CSCI-567: Assignment #3

Due on Saturday, October 17, 2015(One late day used)

Saket Choudhary
2170058637

Contents

Problem 1

Problem 1: (a)

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

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

Observe that $Y=1$ when $b+w^Tx \geq 0$ and $Y=0$ when $b+w^Tx < 0$
Thus,

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

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

Consider $\mathcal{L}(w) = -y(b+w^Tx) + \log(1+e^{(b+w^Tx_i)})$

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

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 continued on next page...

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 1 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 $||w|| \longrightarrow \infty$

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)

$$\mathcal{L}(w) = \sum_{j=1}^{n} \big(-y_j(b + w^T x_j) + \log(1 + e^{(b+w^T x_j)})\big) + \lambda||w||_2^2$$

$$\frac{\partial(\mathcal{L})(w)}{\partial w_i} = \sum_{j=1}^{n} \big(-y_j(x_{ji}) + \frac{x_{ji}e^{(b+w^T x_j)x_{ij}}}{1 + e^{(b+w^T x_j)}}\big) + 2\lambda w_i = 0$$

$$\frac{\partial^2(\mathcal{L})(w)}{\partial w_i^2} = \sum_{j=1}^{n} \big(\frac{x_{ji}^2 e^{(b+w^T x_j)x_{ij}}}{(1 + e^{(b+w^T x_j)})^2}\big) + 2\lambda > 0$$

where the last inequality holds since $\lambda > 0$ Consider $\boldsymbol{f(w_i) = \sum_{j=1}^{n} \big(-y_j(x_{ji}) + \frac{x_{ji}e^{(b+w^T x_j)x_{ij}}}{1+e^{(b+w^T x_j)}}\big) + 2\lambda w_i = 0}$
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, $f'(w_i) = \sum_{j=1}^{n} \big(\frac{x_{ji}^2 e^{(b+w^T x_j)x_{ij}}}{(1+e^{(b+w^T x_j)})^2}\big) + 2\lambda > 0$ 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: (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:

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

Thus,

$$\begin{aligned} f(0) &= 0 \\ f(\epsilon) &= 1 \\ f((1-\lambda)\epsilon) &= 1 \\ f(\lambda \times 0 + (1-\lambda) \times \epsilon) &= 1 \ \forall \ 0 < \lambda < 1 \qquad (2a.1) \\ \lambda f(0) + (1-\lambda) f(\epsilon) &= 1 - \lambda < 1 = f(\lambda \times 0 + (1-\lambda) \times \epsilon) \qquad (2a.2) \end{aligned}$$

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: $0\lambda < 1$

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

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 continued on next page...

Problem 2: (c)

Let's redefine(for the sake of easense)x_i to be column vector i.e $x_i\ is D\times 1\ w^T is 1\times D$ and and $Y = (y_1 \dots y_n)$ the equivalent porblem then becomes:

$$min_w \sum_i (y_i - x_i^w)^2$$

$$\min_w \sum_i (y_i - x_i^T w)^2 + \lambda ||w||_1$$

$$\min_w (y - X^T w)^T (y - X^T w) + \lambda ||w||_1$$

$$\min_w (w^T X X^T w - 2Y^T X w + Y^T Y) + \min_w \lambda ||w||_1$$

$$\min_w (w^T X X^T w - 2Y^T X w) + \min_w \lambda ||w||_1$$

We introduce dummy variables t_i such that:

$$||w_i|| \le t_i \implies t_i \ge w_i \text{and} t_i \ge -w_i$$

Now,

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

Constraint:

$$t_i + w_i \ge 0$$
$$t_i - w_i \ge 0$$

which in the matrix form looks like:

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

Now consider this vector,:

