CSCI-SHU 360 Machine Learning
Solution to homework 2

Josiah Li yl11912@nyu.edu

March 12, 2025

1 Linear Regression and Convexity

We have:

$$\begin{aligned} L(w) &= (y - Xw)^T (y - Xw) \\ &= y^T y - w^T X^T y - y^T Xw + w^T X^T Xw. \end{aligned}$$

We take the directional derivative on direction v:

$$\begin{aligned} D_v L(w) &= \lim_{h \to 0} \frac{L(w + hv) - L(w)}{h} \\ &= \lim_{h \to 0} \frac{y^T y - (w + hv)^T X^T y - y^T X(w + hv)^T X^T X(w + hv) - y^T y + w^T X^T y + y^T Xw - w^T X^T Xw}{h} \\ &= -v^T X^T y - y^T Xv + v^T X^T Xw + W^T X^T Xv \\ &= v^T \cdot (2X^T Xw - 2X^T y) \\ &=< v, 2X^T Xw - 2X^T y > . \end{aligned}$$

Thus, the derivative is $\nabla_w L(w) = 2X^T Xw - 2X^T y$.
The second order derivative of $L(w)$ is $(\nabla_w)^2 L(w) = 2X^T X \geq 0$.
Thus, the loss function is convex as the second order derivative of the function is PSD.

2 Gaussian Distribution and the Curse of Dimensionality

2.1

For $m = 2$, we have:

$$S_1(r) = 2\pi r$$
$$V_2(r) = \pi r^2.$$

And for $m = 3$, we have:

$$S_2(r) = 4\pi r^2$$
$$V_3(r) = \frac{4}{3}\pi r^3$$

2.2

When $m \in 2, 3$, this equation holds: $2\pi r = \frac{d}{dr}(\pi r^2)$, $4\pi r^2 = \frac{4}{3}\pi r^3$.
Intuitively, we can consider the surface area to be $\lim_{\Delta r \to 0} \frac{V_m(r + \Delta r) - V_m(r)}{\Delta r}$. $V_m(r + \Delta r) - V_m(r)$ can be considered as a shell outside the origin volume with the shell thickness be Δr, thus, when the $\Delta r \to 0$, the shell goes to the surface of the sphere.

2.3

As $\bar{S}_{m-1}$ contains the dimensional information about the sphere (per say, the constant value). What we need to care is the influence of r and m. Thus $S_{m-1}(r) = \bar{S}_{m-1} r^{m-1}$

2.4

We have:

$$p(x) = \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right).$$

Then we try to do the integration:

$$\rho_m(r) = \int p(x)dx = \int \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right) dx.$$

As we have $\|x\| = r$, we can get:

$$\begin{aligned}
\rho_m(r) = \int p(x)dx &= \int \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right) dx \\
&= \int \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) dx \\
&= \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) r^{m-1} \bar{S}_{m-1}
\end{aligned}$$

2.5

When m is set, ρ_m only depend on r^{m-1} and $\exp\left(-\frac{r^2}{2\sigma^2}\right)$

Thus we can let $f(r) = r^{m-1}\exp(-\frac{r^2}{2\sigma^2})$, and we have:

$$\begin{aligned} f'(r) &= (m-1)r^{m-2}\exp(-\frac{r^2}{2\sigma^2}) - r^m\exp(-\frac{r^2}{2\sigma^2})\frac{1}{\sigma^2} \\ &= (m-1-\frac{r^2}{\sigma^2})r^{m-2}\exp(-\frac{r^2}{2\sigma^2}) \end{aligned}$$

Thus, when $m-1-\frac{r^2}{\sigma^2} = 0$, the function has the extreme point.
Which is $r = \sqrt{(m-1)\sigma^2}$, when $r \leq \sqrt{(m-1)\sigma^2}$, $f'(r)$ increases, thus, the function has a single maximum value at $r = \sqrt{(m-1)\sigma^2}$, when $m \to \infty$, $m-1 \approx m$, which means for large m, $\rho_m(r)$ has a single maximum value at $\hat{r}$ such that $\hat{r} \approx \sqrt{m}\sigma$.

