CSCI 567-Machine Learning Assignment-3

1 Kernel Methods

1.1 Kernel Function - Symmetric Positive Semi Definite

k is a positive semi-definite matrix. To prove k is a symmetric matrix:

$$k(x_i, x_j) = C = k(x_i, x_j)$$

To prove it is positive semi-definite, let's say v is a vector.

$$v^T k(x_i, x_j) v = C v^T v$$

$v^T v$ is sum of squares for vector v. So,

$$C v^T v \geq 0 \iff C \geq 0$$

2 Support Vector Machines

$$\min_{R,\mathbf{c},\varepsilon} \frac{1}{2} R^2 + C \sum_n \varepsilon_n \tag{1}$$

$$s.t. ||\phi(x_i) - b||_2^2 \leq R^2 + \varepsilon_n \tag{2}$$

$$\varepsilon_n \geq 0 \tag{3}$$

We take the Lagrangian of this primal.

$$\mathcal{L}(R, b, \varepsilon, \alpha, \eta) = \frac{1}{2} R^2 + C \sum_n \varepsilon_n + \sum_n \alpha_n [(\phi(x_n) - b)'(\phi(x_n) - b) - R^2 - \varepsilon_n] - \sum_n \eta_n \varepsilon_n \tag{4}$$

Let's take derivatives of primal variables and equal to 0.

$$\frac{\partial \mathcal{L}}{\partial R} = R - 2R \sum_n \alpha_n = 0 \rightarrow \sum_n \alpha_n = \frac{1}{2} \tag{5}$$

Using this equation,

$$\frac{\partial \mathcal{L}}{\partial b} = 2 \sum_n \alpha_n b - 2 \sum_n \alpha_n \phi(x_n) = 0 \rightarrow b = 2 \sum_n \alpha_n \phi(x_n) \tag{6}$$

$$\frac{\partial \mathcal{L}}{\partial \varepsilon} = C - \alpha_n - \eta_n = 0 \rightarrow \alpha_n + \eta_n = C \tag{7}$$

If we fill this equations in Lagrangian

$$\begin{split}\mathcal{L}(R, b, \varepsilon, \alpha, \eta) = \frac{1}{2}R^2 + C\sum_n \varepsilon_n + \sum_n \alpha_n k(x_n, x_n) - 2\sum_n \alpha_n \phi(x_n)^T b + b^T b \sum_n \alpha_n - \\ R^2 \sum_n \alpha_n - \sum_n (\alpha_n + \eta_n)\epsilon_n\end{split} \tag{8}$$

Using new constraints, we simplify this equation.

$$\mathcal{L}(R, b, \varepsilon, \alpha, \eta) = \sum_n \alpha_n k(x_n, x_n) - \frac{1}{2}b^T b \tag{9}$$

Dual of this LP is

$$\max_\alpha \sum_n \alpha_n k(x_n, x_n) - 2\sum_i \sum_j \alpha_i \alpha_j k(x_i, x_j) \tag{10}$$

$\sum_n \alpha_n = \frac{1}{2}$ and $k(x_n, x_n)$ is a constant. So, our LP is

$$\arg\min_\alpha 2\sum_i \sum_j \alpha_i \alpha_j k(x_i, x_j) \tag{11}$$

$$0 \le \alpha_n \le C, \forall \tag{12}$$

$$\sum \alpha_n = \frac{1}{2} \tag{13}$$

α_n must be less or equal than C since $\alpha_n + \eta_n = C$ and $\eta_n \ge 0$

Let's find R value. $\alpha_n = C$ for abnormal values. And, for any data point that is on the boundary $0 \le \alpha_n \le C$. Also, we can say that $||\phi(x_i) - b||_2^2 = R^2$ for data points on the boundary.

$$||\phi(x_i) - b||_2^2 = R^2 \tag{14}$$

$$b = 2\sum \alpha_n \phi(x_n) \tag{15}$$

$$R^2 = k(x_i, x_i) - 2\sum_j \alpha_j k(x_i, x_j) + \sum_i \sum_j \alpha_i \alpha_j k(x_i, x_j) \tag{16}$$

If the value for a x variable $||\phi(x) - b||_2^2 > R^2$, then it is abnormal.