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

which in short form is :

$$\begin{pmatrix} t \\ w \end{pmatrix}$$

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

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

$$\begin{pmatrix} 1 & (n-1)zeroes\ldots & 1 & 0\ldots & 0 & 0 & 0 \\ 1 & (n-1)zeroes\ldots & -1 & 0\ldots & 0 & 0 & 0 \\ 0 & 1 & (n-1)zeroes & 1 & 0\ldots & 0 & 0 \\ 0 & 1 & (n-1)zeroes & -1 & 0\ldots & 0 & 0 \\ 0 & 0 & 1 & (n-1)zeroes & 1 & 0\ldots & 0 \\ 0 & 0 & 1 & (n-1)zeroes & -1 & 0\ldots & \\ \vdots & \vdots & \vdots & 0\ldots & & & 1 \\ & & 0\ldots & & & & -1 \end{pmatrix}_{2n\times 2n} \begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ \vdots t_n \\ w_1 \\ \vdots w_n \end{pmatrix}_{2n\times 1} \geq 0$$

Our optimisation problem now looks like:

$$\min_{t,w} \begin{pmatrix} t \\ w \end{pmatrix}^T_{1\times 2n} \begin{pmatrix} 0 & 0 \\ 0 & XX^T \end{pmatrix}_{2n\times 2n} \begin{pmatrix} t \\ w \end{pmatrix}_{2n\times 1} + \begin{pmatrix} 1 & \ldots (n-2)\text{times } 1 & 0\ldots & \text{n times } 0\ldots \end{pmatrix}^T_{1\times 2n} \begin{pmatrix} t \\ w \end{pmatrix}_{2n\times 1}$$

with the constraint:

$$\begin{pmatrix} 1 & (n-1)zeroes\ldots & 1 & 0\ldots & 0 & 0 & 0 \\ 1 & (n-1)zeroes\ldots & -1 & 0\ldots & 0 & 0 & 0 \\ 0 & 1 & (n-1)zeroes & 1 & 0\ldots & 0 & 0 \\ 0 & 1 & (n-1)zeroes & -1 & 0\ldots & 0 & 0 \\ 0 & 0 & 1 & (n-1)zeroes & 1 & 0\ldots & 0 \\ 0 & 0 & 1 & (n-1)zeroes & -1 & 0\ldots & \\ \vdots & \vdots & \vdots & 0\ldots & & & 1 \\ & & 0\ldots & & & & -1 \end{pmatrix}_{2n\times 2n} \begin{pmatrix} t \\ w \end{pmatrix}_{2n\times 1} \geq 0$$

which is a QP formulation. of the form:

$$\min_u u^T Q u + c^T u$$
$$Au^T \leq b$$

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:

$$y_{n\times 1} = (y_1\ y_2\ \cdots\ y_n)^T$$
$$X_{n\times D} = (x_1^T\ x_2^T\ \cdots\ x_n^T)^T$$

This notation, reduces the above function to:
$min_w(||y - w^T X||_2^2 + \lambda||w||_2^2)$

$$\begin{aligned} f(w) &= min_w(||y - Xw||_2^2 + \lambda||w||_2^2) \\ &= (y - Xw)^T(y - Xw) + \lambda w^T w \\ &= (y^T - w^T X^T)(y - Xw) + \lambda w^T w \\ &= y^T y - y^T Xw - w^T X^T y + w^T X^T Xw + \lambda w^T w \\ &= y^T y - (X^T y)^T w - w^T X^T y + w^T X^T Xw + \lambda w^T w \\ \frac{\partial f(w)}{\partial w} &= -X^T y - X^T y + 2\lambda w + (X^T Xw + (XX^T w)) = 0 \\ &= 2\lambda w + 2X^T Xw - 2X^T y = 0 \\ \mathbf{w}(\lambda I_D + X^T w) &= X^T y \end{aligned}$$

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

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:

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

Thus,

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

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

Problem 3: (c)

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

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

Now using the property,

$$(A^{-1})^T = (A^T)^{-1}$$

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

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 programing and hence is $O(n^2)$.
Kernel $n \times n$ instead of $d \times d$(for vanilla ridge regression without kernel). In cases where $d < n$ this leads to an additional n operations for calculating K itself.

Problem 4

Given: $k_1(.,.)$ and $k_2(.,.)$ are kernel function. Thus, for any vector $y \in \mathbf{R}$, $y^TKy \geq 0$ where $K_{ij} = k(x_i, x_j)$ Mercer's theorem requires K to be positive semi-definite.

