

CSCI-567: Assignment #3

Due on Friday, October 16, 2015

Saket Choudhary
2170058637

Contents

| | |
|--------------------------|-----------|
| Problem 1 | 3 |
| Problem 1: (a) | 3 |
| Problem 1: (b) | 4 |
| Problem 1: (c) | 4 |
| Problem 1: (d) | 4 |
| Problem 2 | 4 |
| Problem 2: (a) | 5 |
| Problem 2: (b) | 5 |
| Problem 2: (c) | 6 |
| Problem 3 | 7 |
| Problem 3: (a) | 7 |
| Problem 3: (b) | 7 |
| Problem 3: (c) | 8 |
| Problem 3: (d) | 8 |
| Problem 4 | 8 |
| Problem 4: (a) | 9 |
| Problem 4: (b) | 9 |
| Problem 4: (c) | 10 |
| Problem 4: (d) | 10 |
| Problem 4: (e) | 10 |
| Problem 5 | 11 |
| a | 11 |
| | 11 |

Problem 1

Problem 1: (a)

Let $\sigma(a) = \frac{1}{1+e^{-a}}$ and

$$P(Y = 1|X = x) = \sigma(b + w^T x) P(Y = 0|X = x) = 1 - \sigma(b + w^T x)$$

Observe that $Y = 1$ when $b + w^T x \geq 0$ and $Y = 0$ when $b + w^T x < 0$

Thus,

$$\begin{aligned} P(Y = y|X = x) &= \sigma(b + w^T x)^y (1 - \sigma(b + w^T x))^{(1-y)} \\ \log(P(Y = y|X = x)) &= y \log(\sigma(b + w^T x)) + (1 - y) \log(1 - \sigma(b + w^T x)) \\ &= y \log\left(\frac{\sigma(b + w^T x)}{1 - \sigma(b + w^T x)}\right) + \log(1 - \sigma(b + w^T x)) \\ &= y(b + w^T x) + \log\left(\frac{e^{-(b + w^T x)}}{1 + e^{-(b + w^T x)}}\right) \\ &= y(b + w^T x) + \log\left(\frac{1}{1 + e^{(b + w^T x)}}\right) \\ &= y(b + w^T x) - \log(1 + e^{(b + w^T x)}) \end{aligned} \tag{[1.1]}$$

$$\begin{aligned} \mathcal{L}(w) &= -\log\left(\prod_{i=1}^n P(Y = y_i|X = x_i)\right) \\ &= -\sum_{i=1}^n \log(P(Y = y_i|X = x_i)) \\ &= -\sum_{i=1}^n (y_i(b + w^T x_i) - \log(1 + e^{(b + w^T x_i)})) \end{aligned}$$

Consider $(L)(w) = y(b + w^T x) - \log(1 + e^{(b + w^T x)})$

$$\begin{aligned} \frac{\partial \mathcal{L}(w)}{\partial w} &= -(xy^T) + \frac{e^{(b + w^T x)} x}{1 + e^{(b + w^T x)}} \\ \frac{\partial^2 \mathcal{L}(w)}{\partial w^2} &= 0 + \frac{\partial}{\partial w} \left(x - \frac{x}{1 + e^{(b + w^T x)}} \right) \\ \frac{\partial^2 \mathcal{L}(w)}{\partial w^2} &= \frac{x(e^{(b + w^T x)})x^T}{(1 + e^{(b + w^T x)})^2} \geq 0 \quad \forall x \in \mathbf{R} \\ \frac{\partial^2 \mathcal{L}(w)}{\partial w^2} &= x^T \sigma(b + w^T x)(1 - \sigma(b + w^T x))x \geq 0 \end{aligned} \tag{1.2}$$

From (1.2) $\frac{\partial^2 \mathcal{L}(w)}{\partial w^2} \geq 0$ and hence, from the definition of convex functions, $\mathcal{L}(w)$ is indeed a convex function.

Problem 1: (b)

When the data is perfectly linearly separable, (assume first $n/2$ of the n training points belong to class 0 and the remaining to class 1), thus our regression model should assign the first $n/2$ points to class with cent percent certainty or with probability 1 and the remaining $n/2$ to class 0 with probability 1. For this to happen, $P(Y = 1|X = X_1) = 1$ and $P(Y = 0|X = X_0) = 1$ where X_1 is the set of points belonging to class 1 and X_0 is the set of points belonging to class 0.