2.6

In order to derive the the form, we calculate $\frac{\rho_m(\hat{r}+\epsilon)}{\rho_m(\hat{r})}$, which is:

$$\begin{aligned} \frac{\rho_m(\hat{r}+\epsilon)}{\rho_m(\hat{r})} &= \frac{\frac{1}{(2\pi\sigma^2)^{m/2}}\exp\left(-\frac{(\hat{r}+\epsilon)^2}{2\sigma^2}\right)(\hat{r}+\epsilon)^{m-1}\bar{S}_{m-1}}{\frac{1}{(2\pi\sigma^2)^{m/2}}\exp\left(-\frac{(\hat{r})^2}{2\sigma^2}\right)(\hat{r})^{m-1}\bar{S}_{m-1}} \\ &= \left(1+\frac{\epsilon}{\hat{r}}\right)^{m-1} e^{-\frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2}} \end{aligned}$$

In order to get the form we desired, we construct $\left(1+\frac{\epsilon}{\hat{r}}\right)^{m-1} = \exp((m-1)\ln(1+\frac{\epsilon}{\hat{r}}))$.
As we proved in last question, we have $m = \frac{\hat{r}^2}{\sigma^2}$. Thus, when for large m, we have $m-1 \approx \frac{\hat{r}^2}{\sigma^2}$
Thus, we have:

$$\frac{\rho_m(\hat{r}+\epsilon)}{\rho_m(\hat{r})} = \exp\left(\frac{\hat{r}^2}{\sigma^2}\ln(1+\frac{\epsilon}{\hat{r}})\right)\exp\left(-\frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2}\right).$$

By Taylor Expansion, we have $\ln(1+\frac{\epsilon}{\hat{r}}) \approx \frac{\epsilon}{\hat{r}} - \frac{\epsilon^2}{2\hat{r}^2}$
Thus, we can have:

$$\begin{aligned} \frac{\rho_m(\hat{r}+\epsilon)}{\rho_m(\hat{r})} &\approx \exp\left(\frac{\hat{r}\epsilon}{\sigma^2} - \frac{\epsilon^2}{2\sigma^2}\right)\exp\left(-\frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{\epsilon^2}{\sigma^2}\right). \end{aligned}$$

Thus, we have $\rho(\hat{r}+\epsilon) \approx \rho(\hat{r})e^{-\frac{\epsilon^2}{\sigma^2}}$.

2.7

We have $\hat{r} = \sqrt{m}\sigma \gg \sigma$ for large m. Thus, most of the sample points reside outside the high dimension sphere with radius σ. They are reside on the sphere with raidus $\hat{r} \approx \sqrt{m}\sigma$. As for Low dimensional Gaussian, we have $\hat{r} = \sqrt{m-1}\sigma$. When m is small, we can see that most of the points reside inside the sphere with radius σ.

2.8

The probability density at the origin will be $\frac{1}{(2\pi\sigma^2)^{m/2}}$, and for the point on sphere $S_{m-1}(\hat{r})$, we have $p = \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left(-\frac{\hat{r}^2}{2\sigma^2}\right)$. As this seems to say that the probability on the sphere will be smaller that the origin, but we have to notice that, as r increases, the volume of the shell will also increase, lead to a balance at $\hat{r}$.