Problem 4 continued on next page...

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, $\forall y \in \mathbf{R}$,

$$y^T K^{(1)} y \geq 0, \tag{4a.1}$$
where
$$K_{ij}^{(1)} = k_1(x_i, x'_j)$$

Similarly,

$$y^T K^{(2)} y \geq 0, \tag{4a.2}$$
where
$$K_{ij}^{(2)} = k_2(x_i, x'_j)$$

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

$$y^T (K^{(1)} + K^{(2)}) y \geq 0 \ \forall y \in \mathbf{R} \implies$$
$$y^T K^{(3)} y \geq 0 \ \forall y \in \mathbf{R}$$
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 $K^{(4)}$

$$K^{(4)} = \begin{bmatrix} f(x_1)f(x'_1) & f(x_1)f(x'_2) & \cdots & f(x_1)f(x'_n) \\ \vdots & & & \\ f(x_n)f(x'_1) & f(x_n)f(x'_2) & \cdots & f(x_n)f(x'_n) \end{bmatrix}$$
$$K^{(4)} = F(\vec{x})_{n\times 1} F(\vec{x})^T_{1\times n}$$
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 continued on next page...

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,

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

Since K_1 is positive definite, it is possible to compute its diagonal formulation:

$$K_1 = P \Lambda P^{-1}$$

where Λ is a diagonal matrix with the diagonals equal to the eigen values (all non-negative).

$$y^T K^{(5)} y = y^T (P g(\Lambda) P^{-1}) y$$

$$y^T K^{(5)} y = (y_1\ y_2 \cdots y_n) \times \begin{bmatrix} g(\lambda_1) & 0 & \cdots 0 \\ \vdots & & \\ 0 & 0 & \cdots g(\lambda_n) \end{bmatrix} \times \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

where $\lambda_i \geq 0$ and hence $g(\Lambda)$ is semi=positive definite since $g(\lambda_1) \geq 0$ (as $\lambda_i \geq 0$ and g is polynomial with positive coefficients)
Thus,

$$y^T K^{(5)} y \geq 0$$

Problem 4 continued on next page...

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

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

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

Problem 4: (e)

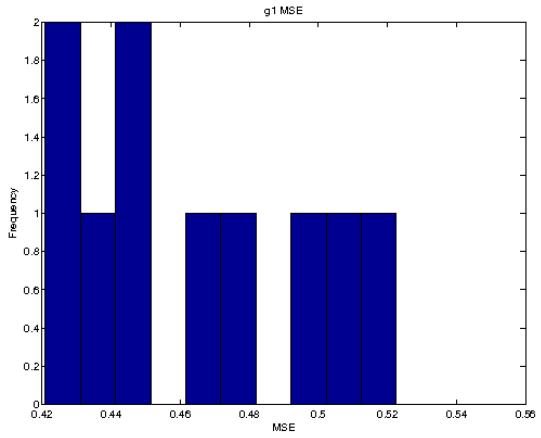
$k_7(x, x') = \exp(k_1(x, x'))$
Just like subpart (c), here $g(x) = exp(x) = 1 + x + x^2/2! + x^3/3! \ldots$ (it's an polynomial with infinite terms and all coefficients are positive)

Problem 5

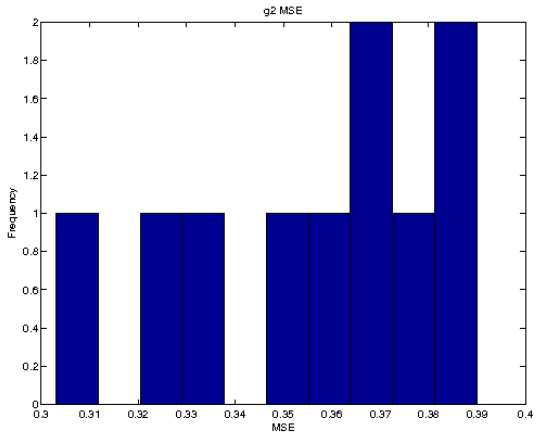
Problem 5: (a)

g	MSE	$Bias^2$	$Variance$
g_1	0.463977	0.108996	0.00
g_2	0.356683	0.002941	0.003295
g_3	0.354618	0.002844	0.007814
g_4	0.004551	0.000151	0.003862
g_5	0.005546	0.000151	0.004782
g_6	0.006223	0.000125	0.005273