Clearly this scenario is possible when $\|w\| \rightarrow \infty$

Problem 1: (c)

A simple example with two points would be $(0, 0)$, $(1, 1)$. Intuitively the step function's step branches (the horizontals of a sigmoid function) will be located at infinity. Also the line separating the points $(0, 0)$ and $(1, 1)$ can be anywhere in between 0 and 1, thus there will be multiple solutions.

Problem 1: (d)

$$\begin{aligned}\mathcal{L}(w) &= \sum_{j=1}^n (-y_j(b + w^T x_j) + \log(1 + e^{(b + w^T x_j)})) + \lambda \|w\|_2^2 \\ \frac{\partial(\mathcal{L})(w)}{\partial w_i} &= \sum_{j=1}^n (-y_j(x_{ji}) + \frac{x_{ji}e^{(b + w^T x_j)x_{ij}}}{1 + e^{(b + w^T x_j)}}) + 2\lambda w_i = 0 \\ \frac{\partial^2(\mathcal{L})(w)}{\partial w_i^2} &= \sum_{j=1}^n (\frac{x_{ji}^2 e^{(b + w^T x_j)x_{ij}}}{(1 + e^{(b + w^T x_j)})^2}) + 2\lambda > 0\end{aligned}$$

where the last inequality holds since $\lambda > 0$ Consider $f(w_i) = \sum_{j=1}^n (-y_j(x_{ji}) + \frac{x_{ji}e^{(b + w^T x_j)x_{ij}}}{1 + e^{(b + w^T x_j)}}) + 2\lambda w_i = 0$

And u, v are the two solutions of $f(w_i) = 0$, i.e. $f(u) = f(v) = 0$ (Without loss of generality, assume $u < v$)

By Rolle's theorem, If $f(u) = f(v) = 0$ then there exists a point in $[u, v]$ say c such that $f'(c) = 0$ for $c \in [u, v]$

But, $f'(w_i) = \sum_{j=1}^n (\frac{x_{ji}^2 e^{(b + w^T x_j)x_{ij}}}{(1 + e^{(b + w^T x_j)})^2}) + 2\lambda > 0$ and hence there exists no such c .

and hence the function is convex, thus the solution to the partial differential $\frac{\partial(\mathcal{L})(w)}{\partial w_i}$ is unique.

Problem 2

Problem 2

Problem 2: (a)

Consider $\|w\|_0 = \#i : w_i \neq 0$ for a 1D case. Where, $x_1 = (0)$ and $x_2 = (\epsilon)$ where $0 < \epsilon < 1$

$$f(w) = \sum_i I\{w_i \neq 0\}$$

Since we are in 1D:

$$f(w) = \begin{cases} 0 & \text{if } w=0 \\ 1 & \text{otherwise} \end{cases}$$

Thus,

$$f(0) = 0$$

$$f(\epsilon) = 1$$

$$f(\lambda \times 0 + (1 - \lambda) \times \epsilon) = 1 \forall 0 < \lambda < 1 \quad (2a.1)$$

$$\lambda f(0) + (1 - \lambda)f(\epsilon) = 1 - \lambda 0 < 1 - \lambda < 1 \quad (2a.2)$$

From (2a.1), (2a.2) we see:

$$f(\lambda \times 0 + (1 - \lambda) \times \epsilon) > \lambda f(0) + (1 - \lambda)f(\epsilon)$$

Thus, $\|w\|_0$ is not a convex function!

Problem 2: (b)

$$\|w\|_1 = \sum_i |w_i|$$

Consider two vectors u, v (same dimension say in \mathbf{R}^D)

Assume: $0 < \lambda < 1$

$$\begin{aligned} \|\lambda u + (1 - \lambda)v\|_1 &= \sum_{i=1}^D |\lambda u_i + (1 - \lambda)v_i| \\ &\leq \sum_{i=1}^D (|\lambda u_i| + |(1 - \lambda)v_i|) \quad (\text{since } |a + b| \leq |a| + |b| \forall a, b \in \mathbf{R}) \\ &= \sum_{i=1}^D |\lambda| |u_i| + \sum_{i=1}^D |1 - \lambda| |v_i| \\ &= \lambda \|u\|_1 + (1 - \lambda) \|v\|_1 \quad \text{since } (0 < \lambda < 1) \end{aligned} \quad (2a.1)$$

From (2b.1), we see. $\|\lambda u + (1 - \lambda)v\|_1 \leq \lambda \|u\|_1 + (1 - \lambda) \|v\|_1$

And hence, $\|w\|_1$ is a convex function.