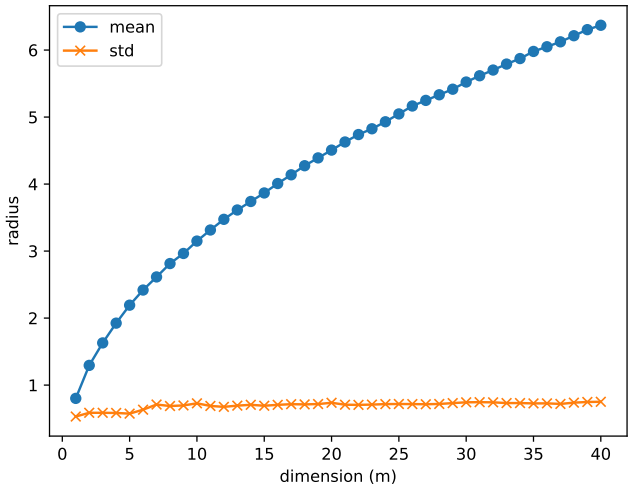

Figure 1: Means and std of radius from m-dimensional Gaussian

3 Ridge Regression

3.1

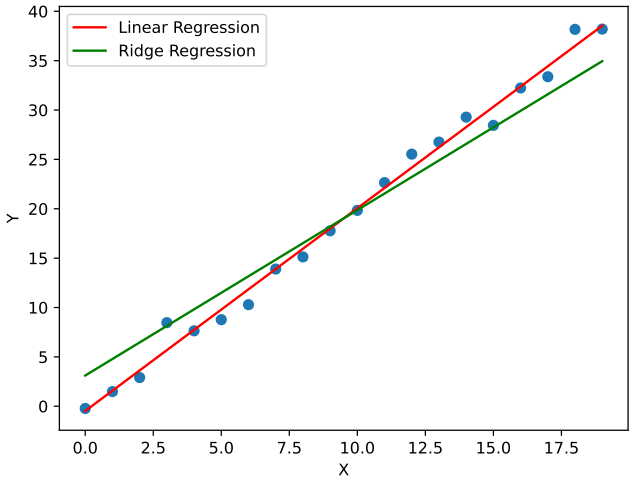

Figure 2: Linear regression and ridge regression on datapoints (X, y)

As we can see, in figure 2, standard linear regression has better performance. This is because the linear relation for the datapoints is strong and there are no outliers in this dataset.

3.2

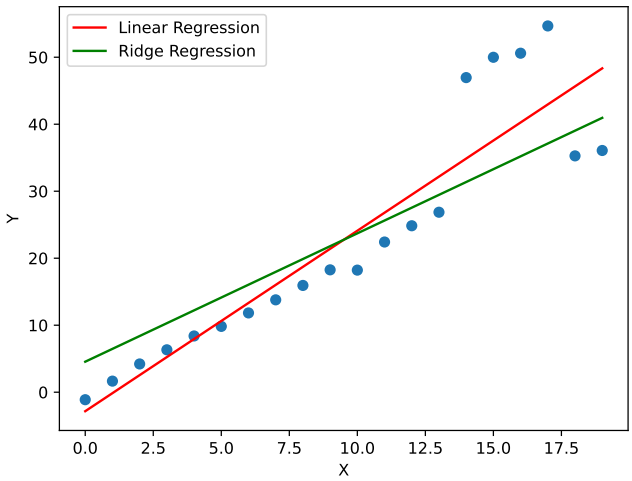

Figure 3: Regressions on dataset with outliers

3.3

By taking the derivative of the original target, we have:

$$D_w F(w) = 2X^T Xw - 2y^T X + \eta w$$

We have the equation:

$$\begin{aligned} 2X^T Xw - 2y^T X + \eta w &= 0 \\ (2X^T X + \eta)w &= 2y^T X \\ w &= (X^T X + \frac{\eta}{2} I)^{-1} X^T y \end{aligned}$$

3.4

(a) In this case, we could not compute the closed-form solution. As when we have a $n \times p$ matrix and $p > n$, we can see that after the multiplication we got a $p \times p$ matrix. The original matrix can have a maximum rank n, leads to a maximum rank n for $X^T X$, thus, the new matrix can not be full rank. Which means the new matrix is not invertible.

