HW1

Brian Allen,ba2542
Serena Zhang,mz2642
Haozheng Ni, hn2318

February 4, 2018

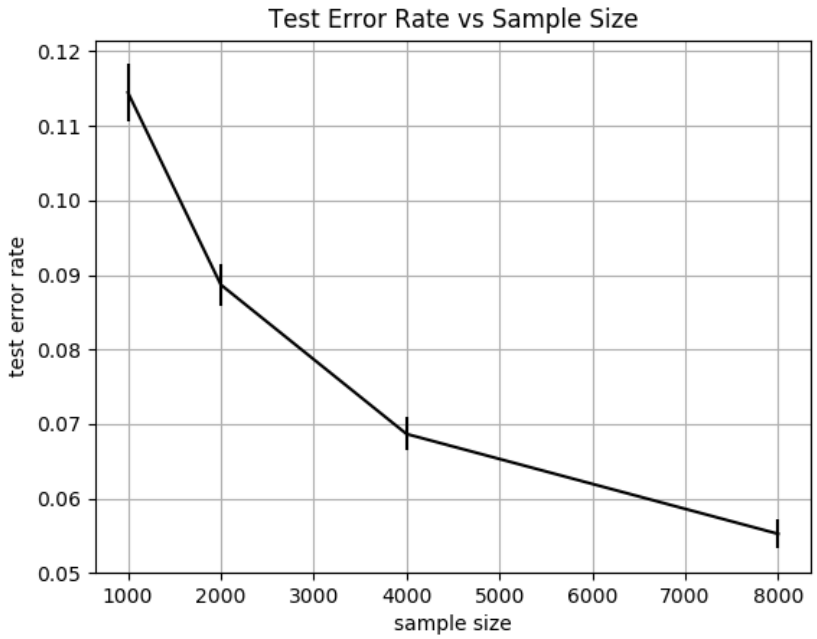
1. (a) To find the vector x_i that has the minimal Euclidean distance from x, it suffices to find x_i that has the minimal squared Euclidean distance. Therefore, for each row t_j in test data, we calculate its squared Euclidean distance to each row x_i in training data and find x_{i^*} that yields the smallest result. To achieve this, we leverage the given formula for the calculation:

   - For each t_j in test data, we get the last term of the formula $X * X.T$ by obtaining the diagnal of the dot product of X and $X.T$.
   - For the second term $-2T.T * X$, we simply take the dot product of X and $T.T$.
   - As the first term is the same for every t_j in the test data, we can ignore this term in our calculation.

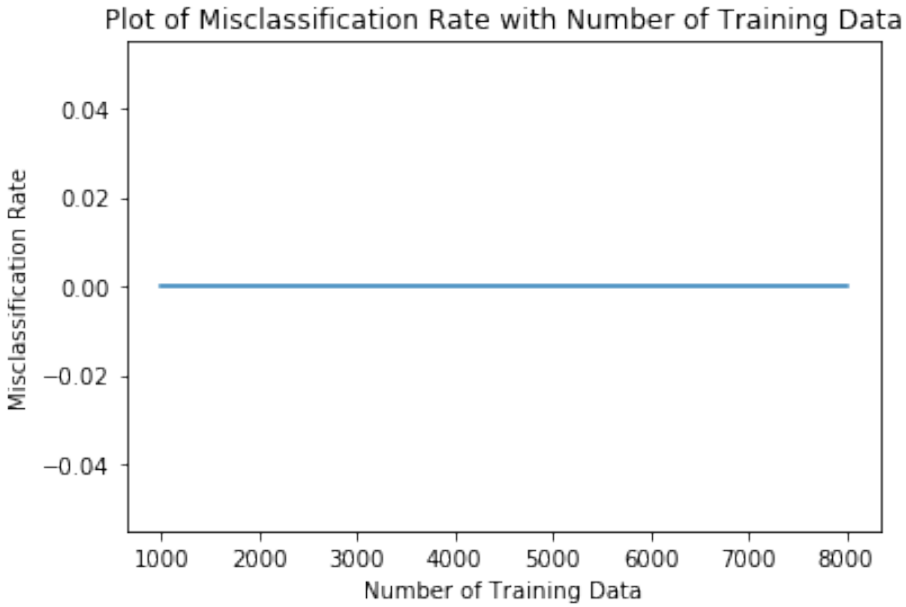
   In the end, for each t_j, we obtain an array of the squared Euclidean distance between t_j and each x_i, and we'd like to pick i^* which corresponds to the smallest element in this array.

