$

We can write this in matrix form as:

$$H = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}J(\boldsymbol{\theta}) = \mathbf{X}^T\mathbf{X}$$

2. Newton's method

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - [H(\boldsymbol{\theta}^{(k)})]^{-1}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^{(k)})$$

Let's apply this update rule for the first iteration, where we initialize $\boldsymbol{\theta}^{(0)}$ to zero. Now, let's compute the update for the first iteration:

$$\begin{aligned}\boldsymbol{\theta}^{(1)} &= \boldsymbol{\theta}^{(0)} - [H(\boldsymbol{\theta}^{(0)})]^{-1}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^{(0)}) \\ &= \mathbf{0} - [\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{0} - \mathbf{y}) \\ &= [\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{y}\end{aligned}$$

This update rule gives us the solution to the least squares problem:

$$\boldsymbol{\theta}^* = [\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{y}$$

Therefore, the first iteration of Newton's method gives us the solution to our least squares problem.

6. Multivariate Least Squares

1. Matrix-Vector Notation

$$\mathbf{x}^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix} \quad X = \begin{bmatrix} \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \cdots & \mathbf{x}^{(m)} \end{bmatrix}^T \quad \mathbf{y}^{(i)} = \begin{bmatrix} y_1^{(i)} \\ y_2^{(i)} \\ \vdots \\ y_p^{(i)} \end{bmatrix} \quad Y = \begin{bmatrix} \mathbf{y}^{(1)} & \mathbf{y}^{(2)} & \cdots & \mathbf{y}^{(m)} \end{bmatrix}^T$$

$$\Theta = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,p} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{n,1} & \theta_{n,2} & \cdots & \theta_{n,p} \end{bmatrix}$$

$$J(\Theta) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{p}((\Theta^T\mathbf{x}^{(i)})_j - (\mathbf{y}^{(i)})_j)^2 = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{p}(\Theta^T\mathbf{x}^{(i)} - \mathbf{y}^{(i)})_j^2$$

let $\mathbf{e}^{(i)} = \Theta^T\mathbf{x}^{(i)} - \mathbf{y}^{(i)}$ be the corresponding error vector. First summation actually calculates the norm of error vector. Then, we can express the cost function as:

$$J(\Theta) = \frac{1}{2}\sum_{i=1}^{m}||\mathbf{e}^{(i)}||^2 = \frac{1}{2}\sum_{i=1}^{m}(\mathbf{e}^{(i)})^T\mathbf{e}^{(i)} = \frac{1}{2}\sum_{i=1}^{m}(\Theta^T\mathbf{x}^{(i)} - \mathbf{y}^{(i)})^T(\Theta^T\mathbf{x}^{(i)} - \mathbf{y}^{(i)})$$

$$\Rightarrow J(\Theta) = \frac{1}{2}\sum_{i=1}^{m}(\Theta^T\mathbf{x}^{(i)})^T(\Theta^T\mathbf{x}^{(i)}) - (\Theta^T\mathbf{x}^{(i)})^T\mathbf{y}^{(i)} - \mathbf{y}^{(i)T}(\Theta^T\mathbf{x}^{(i)}) + \mathbf{y}^{(i)T}\mathbf{y}^{(i)}$$

$$J(\Theta) = \frac{1}{2}\left[\mathrm{Tr}\left(X\Theta\Theta^TX^T\right) - 2\mathrm{Tr}\left(X\Theta Y^T\right) + \mathrm{Tr}\left(YY^T\right)\right]$$

2. closed form solution for Θ^*

$$J(\Theta) = \frac{1}{2}\left[\mathrm{Tr}\left(X\Theta\Theta^TX^T\right) - 2\mathrm{Tr}\left(X\Theta Y^T\right) + \mathrm{Tr}\left(YY^T\right)\right]$$

$$\frac{\partial J(\Theta)}{\partial\Theta} = 0 \Rightarrow \Theta^TX^TX - Y^TX = 0 \Rightarrow \Theta^TX^TX = Y^TX \Rightarrow X^TX\Theta = XY^T$$

$$\Theta^* = (X^TX)^{-1}X^TY$$

3. compare the multivariate solution with independent least squares

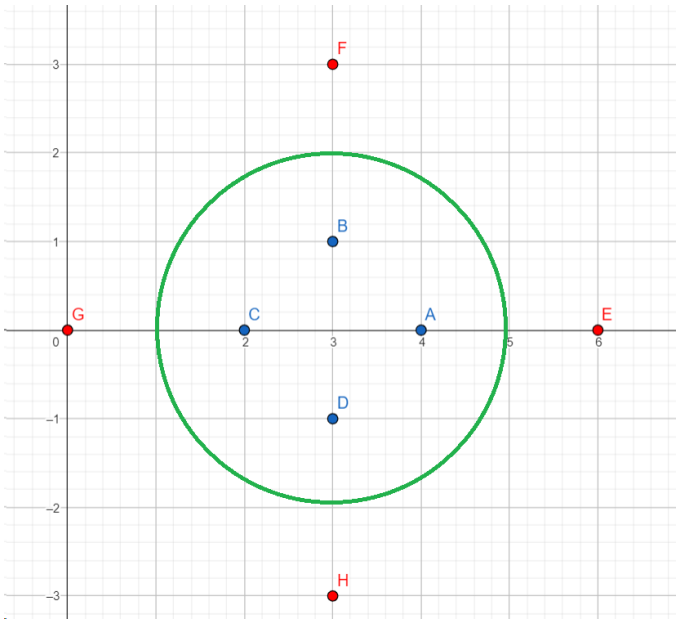
If we solve the p independent linear models $\mathbf{y}_j^{(i)} = \boldsymbol{\theta}_j^T \mathbf{x}^{(i)}$ separately, we will obtain p different parameter vectors $\boldsymbol{\theta}_j$ for each variable. Each parameter vector will minimize the squared error for the corresponding variable, but they may not necessarily minimize the total squared error across all variables.

In other words, the p independent least squares problems give a solution that is optimal for each variable in isolation, but it may not be optimal for all variables taken together. This is because there may be correlations or interactions between the variables that are not taken into account when solving each problem independently.

On the other hand, the multivariate least squares problem takes into account all variables simultaneously and finds the parameter vector $\boldsymbol{\theta}$ that minimizes the total squared error across all variables. This means that the multivariate solution is likely to be more accurate and robust than the p independent solutions.

In summary, the parameters from the p independent least squares problems may not be as accurate as the multivariate solution because they only optimize the error for each variable in isolation and do not take into account the correlations or interactions between variables. The multivariate solution, on the other hand, considers all variables simultaneously and finds the optimal parameter vector that minimizes the total squared error.

7. Choosing A Proper Mapping



proper mapping is $\Phi(x_1, x_2) = [1, (x_1 - 3)^2, x_2^2]$