(b) As we got in the previous question, we have the closed form solution of the ridge regression as:

$$w = (X^T X + \frac{\eta}{2} I)^{-1} X^T y.$$

We can notice that, no matter matrix $X^T X$ is invertible or not, it is always a PSD matrix. Moreover, by the definition of eigenvalue, we have:

$$\begin{aligned} X^T X v &= \lambda v \\ (X^T X + I)v &= (\lambda + 1)v \end{aligned}$$

Thus, as the origin matrix is PSD ($\lambda \geq 0$), we have the eigenvalues for the matrix $X^T X + I$ always larger than 0. And this ensures that the matrix $X^T X + I$ is a positive definite matrix. Which means this matrix is always invertible. And we can calculate the closed-form solution solution for ridge regression.

4 Programming Problem: Draw an Ellipsoid

4.1

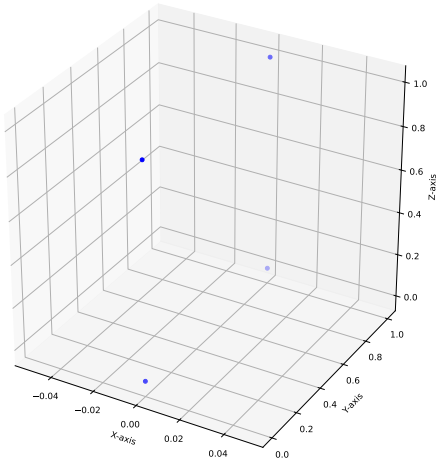

Figure 4: 3D point cloud

4.2

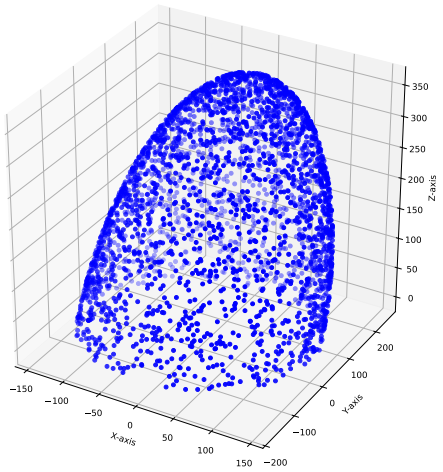

Figure 5: Half ellipsoid

5 Programming Problem: Linear Regression

5.1

By observing the output scatter plots, we pick these three relations as follows.

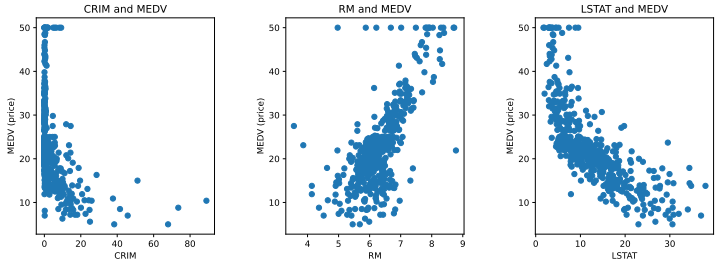

Figure 6: 2D scatter plot for each feature

5.2

By the heatmap below, we can see that the top-3 features that are mostly linearly related to the "MEDV" is "LSTAT", "PTRATIO", and "RM". This is different as the ones in previous question.