   As a final step, we predict y_{i^*} for tj by using index i^*.

   (b)



   (c) It would be a horizontal line at test error rate = 0 as the data we use to evaluate is the same as the data we used to train.



2. (a) Test risks of the ordinary least squares estimator is 0.5398, and that of sparse linear predictor is 0.5565.

   (b)

variable name	coefficient
volatile acidity	-0.2447
sulphates	0.1085
alcohol	0.3736

   The coefficient $\hat{\beta}_0 = 5.8261$.

   (c)

Correlation			
variable with non-zero parameter	rank	variable	correlation
volatile acidity	1	total sulfur dioxide	-0.4220
	2	citric acid	-0.3756
sulphates	1	chlorides	0.4071
	2	fixed acidity	0.2969
alcohol	1	density	-0.6796
	2	residual sugar	-0.3463

3. (a) From the recurrence relation we have

$$\begin{aligned} \beta^{(k)} &= \beta^{(k-1)} + \eta A^T (b - A\beta^{(k-1)}) \\ &= (I - \eta M)\beta^{(k-1)} + \eta v \end{aligned} \tag{1}$$

   Next we would prove the statement by induction:

   - For k = 1, since $\beta^{(0)} = \mathbf{0}$, equ(1) above shows

$$\beta^{(1)} = (I - \eta M)\beta^{(0)} + \eta v = \eta v$$

     which is the same as the result implied by the statement.
   - Assume the statement holds true for some integer n, i.e

$$\beta^{(n)} = \eta \sum_{k=0}^{(n-1)} (I - \eta M)^k v$$

Let's consider N = n+1. Equation(1) above implies

$$\begin{aligned}
\beta^{(n+1)} &= (I - \eta M)\beta^{(n)} + \eta v \\
&= (I - \eta M)\eta \sum_{k=1}^{n-1} (I - \eta M)^k v + \eta v \\
&= \eta \sum_{k=0}^{n-1} (I - \eta M)^{k+1} v + \eta v \\
&= \eta \sum_{k=1}^{n} (I - \eta M)^k v + \eta v \\
&= \eta \sum_{k=1}^{n} (I - \eta M)^k v + \eta (I - \eta M)^0 v \\
&= \eta \sum_{k=0}^{n} (I - \eta M)^k v
\end{aligned}$$

which is the same as the result implied by the statement with $N = n + 1$. Therefore the statement holds true for any integer N by mathematical induction.

(b) Since M is a square matrix, we can apply spectrum decomposition as the following:

$$M = Q^T \Lambda Q$$

where Λ is the diagonal matrix of eigenvalues and the eigenvectors corresponding to each of them is represented by columns of Q. Notice that Q is orthonormal, i.e $Q^T Q = I$)
Substitute M into the target, we have

$$\begin{aligned}
\eta \sum_{k=0}^{N-1} (I - \eta M)^k &= \eta \sum_{k=0}^{N-1} (I - \eta Q^T \Lambda Q)^k \\
&= \eta \sum_{k=0}^{N-1} (Q^T (I - \eta \Lambda) Q)^k \\
&= \eta \sum_{k=0}^{N-1} Q^T (I - \eta \Lambda)^k Q \\
&= Q^T [\eta \sum_{k=0}^{N-1} (I - \eta \Lambda)^k] Q
\end{aligned} \tag{2}$$

Since Λ is a diagonal matrix with diagonal entries $\lambda_1, ..., \lambda_d$, we have

$$(I - \eta \Lambda)^k = \begin{pmatrix} (1 - \eta\lambda_1)^k & 0 & \dots & 0 \\ 0 & (1 - \eta\lambda_2)^k & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & (1 - \eta\lambda_d)^k \end{pmatrix}$$

Therefore

$$\eta \sum_{k=0}^{N-1} (I - \eta \Lambda)^k = \begin{pmatrix} \eta \sum_{k=0}^{N-1} (1 - \eta\lambda_1)^k & 0 & \dots & 0 \\ 0 & \eta \sum_{k=0}^{N-1} (1 - \eta\lambda_2)^k & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \eta \sum_{k=0}^{N-1} (1 - \eta\lambda_d)^k \end{pmatrix}$$