Problem 2: (c)

Let's redefine (for the sake of easense) x_i to be column vector i.e x_i is $D \times 1$ w is $1 \times D$ and $Y = (y_1 \dots y_n)$ the equivalent problem then becomes:

$$\begin{aligned} & \min_w \sum_i (y_i - x_i^T w)^2 \\ & \min_w \sum_i (y_i - x_i^T w)^2 + \lambda \|w\|_1 \\ & \min_w (y - X^T w)^T (y - X^T w) + \lambda \|w\|_1 \\ & \min_w (w^T X X^T w - 2Y^T X w + Y^T Y) + \min_w \lambda \|w\|_1 \\ & \min_w (w^T X X^T w - 2Y^T X w) + \min_w \lambda \|w\|_1 \end{aligned}$$

We introduce dummy variables t_i such that:

$$\|w_i\| \leq t_i \implies t_i \geq w_i \text{ and } t_i \geq -w_i$$

Now,

$$\min_w \lambda \|w\|_1 \leq \lambda (t_1 + t_2 + \dots + t_n)$$

Constraint:

$$\begin{aligned} t_i + w_i &\geq 0 \\ t_i - w_i &\geq 0 \end{aligned}$$

which in the matrix form looks like:

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} t_i \\ w_i \end{pmatrix} \geq 0$$

Now consider this vector:

$$\begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_n \\ w_1 \\ \vdots \\ w_n \end{pmatrix}$$

The matrix A for reducing this constraint to the form $Au < b$ is then given by: Let:

$$B = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & (n-1)\text{zeroes} & \dots & 1 & 0 & \dots \\ 1 & (n-1)\text{zeroes} & \dots & -1 & 0 & \dots \\ 0 & 1 & & (n-1)\text{zeroes} & -1 & \\ & & & & -1 & \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_n \\ w_1 \\ \vdots \\ w_n \end{pmatrix} \geq 0$$

Our optimisation problem now looks like:

$$\begin{pmatrix} t \\ w \end{pmatrix}^T \begin{pmatrix} 0 & 0 \\ 0 & XX^T \end{pmatrix} \begin{pmatrix} t \\ w \end{pmatrix} + (1 \quad 1 \quad \dots (d-2)\text{times } 1 \quad d \text{ times } 0 \dots) \begin{pmatrix} t \\ w \end{pmatrix}$$

Problem 3

Problem 3: (a)

$$\min_w (\sum_i (y_i - w^T x_i)^2 + \lambda \|w\|_2^2)$$

In more compact matrix notation, let:

$$\begin{aligned} y_{n \times 1} &= (y_1 \ y_2 \ \dots \ y_n)^T \\ X_{n \times D} &= (x_1^T \ x_2^T \ \dots \ x_n^T)^T \end{aligned}$$

This notation, reduces the above function to:

$$\min_w (\|y - w^T X\|_2^2 + \lambda \|w\|_2^2)$$

$$\begin{aligned} f(w) &= \min_w (\|y - Xw\|_2^2 + \lambda \|w\|_2^2) \\ &= (y - Xw)^T (y - Xw) + \lambda w^T w \\ &= (y^T - w^T X^T)(y - Xw) + \lambda w^T w \\ &= y^T y - y^T Xw - w^T X^T y + w^T X^T Xw + \lambda w^T w \\ &= y^T y - (X^T y)^T w - w^T X^T y + w^T X^T Xw + \lambda w^T w \end{aligned}$$

$$\begin{aligned} \frac{\partial f(w)}{\partial w} &= -X^T y - X^T y + 2\lambda w + (X^T Xw + (X X^T w)) = 0 \\ &= 2\lambda w + 2X^T Xw - 2X^T y = 0 \end{aligned}$$

$$\mathbf{w}(\lambda I_D + X^T X) = X^T y$$

$$\mathbf{w} = (X^T X + \lambda I_D)^{-1} X^T y$$

Problem 3: (b)