3 Boosting

3.1 Exponential Loss and Logistic Loss

3.1.1 Similarity

For Exponential Loss:

$$\mathcal{L}_e(f) = E_{x,y}[\exp(-yf(x))]$$

$$E_{x,y}[\exp(-yf(x))|x] = p(y=1|x)\exp(-f(x)) + p(y=-1|x)\exp(f(x))$$

$$\frac{\partial E_{x,y}[\exp(-yf(x))|x]}{\partial f(x)} = -p(y=1|x)e^{-f(x)} + p(y=-1|x)e^{f(x)} = 0$$

$$p(y=-1|x)e^{f(x)} = p(y=1|x)e^{-f(x)}$$

$$log(p(y=-1|x)) + f(x) = log(p(y=1|x)) - f(x)$$

$$f(x) = \frac{1}{2}log\frac{p(y=1|x)}{p(y=-1|x)}$$

Let's take exponential of this term:

$$e^{2f(x)} = \frac{p(y=1|x)}{p(y=-1|x)} \rightarrow e^{2f(x)} = \frac{p(y=1|x)}{1-p(y=1|x)}$$

$$e^{2f(x)} - e^{2f(x)}p(y=1|x) - p(y=1|x) = 0$$

$$p(y|x) = p(y=1|x) = \frac{e^{2f(x)}}{1+e^{2f(x)}}$$

Divide both nominator and denominator by $e^{f(x)}$:

$$p(y|x) = \frac{e^{f(x)}}{e^{-f(x)} + e^{f(x)}}$$

For Logistic Loss:

$$\mathcal{L}_\sigma(f) = E_{x,y}log[1+\exp(-2yf(x))]$$

$$E[log(1+e^{-2yf(x)})|x] = p(y=1|x)log(1+e^{-2f(x)}) + p(y=-1|x)log(1+e^{2f(x)})$$

$$\frac{\partial E[log(1+e^{-2yf(x)})|x]}{\partial f(x)} = \frac{-2p(y=1|x)e^{-2f(x)}}{1+e^{-2f(x)}} + \frac{2p(y=-1|x)e^{2f(x)}}{1+e^{2f(x)}} = 0$$

Let's divide the RHS's first term's nominator and denominator by $e^{-2f(x)}$:

$$\frac{\partial E[log(1+e^{-2yf(x)})|x]}{\partial f(x)} = \frac{-2p(y=1|x)}{e^{2f(x)}+1} + \frac{2p(y=-1|x)e^{2f(x)}}{1+e^{2f(x)}} = 0$$

Simplify the equation:

$$e^{2f(x)} - p(y=1|x)(1+e^{2f(x)}) = 0$$

$$p(y|x) = p(y|x=1) = \frac{e^{2f(x)}}{1+e^{2f(x)}} = \frac{e^{f(x)}}{e^{-f(x)}+e^{f(x)}}$$

3.1.2 Difference

3.2 Boosting Using Log Loss

4 Bias-Variance Trade-off

4.1 Closed Form Solution and Distribution of $\hat{\beta_\lambda}$

$$\hat{\beta_\lambda} = \arg\min_{\hat{\beta_\lambda}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda||\beta||_2^2 \right\} \tag{17}$$

Define Y $nx1$ matrix for n data point. X pxn matrix for n data points in $p-1$-dimension. In this case.

$$\hat{\beta_\lambda} = \arg\min_{\hat{\beta_\lambda}} \left\{ (Y - X\beta)'(Y - X\beta) + \beta\lambda I\beta \right\} \tag{18}$$

$$= (Y^T - \beta^T X^T)(Y - X\beta) + \beta^T\lambda I\beta$$

$$= Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta + \beta^T\lambda I\beta$$

$$\frac{\partial\hat{\beta_\lambda}}{\partial\beta} = X^TY - X^TY + 2X^TX\beta + 2\lambda I\beta = 0 \tag{19}$$

$$\hat{\beta_\lambda} = (X^TX + \lambda I)^{-1}X^TY \tag{20}$$

The only random variable in $\hat{\beta_\lambda}$ is Y. And it is Gaussian with mean $X^T\beta^*$, and variance σ^2. In this case,

$$E[\hat{\beta_\lambda}] = E[(X^TX + \lambda I)^{-1}X^TY|X]$$

$$E[\hat{\beta_\lambda}] = (X^TX + \lambda I)^{-1}X^TE[Y|X]$$

$$E[Y|X] = X\beta^*$$

$$E[\hat{\beta_\lambda}] = (X^TX + \lambda I)^{-1}X^TX\beta^*$$

The variance of $\hat{\beta_\lambda}$ is:

$$Var[\hat{\beta_\lambda}] = Var[(X^TX + \lambda I)^{-1}X^TY|X]$$

$$Var[\hat{\beta_\lambda}] = ((X^TX + \lambda I)^{-1}X^T)Var[Y|X]((X^TX + \lambda I)^{-1}X^T)^T$$

$$Var[Y|X] = \sigma^2 I$$

$$Var[\hat{\beta_\lambda}] = ((X^TX + \lambda I)^{-1}X^T)\sigma^2 I((X^TX + \lambda I)^{-1}X^T)^T$$

The distribution of $\hat{\beta_\lambda}$ is Gaussian with mean and variance as stated above.

4.2 Bias Term

$$E[X^T\hat{\beta_\lambda}] - X\beta^* = XE[\hat{\beta_\lambda}] - X\beta^*$$

$$= X(X^TX + \lambda I)^{-1}X^TX\beta^* - X\beta^*$$

$$= X(X^TX + \lambda I)^{-1}(X^TX + \lambda I - \lambda I)\beta^* - X\beta^*$$

$$= X(I - \lambda(X^TX + \lambda I)^{-1})\beta^* - X\beta^* = X\lambda(X^TX + \lambda I)^{-1}\beta^*$$

4.3 Variance Term

The variance term of $\hat{\beta_\lambda}$ is as stated in section 4.1.

$$Var[\hat{\beta_\lambda}] = ((X^TX + \lambda I)^{-1}X^T)\sigma^2 I((X^TX + \lambda I)^{-1}X^T)^T$$

4.4 Impact of λ on Bias and Variance Terms

Let's start with bias term. If λ increases this will result in an increase in bias term. On the other hand, in the variance term, if λ term increases, it will decrease the $Var[\hat{\beta_\lambda}]$.

5 Support Vector Machines-Programming

5.1 Preprocessing

5.2 Coding

5.3 Cross validation for linear SVM

5.3.1 3-fold CV Results

C-Value	4^{-6}	4^{-5}	4^{-4}	4^{-3}	4^{-2}	4^{-1}	4^{0}	4^{1}	4^{2}
Average Acc	0.5312	0.5312	0.5312	0.5312	0.5364	0.5963	0.6328	0.6536	0.6718
Average Time	0.9428	0.4438	0.5761	0.4962	0.4073	0.4662	0.4740	0.5301	0.5502

As value of C increases, average accuracy is steadily increasing until 4^2. Time of Cross Validation has a local minimum at 4^{-5} and slowly increasing with C value after that. It has the local maximum at $^{-6}$, which takes more than double of any other C-value's time.

5.3.2 Best C-Value

According to 3-fold cross-validation the best C-value is 16.

5.3.3 Test Accuracy

Using 16 as C-Value, we get a test accuracy of 0.73107.

5.4 Use linear SVM in LIBSVM

C-Value	4^{-6}	4^{-5}	4^{-4}	4^{-3}	4^{-2}	4^{-1}	4^{0}	4^{1}	4^{2}
Average Acc	53.125	53.125	53.125	53.125	55.469	58.984	60.677	64.453	67.448
Average Time	0.1587	0.1207	0.1202	0.1216	0.1195	0.1208	0.1248	0.1136	0.1236

Cross validation accuracy is pretty similar to 5.3. There are minor changes possibly because of randomness of cross-validation. But, average time spent on cross validation is pretty short comparing to 5.3.

5.5 Use kernel SVM in LIBSVM

5.5.1 Polynomial kernel

For degree = 1:

C-Value	Average Accuracy	Average Time
4^{-3}	53.125	0.12425
4^{-2}	53.125	0.12369
4^{-1}	53.125	0.12384
4^{0}	54.818	0.12315
4^{1}	58.594	0.11923
4^{2}	60.807	0.11688
4^{3}	64.974	0.11716
4^{4}	67.057	0.130757
4^{5}	69.401	0.15034
4^{6}	71.745	0.35991
4^{7}	72.135	0.79701

For degree 2;