For $i = 1, ..., d$,

$$\eta \sum_{k=0}^{N-1} (1 - \eta\lambda_i)^k = \eta \frac{1 - (1 - \eta\lambda_i)^N}{1 - (1 - \eta\lambda_i)}$$
$$= \frac{1 - (1 - \eta\lambda_i)^N}{\lambda_i}$$

Since above Equation(2) again forms a spectral decomposition of $\eta \sum_{k=0}^{N-1} (I - \eta M)^k$, and thus it's eigenvalues are the diagonal entries of $\eta \sum_{k=0}^{N-1} (I - \eta\Lambda)^k$, which are

$$\tilde{\lambda_i} = \frac{1 - (1 - \eta\lambda_i)^N}{\lambda_i}$$

(c) Since the recurrence relation for $\beta^{(k)}$ is given by

$$\beta^{(k)} = \beta^{(k-1)} + \eta A^T (b - A\beta^{(k-1)})$$

substitute to $\|\hat{\beta} - \beta^{(N)}\|_2^2$, we have

$$\begin{aligned}
\|\hat{\beta} - \beta^{(N)}\|_2^2 &= \|\hat{\beta} - \beta^{(N-1)} - \eta A^T (b - A\beta^{(N-1)})\|_2^2 \\
&= \|\hat{\beta} - \beta^{(N-1)} - \eta (v - M\beta^{(N-1)})\|_2^2 \\
&= \|\hat{\beta} - \beta^{(N-1)} - \eta (M\hat{\beta} - M\beta^{(N-1)})\|_2^2 \\
&= \|\hat{\beta} - \beta^{(N-1)} - \eta M(\hat{\beta} - \beta^{(N-1)})\|_2^2 \\
&= \|(I - \eta M)(\hat{\beta} - \beta^{(N-1)})\|_2^2
\end{aligned}$$

If we keep substituting the recurrence relation of β^{N-1}, we can end with

$$\begin{aligned}
\|\hat{\beta} - \beta^{(N)}\|_2^2 &= \|(I - \eta M)^N (\hat{\beta} - \beta^{(0)}\|_2^2 \\
&= \|(I - \eta M)^N \hat{\beta}\|_2^2 \\
&= ((I - \eta M)^N \hat{\beta})^T ((I - \eta M)^N \hat{\beta}) \\
&= \hat{\beta}^T (I - \eta M)^{2N} \hat{\beta}
\end{aligned}$$

The last equation holds because $(I - \eta M)$ is a symmetric matrix.
If we apply spectral decomposition of M, we can obtain

$$\begin{aligned}
\|\hat{\beta} - \beta^{(N)}\|_2^2 &= \hat{\beta}^T (I - \eta Q^T \Lambda Q)^{2N} \hat{\beta} \\
&= \hat{\beta}^T Q^T (I - \eta\Lambda)^{2N} Q\hat{\beta}
\end{aligned}$$

Hence we have

$$\begin{aligned}
\frac{\|\hat{\beta} - \beta^{(N)}\|_2^2}{\|\hat{\beta}\|_2^2} &= \frac{\hat{\beta}^T Q^T (I - \eta\Lambda)^{2N} Q\hat{\beta}}{\hat{\beta}^T \hat{\beta}} \\
&= \frac{S^T (I - \eta\Lambda)^{2N} S}{S^T S}
\end{aligned}$$

where $S \coloneqq Q\hat{\beta}$. Denote each element as $s_i, i = 1, .., d$
Since $(I - \eta\Lambda)^{2N}$ is a diagonal matrix, denote the diagonal entries as $\tilde{\lambda}_i, i = 1, .., d$.
Then above quotient can be written as

$$\begin{aligned}
\frac{\|\hat{\beta} - \beta^{(N)}\|_2^2}{\|\hat{\beta}\|_2^2} &= \frac{S^T \begin{pmatrix} \tilde{\lambda}_1 & 0 & \dots & 0 \\ 0 & \tilde{\lambda}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \end{pmatrix} S}{S^T S} \\
&= \frac{\sum_{i=1}^d \tilde{\lambda}_i s_i^2}{\sum_{i=1}^d s_i^2} \\
&\leq \tilde{\lambda}_{max}
\end{aligned} \tag{3}$$

where $\tilde{\lambda}_{max}$ is the maximum eigenvalue of $(I - \eta\Lambda)^{2N}$.
Since $\tilde{\lambda}_i = (1 - \eta\lambda_i)^{2N}$, we have

$$\tilde{\lambda}_{max} = (1 - \eta\lambda_{min})^{2N}$$

where λ_{min} is the minimum eigenvalue of M. Then we substitute the result in equation (3) above:

$$\begin{aligned}\frac{\|\hat{\beta} - \beta^{(N)}\|_2^2}{\|\hat{\beta}\|_2^2} &\leq \tilde{\lambda}_{max} \\ &\leq (1 - \eta\lambda_{min})^{2N} \\ &\leq e^{-2N\eta\lambda_{min}}\end{aligned}$$