$\min_w (\|y - w^T \Phi\|_2^2 + \lambda \|w\|_2^2)$ From the previous part, the solution should be of similar form:

$$\mathbf{w} = (\Phi^T \Phi + \lambda I_D)^{-1} \Phi^T y$$

Using the identity:

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = P B^T (B P B^T + R)^{-1}$$

Thus,

$$((\lambda I_D + \Phi^T \Phi)^{-1}) \Phi^T y = \Phi^T (\Phi \Phi^T + \lambda I_N)^{-1} y$$

$$w^* = \Phi^T (\Phi \Phi^T + \lambda I_N)^{-1} y$$

Problem 3: (c)

$$\hat{y} = w^{*T} \Phi(x)$$

$$\hat{y} = (\Phi^T(\Phi\Phi^T + \lambda I_N)^{-1}y)^T \Phi(x) = y^T ((\Phi\Phi^T + \lambda I_N)^{-1})^T \Phi^T \Phi(x)$$

Now using the property, $(A^{-1})^T = (A^T)^{-1}$

$$\begin{aligned} \hat{y} &= y^T ((\Phi\Phi^T + \lambda I_N)^{-1})^T \Phi^T \Phi(x) \\ &= y^T ((\Phi\Phi^T + \lambda I_N)^T)^{-1} \Phi^T \Phi(x) \\ &= y^T ((\Phi^T \Phi + \lambda I_N))^{-1} \Phi^T \Phi(x) \\ &= y^T (K + \lambda I_N)^{-1} \kappa(x) \end{aligned}$$

Where $K_{ij} = \Phi_i^T \Phi_j$ and $\kappa(x) = \phi^T \phi^T(x)$

Problem 3: (d)

Kernel ridge regression is $O(n^3)$ for n data points. Linear regression was formulated as quadratic programming and hence is $O(n^2)$.

The extra n factor comes from the formation of kernel matrix K .

Problem 4

Given: $k_1(.,.)$ and $k_2(.,.)$ are kernel function. Thus, for any vector $y \in \mathbf{R}$, $y^T K y \geq 0$ where $K_{ij} = k(x_i, x_j)$ Mercer's theorem requires K to be positive semi-definite.

Problem 4: (a)

$k_3(x, x') = a_1 k_1(x, x') + a_2 k_2(x, x')$ where $a_1, a_2 \geq 0$

Since $k_1(x, x')$ is positive definite, $\forall y \in \mathbf{R}$,

$$y^T K^{(1)} y \geq 0, \quad (4a.1)$$

where

$$K_{ij}^{(1)} = k_1(x_i, x'_j)$$

Similarly,

$$y^T K^{(2)} y \geq 0, \quad (4a.2)$$

where

$$K_{ij}^{(2)} = k_2(x_i, x'_j)$$

Thus, from (4a.1) and (4a.2), we get

$$y^T (K^{(1)} + K^{(2)}) y \geq 0 \quad \forall y \in \mathbf{R} \implies$$

$$y^T K^{(3)} y \geq 0 \quad \forall y \in \mathbf{R}$$

where

$$K_{ij}^{(3)} = k_3(x_i, x'_j)$$

Problem 4: (b)

$k_4(x, x') = f(x)f(x')$ Let $K_{ij}^{(4)} = k_4(x_i, x_j) = f(x_i)f(x'_j)$

Since $f(x)$ is a real valued function, consider $K^{(4)}$

$$K^{(4)} = \begin{bmatrix} f(x_1)f(x'_1) & f(x_1)f(x'_2) & \cdots & f(x_1)f(x'_n) \\ \vdots & \vdots & \ddots & \vdots \\ f(x_n)f(x'_1) & f(x_n)f(x'_2) & \cdots & f(x_n)f(x'_n) \end{bmatrix}$$

$$K^{(4)} = F(x)_{n \times 1} F(x)^T_{1 \times n}$$

where

$$F(x)^T_{1 \times n} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix}$$

Now consider $y^T K^{(4)} y = y^T F(x) F(x)^T y = y^T F(x) (y^T F(x))^T = \|y^T F(x)\|_2^2 \geq 0$

Thus, $k_2(., .)$ is a valid kernel function!.

Problem 4: (c)

$k_5(x, x') = g(k_1(x, x'))$ where g is a polynomial with positive coefficients.

