CSCI-SHU 360 Machine Learning
Solution to homework 2

Yufeng Xu yx3038@nyu.edu

March 2, 2024

1 Linear Regression and Convexity

The loss function of linear regression is

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

hence

$$\begin{aligned} &D_v L(w) = \lim_{h\to 0} \frac{L(w + hv) - L(w)}{h} \\ &= \lim_{h\to 0} \frac{(w + hv)^T X^T X(w + hv) - (w + hv)^T X^T y - y^T X(w + hv) + y^T y - w^T X^T Xw + w^T X^T y + y^T Xw - y^T y}{h} \\ &= \lim_{h\to 0} \frac{(hv)^T X^T Xw + w^T X^T Xhv + (hv)^T X^T Xhv - (hv)^T X^T y - y^T Xhv}{h} \\ &= v^T X^T Xw + w^T X^T Xv - v^T X^T y - y^T Xv + \lim_{h\to 0} hv^T X^T Xv = \nabla_w L(w) \cdot v = (\nabla_w L(w))^T v \end{aligned}$$

Therefore, $\nabla_w L(w) = 2X^T Xw - 2X^T y$, $\nabla_w^2 L(w) = \nabla_w(\nabla_w L(w)) = 2X^T X \geq 0$. Hence, $L(w) = ||y - Xw||_2^2$ is a convex function.

2 Gaussian Distribution and the Curse of Dimensionality

2.1

$S_{2-1}(r) = 2\pi r, V_2(r) = \pi r^2, S_{3-1}(r) = 4\pi r^2, V_3(r) = \frac{4}{3}\pi r^3$

2.2

The equation $S_{m-1}(r) = \frac{d}{dr} V_m(r)$ works for $m \in \{2, 3\}$, as $\frac{d}{dr} V_2(r) = \frac{d}{dr}\pi r^2 = 2\pi r = S_{2-1}(r)$, $\frac{d}{dr} V_3(r) = \frac{d}{dr}\frac{4}{3}\pi r^3 = 4\pi r^2 = S_{3-1}(r)$.
Intuitively, this equation should hold for $\forall m \in \mathbb{N}, n \geq 2$. Consider $V_m(r+\Delta r) - V_m(r)$, which is equivalent to the volume of an m-d spherical shell that is outside of the sphere with radius r and inside of the sphere with radius $r+\Delta r$. When $\Delta r \to 0$, the shell can be approximated by a plate with base area $S_{m-1}(r)$ and thickness Δr, i.e., $V_m(r + \Delta r) - V_m(r) \to S_{m-1}(r)\Delta r$ when $r \to 0$, therefore $S_{m-1}(r) = \lim_{r\to 0} \frac{V_m(r+\Delta r) - V_m(r)}{\Delta r} = \frac{d}{dr} V_m(r)$.

2.3

We know that $V_m(r)$ is only dependent on r^m, in other words, $V_m(r) = \frac{r^m}{1^m} V_m(1) = r^m V_m(1)$. We also know from 2.2 that $S_{m-1}(r) = \frac{d}{dr} V_m(r) = \frac{d}{dr}(r^m V_m(1)) = r^m \frac{d}{dr} V_m(1) + mr^{m-1} V_m(1) = mr^{m-1} V m(1)$.
When $r = 1$, $\bar{S}_{m-1} = S_{m-1}(r) = mV_m(1)$, hence $S_{m-1}(r) = r^{m-1}(mV_m(1)) = r^{m-1}\bar{S}_{m-1}$.

2.4

Because $||x||_2 = r$,

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

2.5

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

Let $\frac{d}{dr}\rho_m(r) = 0$, $r = \sqrt{m-1}\sigma = \hat{r}$. When $r < \hat{r}$, $\frac{d}{dr}\rho_m(r) > 0$, $\rho_m(r) \nearrow$ as $r \nearrow$; when $r > \hat{r}$, $\frac{d}{dr}\rho_m(r) < 0$, $\rho_m(r) \searrow$ as $r \nearrow$. Therefore, $\rho_m(r)$ is maximal if and only if $r = \hat{r}$.
On the other hand, when $m \to \infty$, $\sqrt{m-1} \to \sqrt{m}$, hence $\hat{r} = \sqrt{m-1}\sigma \to \sqrt{m}\sigma$.