Therefore we have proved the statement.
Comment: The question states that λ_{min} should be non-zero. However, this condition seems to be too strong, and could make the statement invalid. Next we will present a counter-example. All the norms below are L_2 norm for vector. Suppose the positive semi-definite matrix M being

$$M = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Then the eigenvalues of M are 0, 2.
Suppose $\hat{\beta} = \begin{pmatrix} 2 & 2 \end{pmatrix}^T, \eta = \frac{1}{3}$ (This is a valid assumption as $\lambda_i < \frac{1}{\eta}$), and we just try $N = 3$. Then we have

$$\begin{aligned}\|\hat{\beta} - \beta^{(3)}\| &= \|(I - \eta M)^3\hat{\beta}\| \\ &= 0.1048\end{aligned}$$

While the right hand side of the inequality is

$$e^{-2N\eta\lambda_{min}}\|\hat{\beta}\| = e^{-2*3*\frac{1}{3}*2}\|\hat{\beta}\| = e^{-4}\|\hat{\beta}\| = 0.0518$$

Therefore we have

$$\|\hat{\beta} - \beta^{(3)}\| > e^{-2*3\eta\lambda_{min}}\|\hat{\beta}\|$$

which contradicts the statement.

4. (a)

$$\begin{aligned}\int_0^\infty x^2 e^{-x/\theta} dx &= \theta \int_0^\infty x^2 \frac{1}{\theta} e^{-x/\theta} dx \\ &= \theta E(Y^2) \text{ for } Y \sim Exp(\frac{1}{\theta}) \\ &= \theta(\theta^2 + \theta^2) \\ &= 2\theta^3\end{aligned} \tag{4}$$

Therefore the pdf of X is

$$f_\theta(x) = \frac{1}{2\theta^3} x^2 e^{-x/\theta}, x > 0$$

Then we have the likelihood function $L(\theta)$ and log likelihood function $l(\theta)$ as

$$L(\theta) = \prod_{i=1}^n f_\theta(x_i) = 2^{-n}\theta^{-3n}(\prod_{i=1}^n x_i)^2 e^{\sum_{i=1}^n x_i/\theta}$$

$$l(\theta) = -n\log 2 - 3n\log\theta + 2\sum_{i=1}^{n}\log x_i - \frac{\sum_{i=1}^{n} x_i}{\theta}$$

Taking derivative of $l(\theta)$ and set it to be 0, we have

$$0 = \frac{-3n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2}$$

which implies

$$\hat{\theta}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{3n}$$

(b)

$$\int_0^\infty e^{-x/\theta}dx = \theta$$

Therefore the pdf of X is

$$f_\theta(x) = \frac{1}{\theta}e^{-x/\theta}, x \geq 0$$

Then we have the likelihood function $L(\theta)$ and log likelihood function $l(\theta)$ as

$$L(\theta) = \theta^{-n}e^{-\sum_{i=1}^{n} x_i/\theta}$$

$$l(\theta) = -n\log\theta - \frac{\sum_{i=1}^{n}}{\theta}$$

Taking derivative of $l(\theta)$ and set it to be 0, we have

$$0 = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2}$$

which implies

$$\hat{\theta}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(c)

$$\int_0^\theta 1dx = \theta$$

Therefore the pdf of X is

$$f_\theta(x) = \frac{1}{\theta}, x \in [0, \theta]$$

Then we have the likelihood function $L(\theta)$ as

$$\begin{aligned} L(\theta) &= \frac{1}{\theta^n}\mathbf{1}(x_i \in [0,\theta]) \\ &= \frac{1}{\theta^n}\mathbf{1}(\theta \geq x_{max}) \end{aligned}$$

where $\mathbf{1}(.)$ is an indicator function, and x_{max} is the largest number in the sample.
Clearly the likelihood function is monotonically decreasing with θ, and the minimum θ will maximize the function value. Since $\theta \geq x_{max}$, the possible smallest value of θ is x_{max}, and thus

$$\hat{\theta}_{MLE} = x_{max}$$