Since g has positive coefficients, $g(x) \geq 0 \forall x \geq 0$

Now consider,

$$y^T K^{(5)} y = (y_1 \ y_2 \ \cdots \ y_n) \times \begin{bmatrix} g(k_1(x_1, x'_1)) & g(k_1(x_1, x'_2)) & \cdots g(k_1(x_1, x'_n)) \\ \vdots & & \\ g(k_1(x_n, x'_1)) & g(k_1(x_n, x'_2)) & \cdots g(k_1(x_n, x'_n)) \end{bmatrix} \times \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$y^T K^{(5)} y = y_1 g(k_1(x_1, x'_1)) y_1 + y_2 g(k_1(x_1, x'_2)) y_2 + \cdots y_n g(k_1(x_n, x'_n)) y_n$$

Since $g(k_1(x_i, x_j)) \geq 0$

$$y^T K^{(5)} y \geq 0 \forall y \in \mathbf{R}$$

Thus k_5 is a kernel

Problem 4: (d)

$$k_6(x, x') = k_1(x, x') k_2(x, x')$$

Thus, in terms of our earlier defined matrix notation, $K^{(6)} = K^{(1)} \circ K^{(2)}$ where \circ denotes element wise multiplication (also known as the Hadamard product).

Since, k_1 and k_2 are valid kernel function $\exists v_i w_j$ the eigen vectors of matrix K_1 and K_2 defines such that:

$$K^{(1)} = \sum_i \lambda_i v_i v_i^T \text{ and } K^{(2)} = \sum_j \mu_j w_j w_j^T$$

Now,

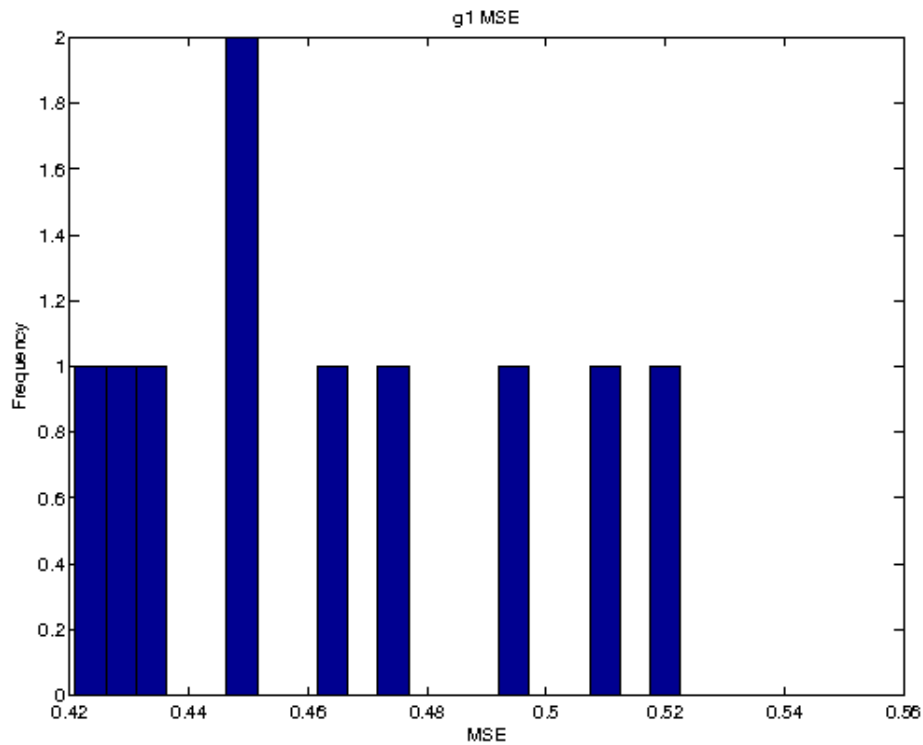
$$\begin{aligned} K^{(6)} &= K^{(1)} \circ K^{(2)} \\ &= \sum_i \lambda_i v_i v_i^T \circ \sum_j \mu_j w_j w_j^T \\ &= \sum_{i,j} \lambda_i \mu_j (v_i v_i^T) \circ w_j w_j^T \\ &= \sum_{i,j} \lambda_i \mu_j (v_i \circ w_j)(v_i \circ w_j)^T \\ &\geq 0 \end{aligned}$$

Because $(v_i \circ w_j)(v_i \circ w_j)^T = \|v_i w_j\|_2^2 \geq 0$

Problem 4: (e)

$$k_7(x, x') = \exp(k_1(x, x'))$$

Just like subpart (c), here $g(x) = \exp(x)$ (it's not a polynomial, though that does not affect the derivation we came up with in part (c)). So this is immediate from part (c).

