HW2

Karolina Vandekerkhove- kvv2005

March 2024

1 Linear Regression and Convexity

y= n-dimensional vector X= n x d matrix of full rank w= d- dimensional vector
A function is convex if its Hessian matrix is PSD
$\nabla L(w) = \nabla y - Xw_2^2$
$\nabla L(w) = \nabla((y - Xw)^T(y - Xw))$
$= \nabla L(w) = \nabla(y^T y - y^T Xw - X^T w^T y + X^T w^T Xw$

In respect to w the derivatives are
$y^T y = 0$
$y^T Xw = -X^T y$
$X^T w^T y = -X^T y$
$X^T w^T Xw = 2X^T Xw$
$= \nabla L(w) = (0 - X^T y - X^T y + 2X^T Xw)$
$= \nabla L(w) = -2X^T y + 2X^T Xw$
$= \nabla L(w) = -2X^T(y - Xw)$
$\nabla^2 L(w) = \nabla(-2X^T(y - Xw))$
$\nabla^2 L(w) = 2X^T X$

Since X is full rank, we conclude that $X^T X$ is positive definite, this also implies that all eigenvalues are non-negative. Therefore, the Hessian form is PSD which implies that $L(w)$ is convex.

2 Problem 2: Gaussian Distribution and the Curse of Dimensionality

2.1

2D cases:
$\mathbf{S}_1(r) = 2\pi r$
$S_2(r) = \pi r^2 (Area)$
3D cases:

$\mathbf{V}_1(r) = 4\pi r^2$
$V_2(r) = 4/3\pi r^3$

2.2

$\mathbf{S}_{m-1}(r) = \frac{d}{dr}V_m(r)$ **The equation describes that the surface area is equal to the derivation of the volume. Intuitively speaking, we know that that an integral is the area under the curve, and the area in question here is the volume. Under the surface of the sphere, we have its volume. Taking a real life example, let's say we wanted to paint a sphere ball. Every time a layer of paint is placed, the volume increases by the surface area we just painted on. So, there is a relation between surface area and volume as the more layers you paint (always the whole surface area) the more the volume will increase. The rate of increase in surface area is directly related to the rate increase in volume. Let us use the 2D/3D cases as verification:**
2D:
When you increase the radius of a circle, the circumference increases (since $2\pi r$). Increasing circumference will also increase the area. The relation can be further emphasised if we derive the equation for the area in respect to radius πr^2 we get $2\pi r$ which is equivalent to the circumference.
3D: Increasing radius will lead to an increase in surface area, corresponding to an increase in volume of that surface area. Looking at the equation derivatives again, the surface area ($4\pi r^2$) when derived gives us $4/3\pi r^3$, further emphasizing the relationship

2.3

We know $V_m(r)$ is dependent on r^m since: m determines how many dimensions the sphere spans as well as its degrees of freedom, and that radius determines volume which expands to r^m in higher dimensions. Since r=1, surface area remains constant
$\mathbf{S}_{m-1}(1) = (\bar{S}_{m-1})(r^{m-1})$

2.4

$\mathbf{p}_m(r) = \int_{S^{m-1}} rp(x)\, dA$
$\int_{S^{m-1}} rp(x)\, dA$ **= p(x) *** $\bar{S}_{m-1}$
$p_m(r) = p(x) * \bar{S}_{m-1} * r^{m-1}$

2.5

$\mathbf{p}_m(r) = p(x) * \bar{S}_{m-1} * r^{m-1} = \frac{1}{(2\pi\sigma^2)^{\frac{m}{2}}} exp(-\frac{r^2}{2\sigma^2}) * \bar{S}_{m-1} * r^{m-1}$
$\frac{\bar{S}_{m-1}}{(2\pi\sigma^2)^{\frac{m}{2}}}(exp - (\frac{r^2}{2\sigma^2}) * r^{m-1})$

$\frac{d\rho_m(r)}{dr} = \frac{\bar{S}_{m-1}}{(2\pi\sigma^2)^{\frac{m}{2}}} = exp(\frac{r^2}{2\sigma^2}) * (-\frac{r}{\sigma^2} * r^{m-1} + (m+1)r^{m-2}) = 0$
$\frac{r}{\sigma^2} * r^{m-1} + (m-1)r^{m-2} = 0$
$\frac{r^m}{\sigma^2} = (m-1)r^{m-2}$
$\frac{r^m}{r^{m-2}} = (m-1)\sigma^2$
$r^2 = (m-1)\sigma^2$ **As m increases, -1 becomes insignificant so we have** $\hat{r} \approx \sqrt{m\sigma}$