2.6

We know $\frac{\rho_m(\hat{r}+\epsilon)}{\rho_m(\hat{r})} = \frac{\exp(-\frac{(\hat{r}+\epsilon)^2}{2\sigma^2})}{\exp(-\frac{\hat{r}^2}{2\sigma^2})} \cdot \frac{(\hat{r}+\epsilon)^{m-1}}{\hat{r}^{m-1}} = \exp(-\frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2}) \cdot (1+\frac{\epsilon}{\hat{r}})^{m-1}$, where $m \to \infty$, hence $(1+\frac{\epsilon}{\hat{r}})^{m-1} \to (1+\frac{\epsilon}{\sqrt{m}\sigma})^m = \left((1+\frac{\epsilon}{\sqrt{m}\sigma})^{\sqrt{m}}\right)^{\sqrt{m}}$, where $(1+\frac{\epsilon}{\sqrt{m}\sigma})^{\sqrt{m}} = \sum_{n=1}^{\sqrt{m}} \binom{\sqrt{m}}{n} \left(\frac{\epsilon}{\sqrt{m}\sigma}\right)^n \to \sum_{n=1}^{\sqrt{m}} \frac{(\sqrt{m})^n}{n!} \left(\frac{\epsilon}{\sqrt{m}\sigma}\right)^n = \sum_{n=1}^{\sqrt{m}} \frac{1}{n!} \left(\frac{\epsilon}{\sigma}\right)^n \to \exp(\frac{\epsilon}{\sigma})$ as $m \to \infty$.
Therefore, $\frac{\rho_m(\hat{r}+\epsilon)}{\rho_m(\hat{r})} \to \exp(-\frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2}) \cdot \left(\exp(\frac{\epsilon}{\sigma})\right)^{\sqrt{m}} \to \exp(-\frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2}) \cdot \left(\exp(\frac{\epsilon}{\sigma})\right)^{\frac{\hat{r}}{\sigma}} = \exp(-\frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2} + \frac{\epsilon\hat{r}}{\sigma^2}) = \exp(-\frac{\epsilon^2}{2\sigma^2})$, hence $\rho(\hat{r}+\epsilon) \approx \rho(\hat{r}) \exp(-\frac{\epsilon^2}{2\sigma^2})$

2.7

As we learned in 2.5, when we are sampling from a high-dimensional Gaussian distribution, i.e., m is large enough, $\rho_m(r)$ is maximal when $r = \hat{r} \approx \sqrt{m}\sigma > \sigma$, hence most of the sampled points reside out of the σ neighborhood, at radius $\hat{r} \approx \sqrt{m}\sigma$.
When we sample from a low-dimensional Gaussian distribution, $\rho_m(r)$ is maximal when $r = \sqrt{m-1}\sigma$. When $m \in \{1, 2\}$, $0 \leq \sqrt{m-1}\sigma \leq \sigma$, hence most of the sampled points reside within the σ neighborhood.

2.8

When x is at the origin, $p_0(x) = \frac{1}{(2\pi\sigma)^{m/2}}$; when x is on the sphere of radius $\hat{r} = \sqrt{m}\sigma$, $p_{\hat{r}}(x) = \frac{1}{(2\pi\sigma)^{m/2}} \exp(-\frac{m}{2}) < \frac{1}{(2\pi\sigma)^{m/2}} = p_0(x)$. The probability density at $||x||_2 = \hat{r}$ is much smaller than that of $||x||_2 = 0$. However, $\rho_m(\hat{r}) > \rho_m(0)$ because as $r \nearrow$, $S_{m-1}(r)$ grows much faster than $p_r(x)$ decreases.
To verify my conjecture, I sampled 100 points from Gaussian distributions $N_m(0, 1)$ where $m = 1, 2 \ldots 40$, calculating the means and standard deviations in each group and plotted the two metrics as functions of m. The results are as follows:
From Figure 1, we observe that $Avg\{||x||_2\} \propto \sqrt{m}$ (actually, $Avg\{||x||_2\} \approx \sqrt{m}$), which is consistent with our conjecture that most of the sampled points reside around the radii $\sqrt{m}\sigma$ for any m.