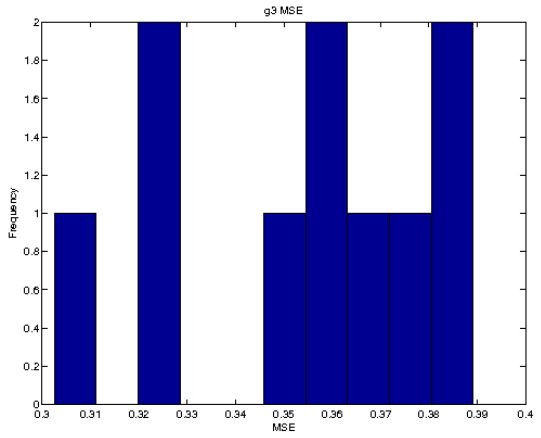
Problem 5 continued on next page...



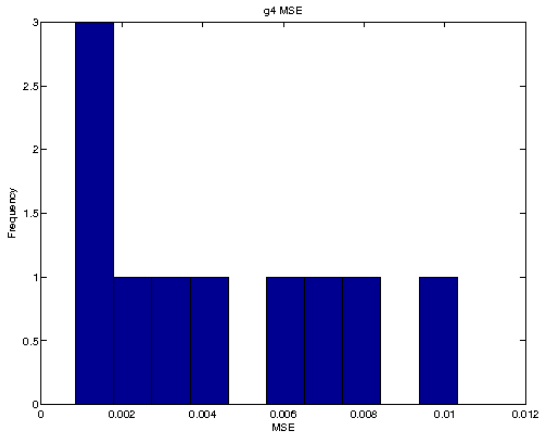
(a) Problem 5.a g_1 MSE



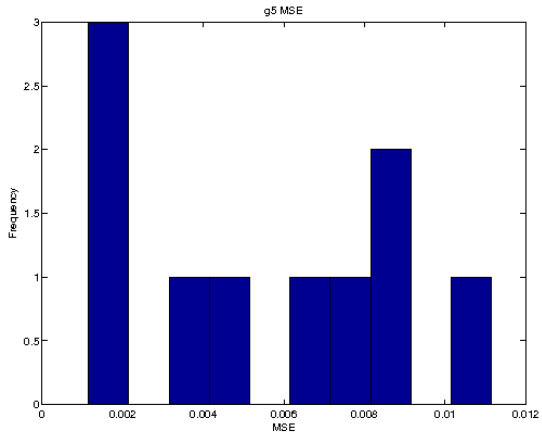
(b) Problem 5.a g_2 MSE



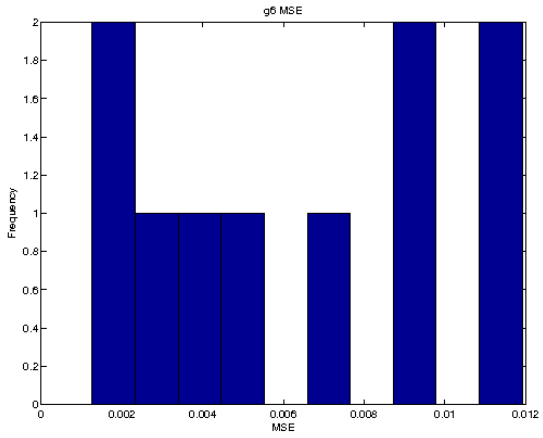
(a) Problem 5.a g_3 MSE



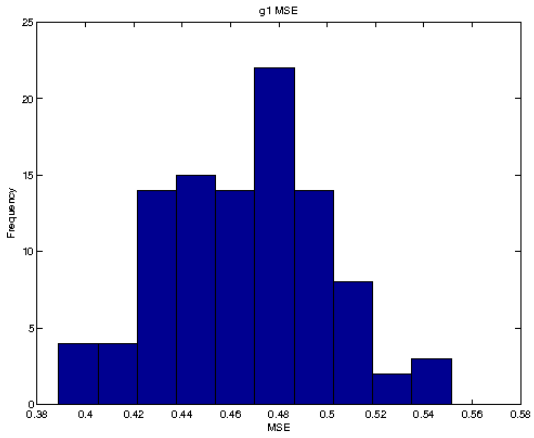
(b) Problem 5.a g_4 MSE



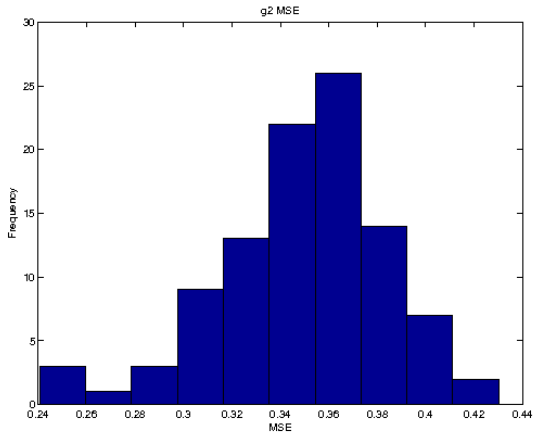
(a) Problem 5.a g_5 MSE



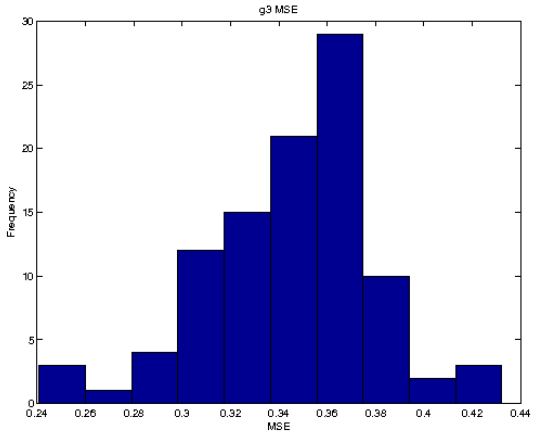
(b) Problem 5.a g_6 MSE



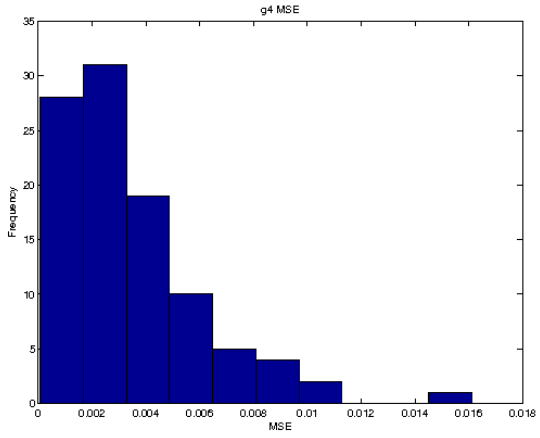
(a) Problem 5.b g_1 MSE



(b) Problem 5.b g_2 MSE



(a) Problem 5.b g_3 MSE

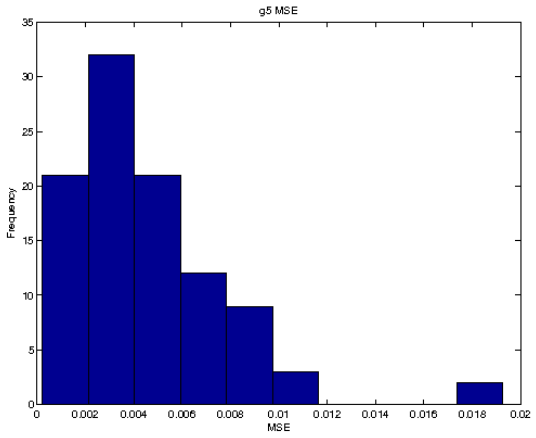


(b) Problem 5.b g_4 MSE

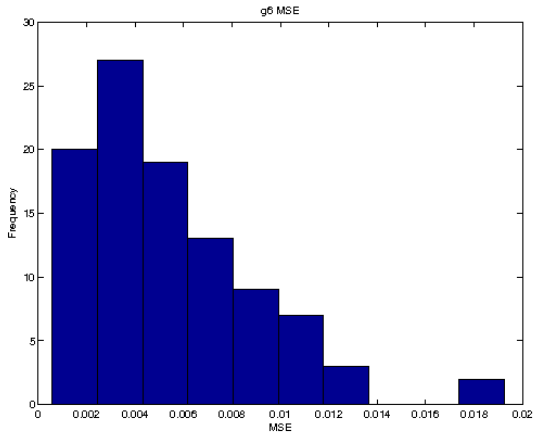
Problem 5: (b)

g	MSE	$Bias^2$	$Variance$
g_1	0.466517	0.118490	0.00
g_2	0.349509	0.003121	0.004464
g_3	0.345009	0.003141	0.010835
g_4	0.003469	0.000002	0.003411
g_5	0.004510	0.000002	0.004441
g_6	0.005375	0.000004	0.005292

Problem 5 continued on next page...



(a) Problem 5.b g_5 MSE



(b) Problem 5.b g_6 MSE

Problem 5: (c)

As the model complexity increase the squared bias decreases and the variance increases and the mean squared error decreases. 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.