C-Value	Average Accuracy	Average Time
4^{-3}	53.125	0.12565
4^{-2}	53.125	0.12488
4^{-1}	53.125	0.12614
4^{0}	53.125	0.126
4^{1}	56.38	0.1255
4^{2}	58.464	0.12166
4^{3}	60.938	0.11653
4^{4}	65.104	0.11652
4^{5}	70.573	0.12102
4^{6}	72.786	0.15023
4^{7}	74.349	0.3002

For degree = 3;

C-Value	Average Accuracy	Average Time
4^{-3}	53.125	0.12275
4^{-2}	53.125	0.12287
4^{-1}	53.125	0.12535
4^{0}	53.125	0.1248
4^{1}	53.125	0.12619
4^{2}	56.771	0.1239
4^{3}	59.245	0.12269
4^{4}	60.547	0.11843
4^{5}	65.104	0.11745
4^{6}	71.094	0.12091
4^{7}	72.266	0.16379

5.5.2 RBF kernel

For gamma = 4^{-7};

C-Value	Average Accuracy	Average Time
4^{-3}	53.125	0.15118
4^{-2}	53.125	0.15177
4^{-1}	53.125	0.15026
4^{0}	53.125	0.14968
4^{1}	53.125	0.15068
4^{2}	53.125	0.15243
4^{3}	53.125	0.154
4^{4}	53.125	0.15222
4^{5}	57.292	0.15112
4^{6}	59.766	0.14909
4^{7}	62.63	0.14367

For gamma = 4^{-6};

C-Value	Average Accuracy	Average Time
4^{-3}	53.125	0.15134
4^{-2}	53.125	0.14994
4^{-1}	53.125	0.15141
4^{0}	53.125	0.15205
4^{1}	53.125	0.15751
4^{2}	53.125	0.15124
4^{3}	53.125	0.14967
4^{4}	57.292	0.15136
4^{5}	59.766	0.14486
4^{6}	62.63	0.1437
4^{7}	65.885	0.14523

For gamma = 4^{-5};

C-Value	Average Accuracy	Average Time
4^{-3}	53.125	0.14931
4^{-2}	53.125	0.149
4^{-1}	53.125	0.15068
4^{0}	53.125	0.14984
4^{1}	53.125	0.15004
4^{2}	53.125	0.1506
4^{3}	57.292	0.15056
4^{4}	60.026	0.14564
4^{5}	62.5	0.14215
4^{6}	65.885	0.14173
4^{7}	68.88	0.15487

For gamma = 4^{-4};

C-Value	Average Accuracy	Average Time
4^{-3}	53.125	0.14985
4^{-2}	53.125	0.14978
4^{-1}	53.125	0.15052
4^{0}	53.125	0.15128
4^{1}	53.125	0.15209
4^{2}	57.292	0.15085
4^{3}	60.156	0.15056
4^{4}	62.24	0.14143
4^{5}	65.885	0.14019
4^{6}	68.88	0.1451
4^{7}	71.484	0.18975

For gamma = 4^{-3};

C-Value	Average Accuracy	Average Time
4^{-3}	53.125	0.15312
4^{-2}	53.125	0.15268
4^{-1}	53.125	0.15031
4^{0}	53.125	0.15079
4^{1}	57.422	0.14918
4^{2}	59.896	0.14707
4^{3}	62.37	0.14279
4^{4}	66.276	0.14117
4^{5}	69.922	0.14522
4^{6}	72.917	0.17701
4^{7}	73.177	0.43309

For gamma = 4^{-2};

C-Value	Average Accuracy	Average Time
4^{-3}	53.125	0.15612
4^{-2}	53.125	0.16143
4^{-1}	53.125	0.40765
4^{0}	57.682	0.238
4^{1}	58.594	0.16691
4^{2}	61.979	0.14471
4^{3}	67.318	0.16136
4^{4}	70.313	0.13959
4^{5}	74.089	0.1862
4^{6}	74.219	0.29778
4^{7}	73.828	0.8001

For gamma = 4^{-1};

C-Value	Average Accuracy	Average Time
4^{-3}	53.125	0.15304
4^{-2}	53.125	0.14935
4^{-1}	57.422	0.1543
4^{0}	58.333	0.1465
4^{1}	62.5	0.14031
4^{2}	66.797	0.13641
4^{3}	71.875	0.16753
4^{4}	73.828	0.3041
4^{5}	73.828	0.1862
4^{6}	72.786	0.79074
4^{7}	71.094	2.2725