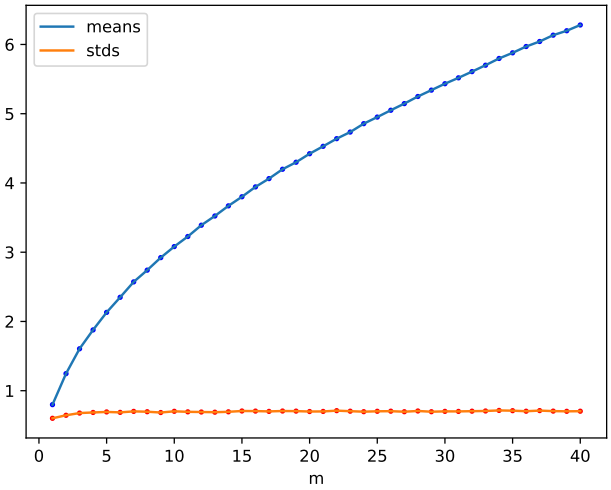

Figure 1: means and standard deviations of 100 sampled points from Gaussian distributions of dimensions 1-40.

The standard deviation of the norms does not change significantly as m increases, which implies that the standard deviation is likely to be independent of the dimension of the distribution.

3 Ridge Regression

3.1

When (X, y) are strongly linearly correlated, standard linear regression is preferrable over ridge regression. The illustration is as follows:

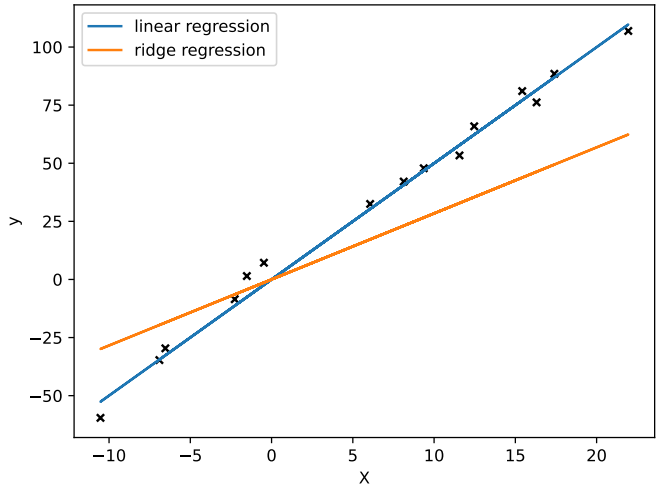

Figure 2: Setting 1 of (X, y) and the corresponding linear regression and ridge regression model.

This is because the fitting of data can be nicely down simply by minimizing the objective $F(w) = ||Xw - y||_2^2$,

whereas the penalty of ridge regression on the norms of weights prevents the minimization, which results in a worse fit of the data points.

3.2

When there are outliers in (X, y), the linear regression model can be highly sensitive to the outliers, while the ridge regression model remains robust and fit the majority of the points (as shown in Figure 3). Therefore, ridge regression is preferrable over linear regression when there does not exist a strong linear relation among all the data in (X, y).

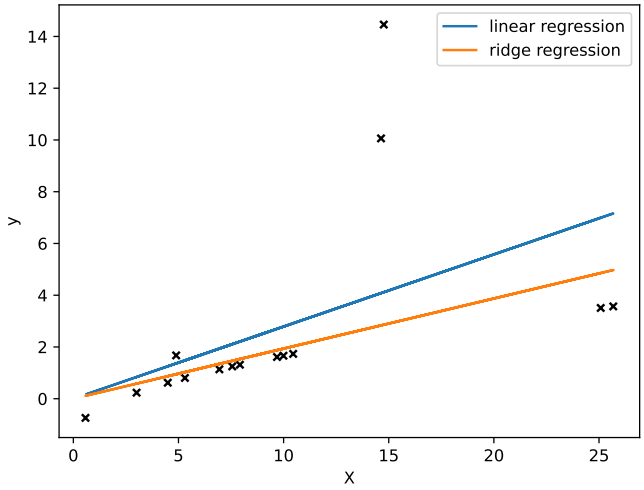

Figure 3: Setting 2 of (X, y) and the corresponding linear regression and ridge regression model.