Figure 1: Problem 5.a g_1 MSE

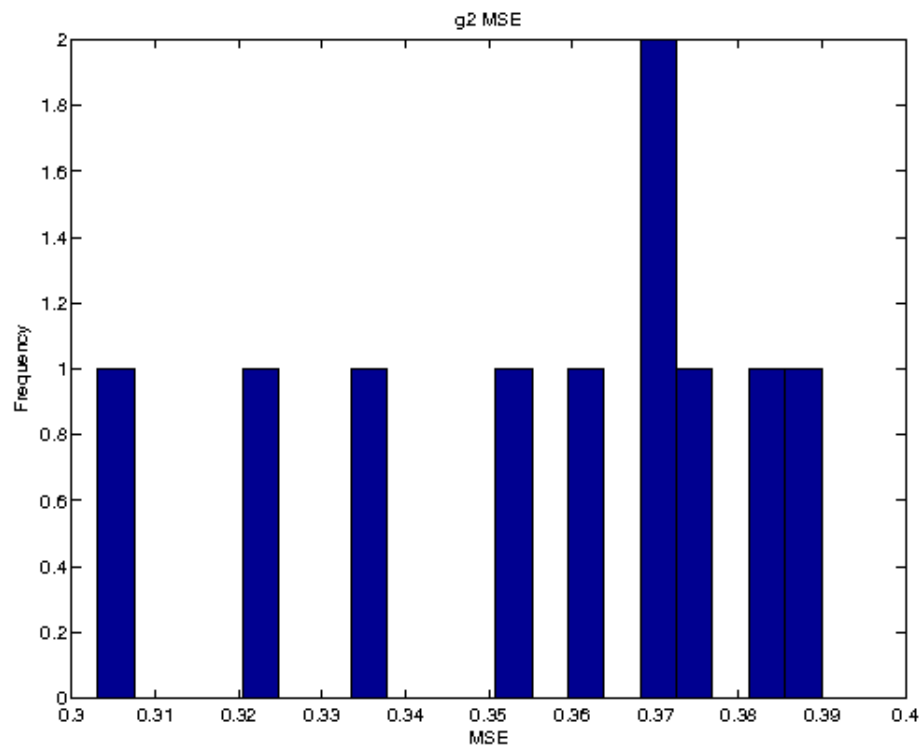
Problem 5

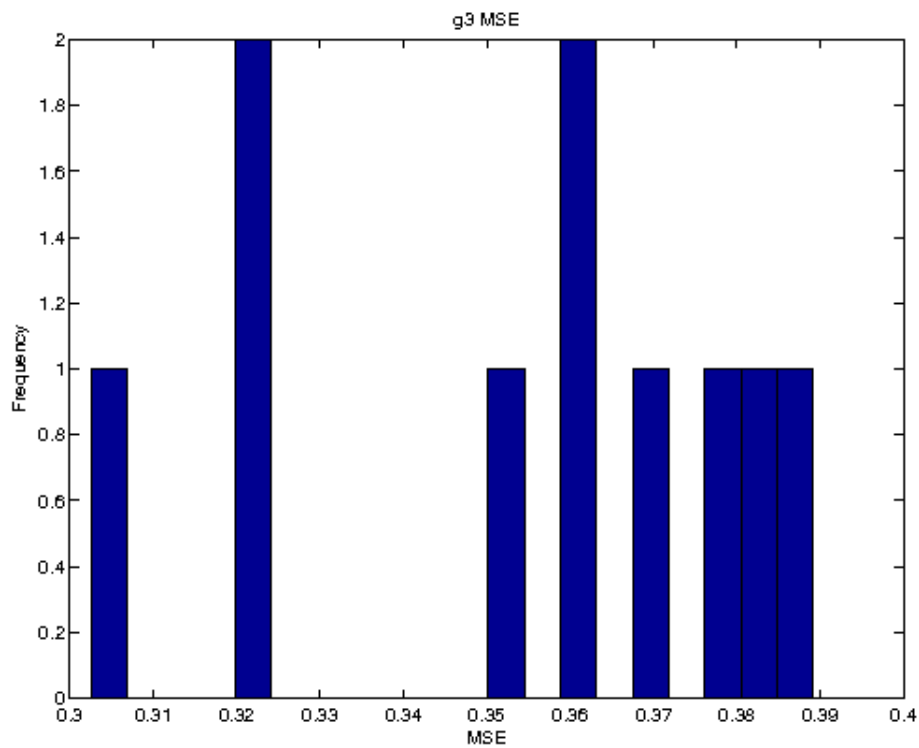
a

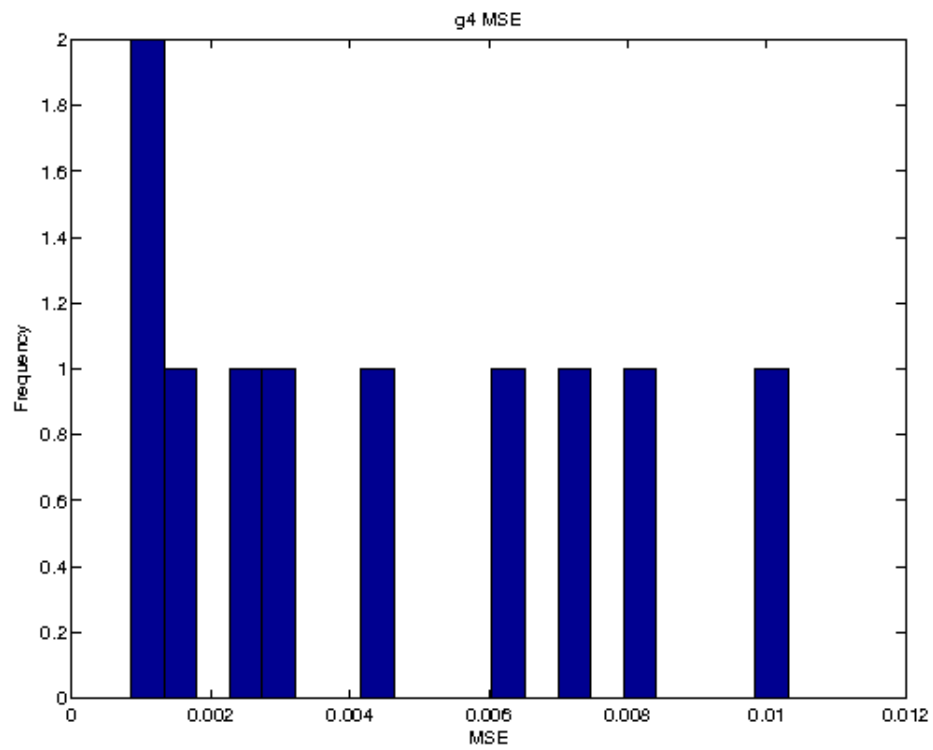
| | |
|-----------|---|
| — g_1 — | MSE:0.463977 Bias Sq:0.108996 Variance:0.000000 |
| — g_2 — | MSE:0.356683 Bias Sq:0.002941 Variance:0.003295 |
| — g_3 — | MSE:0.354618 Bias Sq:0.002844 Variance:0.007814 |
| — g_4 — | MSE:0.004551 Bias Sq:0.000151 Variance:0.003862 |
| — g_5 — | MSE:0.005546 Bias Sq:0.000151 Variance:0.004782 |
| — g_6 — | MSE:0.006223 Bias Sq:0.000125 Variance:0.005273 |