For gamma = 4^{-0};

C-Value	Average Accuracy	Average Time
4^{-3}	53.125	0.15294
4^{-2}	54.5575	0.15293
4^{-1}	57.422	0.14791
4^{0}	61.198	0.13945
4^{1}	67.057	0.14016
4^{2}	70.443	0.13588
4^{3}	70.443	0.15595
4^{4}	70.573	0.25515
4^{5}	70.964	0.53631
4^{6}	70.573	1.713
4^{7}	68.88	5.9055

For gamma = 4^1;

C-Value	Average Accuracy	Average Time
4^{-3}	53.125	0.14965
4^{-2}	55.99	0.14971
4^{-1}	59.766	0.14522
4^{0}	64.453	0.1391
4^{1}	66.536	0.13566
4^{2}	69.01	0.14657
4^{3}	67.188	0.22412
4^{4}	65.365	0.41951
4^{5}	64.453	0.77028
4^{6}	63.932	1.2627
4^{7}	65.755	1.6428

For gamma = 4^2;

C-Value	Average Accuracy	Average Time
4^{-3}	53.125	0.15376
4^{-2}	53.125	0.15469
4^{-1}	62.37	0.15253
4^{0}	63.542	0.14972
4^{1}	65.885	0.15617
4^{2}	63.932	0.19865
4^{3}	63.151	0.25659
4^{4}	62.5	0.28378
4^{5}	62.63	0.28759
4^{6}	62.63	0.29117
4^{7}	62.63	0.2832

Based on these results the best CV value is given by using polynomial kernel with parameters; C = 4^7 and degree = 2. These values give 78.329 testing accuracy.

6 Bias/Variance Trade-off Programming

6.0.1 Dataset size = 10

Function	$Bias^2$	Variance
g_1	0.4864	0
g_2	0.33498	0.049629
g_3	0.27581	0.14005
g_4	0.040733	0.030575
g_5	0.040733	0.041986
g_6	0.051124	0.28389