3.3

Let $L_{Ridge}(w) = ||Xw - y||_2^2 + \frac{\eta}{2}||w||_2^2$, from 1, we know $\nabla_w L_{Ridge}(w) = 2X^T(Xw - y) + \eta I w$ where $\eta > 0$, thus $\nabla_w^2 L_{Ridge}(w) = \nabla_w(\nabla_w L_{Ridge}(w)) = 2X^T X + \eta I > 0$, so $L_{Ridge}(w)$ is a convex function.
Let $\nabla_w L_{Ridge}(w) = 0$, then we have $(2X^T X + \eta I)w = 2X^T y$, hence $w = \frac{2X^T y}{2X^T X + \eta I}$ Therefore, the close-form solution of ridge regression is $(2X^T X + \eta I)^{-1} 2X^T y$.

3.4

(a)

Under the extreme case of multicollinearity, where some features are identical to others, the columns of the matrix X^T will no longer be linearly independent. Consequently, $\det(X^T X) = \det(X^T)\det(X) = 0$, therefore $X^T X$ is not invertible. Considering the closed-form solution to vanilla linear regression $(X^T X)^{-1} X^T y$ requires to take the inversion of $X^T X$, we will no longer be able to compute this solution.

(b)

Because $(X^T X)^T = (X)^T (X^T)^T = X^T X$, $X^T X$ is a symmetric matrix, hence $X^T X$ is orthogonally diagonalizable, i.e., $X^T X$ can be decomposed as $X^T X = V^{-1} \Sigma V$ where V is orthogonal and Σ is a diagonal matrix.
On the other hand, $V^{-1} V = V^{-1} I V$, hence $2X^T X + \eta I = V^{-1}(2\Sigma + \eta I)V$. We know from the last problem that $\det(X^T X)$ can be 0, which makes it impossible to calculate the closed-form solution to vanilla linear regression.
Therefore, $\det(X^T X) = \det(V^{-1})\det(\Sigma)\det(V) = 0$, where $\det(V^{-1}), \det(V) \neq 0$, hence $\det(\Sigma) = 0$.

Let $\Sigma = \begin{pmatrix} \Sigma_1 & & & \\ & \Sigma_2 & & \\ & & \ddots & \\ & & & \Sigma_k \end{pmatrix}$, where $\Pi_{i=1}^{k}\Sigma_i = 0$, then $2\Sigma + \eta I = \begin{pmatrix} 2\Sigma_1 + \eta & & & \\ & 2\Sigma_2 + \eta & & \\ & & \ddots & \\ & & & 2\Sigma_k + \eta \end{pmatrix}$.
When η is large enough ($\eta > -2\min_i \Sigma_i$), $\det(\Sigma + \eta I) = \Pi_{i=1}^{k}(2\Sigma_i + \eta) > 0$, hence $\det(2X^TX + \eta I) = \det(V^{-1})\det(2\Sigma + \eta I)\det(V) > 0$, $2X^TX + \eta I$ is invertible, hence the closed-form solution of ridge regression can always be obtained when η is large enough.
This implies another benefit of using ridge regression is that when the dataset suffers from multicollinearity, ridge regression can always be used to obtain a solution.

4 Locality Sensitive Hashing (LSH)

4.1

4.2

4.3

4.4

4.5

4.6

5 Programming Problem: Linear Regression

5.1

After checking all the scatter plots, we picked out the three features that look the most linearly related to price on the scatter plots: **LSTAT**, **RM**, and **INDUS**. The plots are as follows:

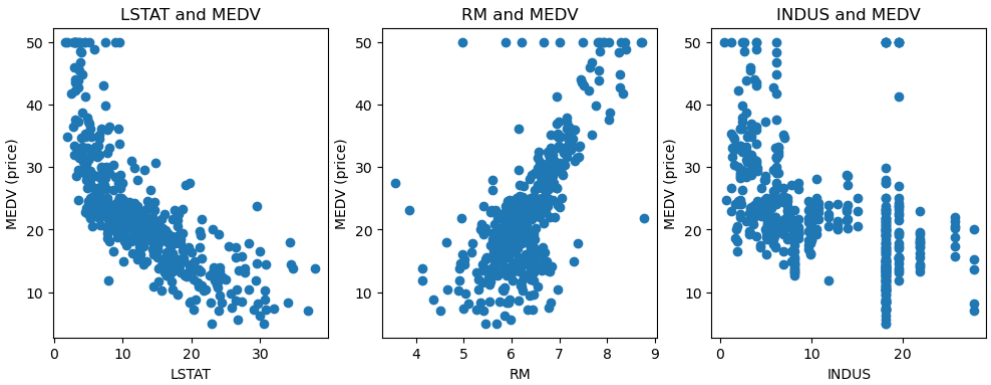

Figure 4: Scatter plots of LSTAT vs MEDV(left), RM vs MEDV(middle), and INDUS vs MEDV(right)

5.2

According to the correlation matrix, the 3 features that are most linearly related to the house price are: **LSTAT**($r = -0.74$), **RM**($r = 0.7$), and **PTRATIO**($r = -0.51$). This is slightly different from our results in 5.1, but the discrepancy is understandable, as the difference between the Pearson scores of **PTRATIO**($r = -0.51$) and **INDUS**($r = -0.48$) is very small.