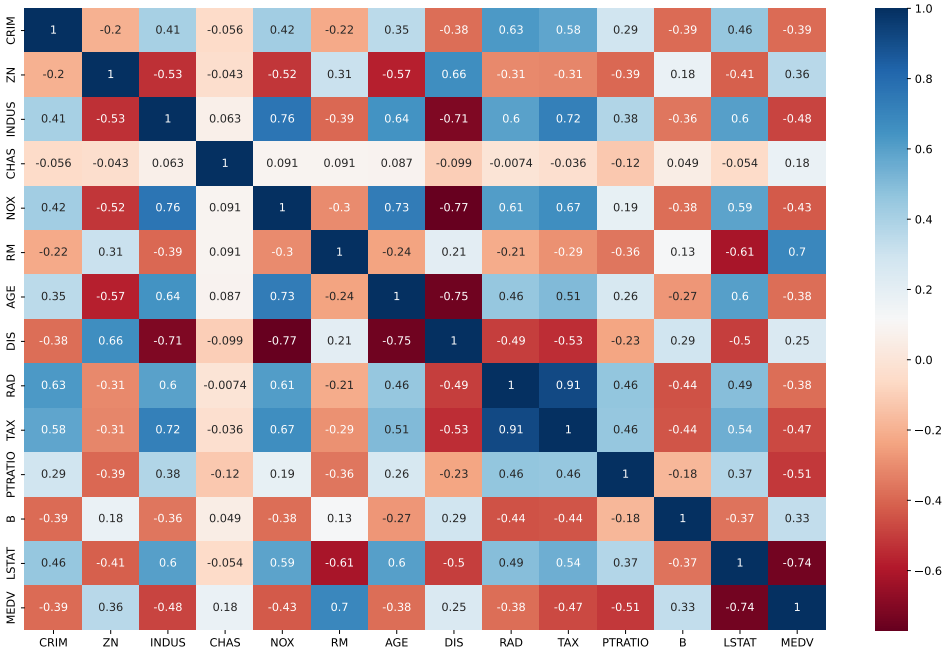

Figure 7: The heatmap of correlation matrix

5.3

Feature	Coeff	Coeff_eta_15.0	Coeff_eta_30.0	Coeff_eta_45.0
CRIM	-0.099324	-0.100648	-0.101157	-0.101396
ZN	0.052251	0.054632	0.056906	0.059028
INDUS	0.004516	0.012958	0.016161	0.018062
CHAS	2.957261	2.272783	1.854801	1.575958
NOX	1.127938	0.457674	0.380824	0.343826
RM	5.854198	5.728152	5.574767	5.424074
AGE	-0.014957	-0.010094	-0.006243	-0.002772
DIS	-0.920844	-0.896985	-0.869514	-0.842988
RAD	0.159519	0.163084	0.164159	0.164232
TAX	-0.008934	-0.008982	-0.008988	-0.008940
PTRATIO	-0.435674	-0.406149	-0.375560	-0.345226
B	0.014905	0.015518	0.015989	0.016406
LSTAT	-0.474751	-0.484274	-0.495622	-0.506287

Table 1: The coeffcients correspondig to OLR and ridge regression

5.4

Model	RMSE_test	RMSE_train
OLR	5.209218	4.820627
Ridge_eta_15	5.191204	4.826364
Ridge_eta_30	5.187847	4.837923
Ridge_eta_45	5.189540	4.852552

Table 2: The RMSE on test and train dataset

From the table, we can see that as η increases, the RMSE on training dataset gets larger (we consider the ordinary linear regression as $\eta = 0$). However, we can notice that, on the test dataset, ridge regression do decreases the RMSE for proper eta. This is because ridge regression can prevent the model from overfitting to the training dataset, lead to a better performance on the test dataset.

5.5

Model	RMSE_test	RMSE_train
Linear Regression	5.494724	5.273362
Ridge Regression	5.481155	5.275046

Table 3: The OLR and ridge regression by only using the top-3 features

From this table we can notice that although we are just using the top-3 features, the RMSE is quite close to the RMSE which applying all features.

6 Bonus: Locality Sensitive Hashing

6.1

In order to find exactly the nearest neighbor point in X, we set $c = 1$, which means the oracle will degraded into a simple oracle that tells us whether there are any point satisfying $d(x, q) \leq r$. Thus, we can apply binary search, which is: every time we update the upper bound and lower bound of the search space by using the returned value of the oracle, and each time we set r as the middle point of the searching space. In this approach, we can have a time complexity at $O(\log n)$.