2016/CSCI 567/HW3/libsvm-3.21/matlab/question6/dataset10.jpg

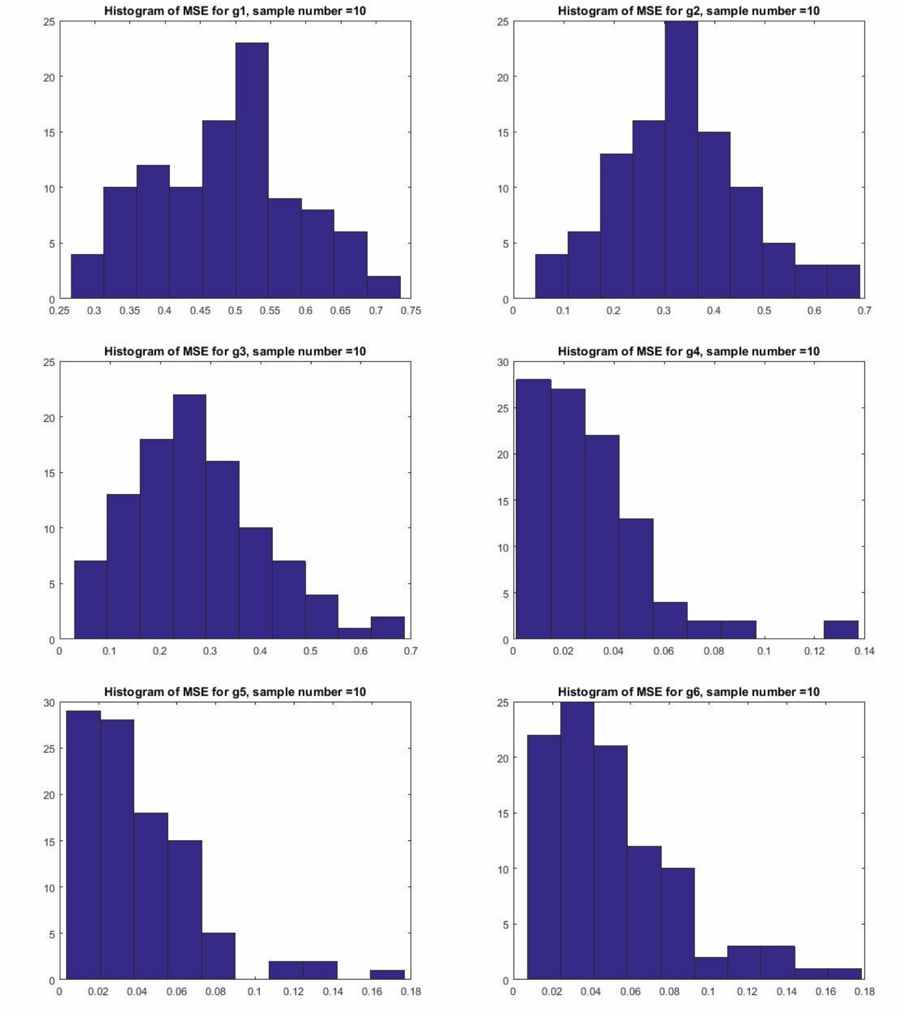

Figure 1: MSE Histograms for n = 10

6.0.2 Dataset size = 100

Function	$Bias^2$	Variance
g_1	0.46664	0
g_2	0.3533	0.0047887
g_3	0.34882	0.011171
g_4	0.0039551	0.0028753
g_5	0.0039551	0.0038822
g_6	0.0048682	0.0047759

2016/CSCI 567/HW3/libsvm-3.21/matlab/question6/dataset100.jpg

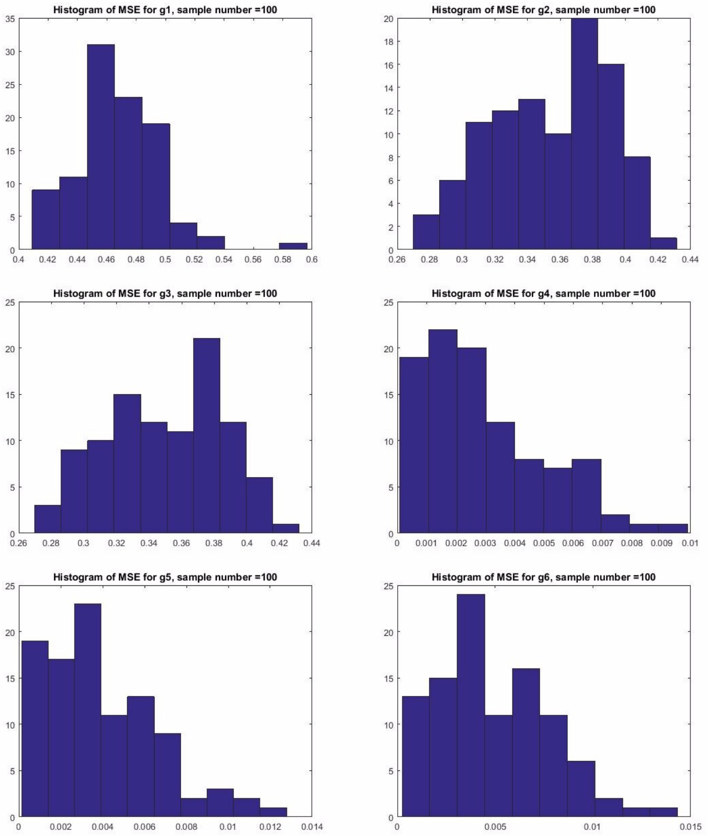

Figure 2: MSE Histograms for n = 100

6.0.3 Model Complexity and Sample Size Impact on Bias and Variance

Model Complexity: Given the results in (a) and (b) sections, we can clearly see that, as the complexity of the model increases, bias term decreases and when model becomes too complex , it does not have any more effect.

On the other hand, variance term increases until model 3 and after that, it fluctuates. It is expected to increase with the model complexity, but maybe because of randomness, it doesn't have a steady increase.

Dataset Size: Dataset size doesn't have a very positive effect for the first 3 models. But, when we use complex models, or the same model given (g4), a high number of data points help us to decrease bias significantly.

To reduce the variance, a high number of dataset is always useful and has a great positive impact on variance, which decreases it.

6.0.4 Regularized Bias and Variance

lambda	$Bias^2$	Variance
0.01	0.0032	0.0028
0.1	0.0029	0.0030
1	0.0062	0.0029
10	0.0916	0.0044

As we can see above, bias term increases as λ value increases. It is expected to have a lower variance with terms, which has a greater lambda. But, in this example variance term is already so low and because of that it doesn't change much.