5.3

According to 3.3, we can know the closed-form solution to linear regression $\min_w ||Xw - y||_2^2$ and ridge regression $\min_w ||Xw - y||_2^2 + \frac{\eta}{2}||w||_2^2$ are $w = (X^T X)^{-1} X^T y$ and $w = (2X^T X + \eta I)2X^T y$, respectively. After implementing these two solutions in Python, we obtained the coefficients corresponding to each feature as follows:

η / features	linear regression	ridge regression		
	0	15.0	45.0	90.0
CRIM	-0.099324	-0.100648	-0.101396	-0.101484
ZN	0.052251	0.054632	0.059028	0.062642
INDUS	**0.004516**	**0.012958**	**0.018062**	**0.020644**
CHAS	**2.957261**	**2.272783**	**1.575958**	**1.107609**
NOX	1.127938	0.457674	0.343826	0.287127
RM	**5.854198**	**5.728152**	**5.424074**	**5.008160**
AGE	-0.014957	-0.010094	-0.002772	0.006178
DIS	-0.920844	-0.896985	-0.842988	-0.770484
RAD	0.159519	0.163084	0.164232	0.162159
TAX	-0.008934	-0.008982	-0.008940	-0.008670
PTRATIO	-0.435674	-0.406149	-0.345226	-0.260870
B	0.014905	0.015518	0.016406	0.017465
LSTAT	-0.474751	-0.484274	-0.506287	-0.534369

Table 1: The coefficients corresponding to different features under different η's. Note linear regression can be viewed as a special case of ridge regression where $\eta = 0$.

Consider the absolute values of the coefficients, as $\eta \nearrow$, the larger absolute values get smaller (like **CHAS** and **RM**), whereas the smaller absolute values get larger (like **INDUS**). In other words, larger η leads to the averaging of the norms of the regression weights.

5.4

We calculated the root mean square error (RMSE) of train and test set under different η's according to the formula $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$. The result are as follows:

η / dataset	linear regression	ridge regression		
	0	15.0	45.0	90.0
train set	4.8206	4.8263	4.8526	4.9076
test set	5.2092	5.1912	5.1895	5.2128

Table 2: RMSE of train set and test set under different η's.Note linear regression can be viewed as a special case of ridge regression where $\eta = 0$.

It is worth noticing that as η gets larger, the train set $RMSE$ gets larger as well, whereas the test set $RMSE$ decreases at first and increases at the end.
A possible explanation to this phenomenon is: on the train set, $RMSE$ is perfectly consistent with the objective of linear regression, therefore linear regression results in a smaller RMSE than ridge regression; on the test set, the penalty of ridge regression on large weights improves the generalizability of the model, therefore the $RMSE$ on test set is smaller when η gets larger. However, when η is too large, the model will focus too much on minimizing the weights instead of fitting the data points, resulting in high $RMSE$ on both train set and test set.

5.5

We picked out the 3 most significant features as noted in 5.1 and 5.3, clipped the data by keeping only those 3 features, and trained a linear regression model and a ridge regression model ($\eta = 45.0$) on the clipped data. Afterwards, we calculated the $RMSE$ on train and test set under the new model. The results are as follows:

η / dataset	linear regression 0	ridge regression 45.0
train set	5.4798	5.4807
test set	5.6279	5.6123

Table 3: RMSE of train set and test set under different η's.Note linear regression can be viewed as a special case of ridge regression where $\eta = 0$.

Compared to the $RMSE$ we obtained in 5.4 where we used all 13 features for training and prediction, the $RMSE$ obtained with only 3 features does increases by at most 13.7%.
This implies by using only the top3 most significant features to predict the house prices, we can still obtain a comaparable performance compared to that of using all features, while cutting down the dimension of the feature space and saving computing power considerably.