6.2

In this case, for given x_i and x_j, we have $d(x_i, x_j) \leq r$. And this indicates that x_i and x_j have at most r different digits and at least 0 different digits. As the function h mainly just pick one random digit from the number, we can see that for every two x, there are $\frac{m-r}{m}$ chances that they are the same value on the corresponding coordinate. Thus, $p_1 = 1 - \frac{r}{m}$.
When $d(x_i, x_j) \geq cr$, by following the same approach, we can see that there are at least cr digits that are not the same, thus, there are $m - cr$ digits that are the same on the same coordinate. From this, we can deduce that in this case, there are $\frac{m-cr}{m}$ chances that they are the same on the same coordinate. Which means $p_2 = 1 - \frac{cr}{m}$

6.3

In this occasion, we can see that $g(x_i)$ can be considered as a intersection of k i.i.d. random variable. When $d(x_i, x_j) \leq r$, $p_1 = 1 - \frac{r}{m}$. When $d(x_i, x_j) \geq cr$, $p_2 = 1 - \frac{cr}{m}$. And when the function $g(x_i)$ and $g(x_j)$ have the same value, all the functions $h_1(x_i), h_2(x_i) \cdots h_k(x_i)$ have the same values to $h_1(x_j), h_2(x_j) \cdots h_k(x_j)$ correspondingly. Which means, when $d(x_i, x_j) \leq r$, $Pr(g(x_i) = g(x_j)) \geq p_1^k$. And when $d(x_i, x_j) \geq cr$, $Pr(g(x_i) = g(x_j)) \leq p_2^k$

6.4

In this case, all the function g are i.i.d., and here we only need there to be at least one b to satisfy the requirement,thus, we have $Pr(\exists b, g_b(x_i) = g_b(x_j)) \geq 1 - (1 - p_1^k)^l$ when $d(x_i, x_j) \leq r$. And we have $Pr(\exists b, g_b(x_i) = g_b(x_j)) \leq 1 - (1 - p_2^k)^l$ when $d(x_i, x_j) \geq cr$.

6.5

(a) According to problem 6.4, as $d(x', q) \leq r$, we have $Pr(\exists b, g_b(x') = g_b(q)) \geq 1 - (1 - p_1^{\frac{\ln(n)}{\ln(1/p_2)}})^{n^{\frac{\ln(p_1)}{\ln(p_2)}}}$

In order to prove the inequality we desired, we only need to prove $(1 - p_1^{\frac{\ln(n)}{\ln(1/p_2)}})^{n^{\frac{\ln(p_1)}{\ln(p_2)}}} \leq e^{-1}$. As we have $p_1^{\frac{\ln(n)}{\ln(1/p_2)}} = p_1^{-\frac{\ln(n)}{\ln(p_2)}} = e^{-\frac{\ln(p_1)\ln(n)}{\ln(p_2)}} = (\frac{1}{n})^{\frac{\ln(p_1)}{\ln(p_2)}}$. We can see that the original inequality is the same as:

$$\left(1 - \left(\frac{1}{n}\right)^{\frac{\ln(p_1)}{\ln(p_2)}}\right)^{n^{\frac{\ln(p_1)}{\ln(p_2)}}} \leq e^{-1}$$

And by the inequality provided, we set $k = n^{\frac{\ln(p_1)}{\ln(p_2)}}$, we can see that the inequality we desired is proved.

(b) According to Markov's inequality, we have $P(X \geq a) \leq \frac{E(X)}{a}$, which means for the event, we have $P(X \geq 4l) \leq \frac{E(X)}{4l}$. As $d(x, q) \geq cr$, we have $E(X) = n \cdot Pr(\exists b, g_b(x') = g_b(q)) \leq n \cdot (1 - (1 -$

$p_2^{\frac{\ln(n)}{\ln(1/p_2)}})^l = n \cdot (1 - (1 - \frac{1}{n})^l) \leq n \cdot (1 - (1 - \frac{l}{n})) = l$. Thus, $\frac{E(X)}{4l} \leq \frac{l}{4l} = \frac{1}{4}$. Thus, we have $P(X \leq 4l) = 1 - P(X \geq 4l) \geq 1 - 1/4 = \frac{3}{4}$. Which means the second event happens with probability at least $\frac{3}{4}$.

(c) By the inequality we proved, we can see that we have the lower bound for each two events. As the two events are independent, we can see that the lower bound that both event happens is $\frac{3}{4}(1 - e^{-1})$.