Another point to realise is that the variance is decreases for g_4 since it is a second order polynomial just like $f(x) = 2x^2$.

Problem 5: (d)

$lambda$	MSE	$Bias^2$	$Variance$
0.01	0.006682	0.000356	0.003399
0.1	0.014202	0.001268	0.003537
1	0.024060	0.002480	0.003716
10	0.035110	0.003843	0.003891

Thus, as the λ increases the bias increases and the variance seems to increase too(unexpected, but anyway it is marginal). Since this is regularization problem, a higher λ will try to further penalise the larger coefficients and hence the coefficients tend to be close to zero, thus the bias increases. Ideally the variance should have decreased(again because the coefficients are now smaller!) but this does not seem to be reflected in my simulation.

Problem 6

Linear Ridge Regression

Split	Optimal λ	MSE
1	0.01	0.016042
2	0.0001	0.016664
3	0	0.017038

Mean test error: 0.016581

Problem 6 continued on next page...

Linear Kernel Ridge Regression

Split	Optimal λ	MSE
1	0.01	0.016101
2	0.0001	0.016829
3	0.0	0.017040

Mean test error: 0.016657

Polynomial Kernel Ridge Regression

Split	Optimal λ	Optimal a	Optimal b	MSE
1	0.01	0.5	2	$1.268206 * 10^{-02}$
2	1	0.0	3	$1.227860 * 10^{-02}$
3	10	1.0	3	$1.286726 * 10^{-02}$

Mean test Error: 0.012609

RBF Gaussian Kernel Ridge Regression

Split	Optimal λ	Optimal σ^2	MSE
1	0.001	8	0.013382
2	0.01	8	0.012444
3	0.01	8	0.012080

Mean test Error: 0.012635

Comparison

Linear Ridge Regression mean test error : 0.016581
Linear Kernel Ridge Regression mean test error : 0.016657
Kernel ridge regression with linear kernel does not give the "same" results(they are very close though), and the thing to realise in this case is that linear kernel projects the data into $N \times N$ dimensions, while the ridge regression still has the 'kernel' in $D \times D$ dimensions. There is extra information being used here (in cases where $N > D$) In a a situation where $D > N$ the linear kernl might perform better.(I don't have a proof for this)

Kernel	Mean Test Error
Linear no Kernel	0.016581
Linear Kernel	0.016657
Polynomial Kernel	0.012609
Gaussian Kernel	0.012635

From the table we see, the Polynomial kernel with an average polynomial degree of 2.5 performs the best, though Gaussian comes close. It looks like the poylnomial kernel in this case is better able to capture the non linearity in 2 or 3 dimensions, while the gaussian uses infinite dimensions.