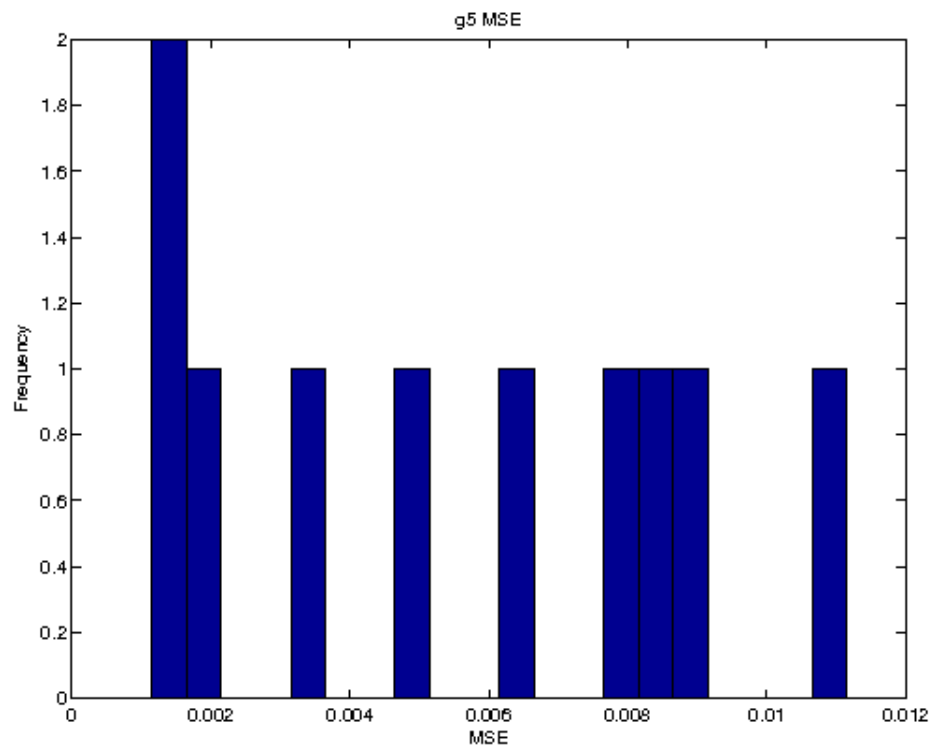
As the model complexity increase the squared bias decreases and the variance increases. However for some reason, the variance attributed with g_3 is a bit more than the normal trend. I could not think of a possible explanation for this.

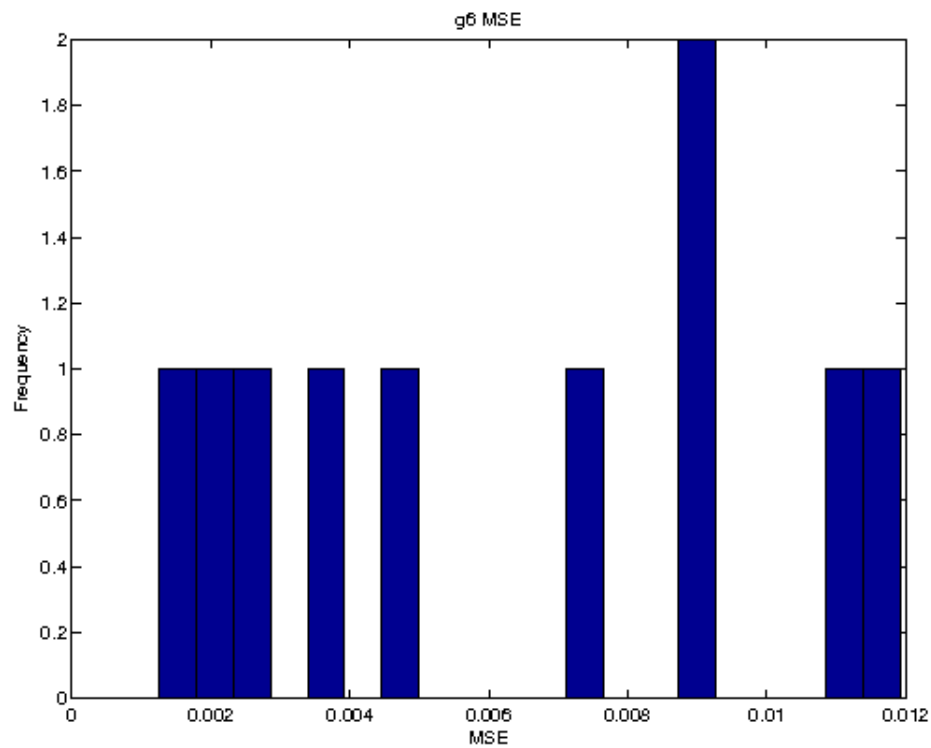
content...

Figure 2: Problem 5.a g_2 MSE

Figure 3: Problem 5.a g_3 MSE

Figure 4: Problem 5.a g_4 MSE

Figure 5: Problem 5.a g_5 MSE

Figure 6: Problem 5.a g_6 MSE