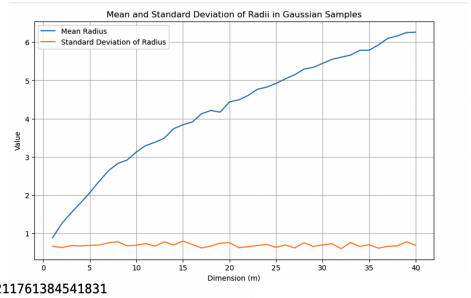
2.6

$$\rho_m(\hat{r}+\epsilon) = \frac{\bar{S}_{m-1}}{(2\pi\sigma^2)^{\frac{m}{2}}} \cdot \exp\left(-\frac{(\hat{r}+\epsilon)^2}{2\sigma^2}\right) \cdot (\hat{r}+\epsilon)^{m-1}$$
$$\frac{\bar{S}_{m-1}}{(2\pi\sigma^2)^{\frac{m}{2}}} \cdot \exp\left(-\frac{(\hat{r}^2+\epsilon^2+2\hat{r}\epsilon)}{2\sigma^2}\right) \cdot \hat{r}^{m-1} \cdot (1+\epsilon\hat{r})^{m-1}$$
$$\rho_m(\hat{r}) \cdot \exp\left(-\frac{\epsilon^2+2\hat{r}\epsilon}{2\sigma^2}\right) \cdot (1+\frac{\epsilon}{\hat{r}})^{m-1}$$
$$\rho_m(\hat{r}) \cdot \exp\left(-\frac{\epsilon^2}{2\sigma^2} - \frac{\hat{r}\epsilon}{\sigma^2} + \frac{m\epsilon}{\hat{r}} - \frac{m\epsilon^2}{2\hat{r}^2}\right)$$
$$\rho_m(\hat{r}) \cdot \exp\left(-\frac{\epsilon^2}{2\sigma^2} - \frac{\hat{r}\epsilon}{\sigma^2} + \frac{\hat{r}\epsilon}{\sigma^2} - \frac{\epsilon^2}{2\sigma^2}\right)$$
$$\rho_m(\hat{r}) \cdot \exp\left(-\frac{\epsilon^2}{\sigma^2}\right)$$

2.7

As we increase dimensions, most of the points would reside along the edge of the radius $\hat{r}$ and the sphere, so you could say the skin of the fruit. When we have lower dimensions, we notice that most points are now residing near the center.

2.8



PDF at the origin: 0.00010211761384541831
PDF at a random point on the sphere: 1.382011619321673e-05

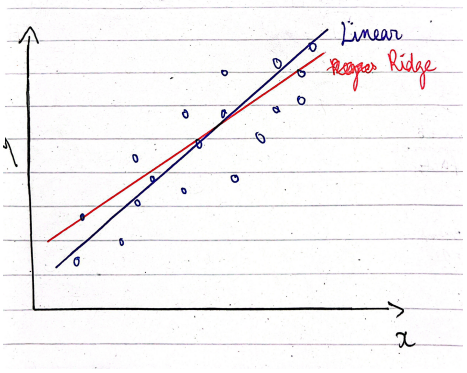
From the figure and calculate density , we can see how the mean and standard deviations change at different radii. The origin has a significantly higher density than a point at a sphere. As we tend to move away, density tends to decrease with increased dimensions. This graph is consistent with previous observations. Since mean increase as dimensions do, more and more points will be on the skin

of the sphere. Increasing dimensions we also increase our distances between points which is why standard deviations would change a lot.

3 Problem 3: Ridge Regression

3.1

Standard linear regression is usually favored when the data does not have or has little multicollinearity. This phenomenon undermines statistical significance since it is hard to determine the impact of individual variables on data. Moreover, linear does not provide regularization to calibrate and prevent under/overfitting. This means that for data sets with low multicollinearity, linear can be a good option. The diagram below is an example of linear regression winning over ridge:

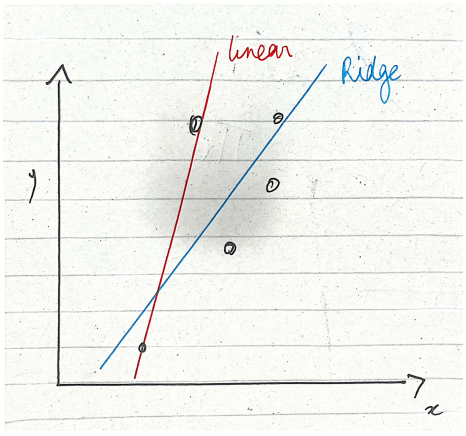


Linear over Ridge

Here, linear regression is favored since the data is relatively straightforward, and adding penalty like ridge does is unnecessary and counter-productive. With little multicollinearity, the risk of overfitting is very low, favoring linear. Also, we can see in the following example that ridge flattens the curve when that is not necessary in this case. It regularizes the data when there is no clear need for it.

3.2

In contrast, a setting where ridge regression is when the data points are high in multicollinearity or contain several outliers. Ridge adds bias to the estimates. It also uses regularization to solve overfitting. As seen from the graph below:



Enter Caption

We can see that Linear takes into account outliers, which potentially skew the result and provide a misinterpretation of the data. It uses a penalty term to lower the high coefficients near zero, which in certain cases are valued over linear regression where this does not occur.

3.3

Ridge regression is given as: min= $\| Xw - y \|_2^2 + \frac{\eta}{2} \| w \|_2^2$
η **=regularization parameter**
$0 = (y - Xw)^T(y - Xw) + \frac{\eta}{2} w^T w$
$0 = y^T y - 2w^T X^T y + w^T X^T X w + \frac{\eta}{2} w^T w$
$0 = -2X^T y + 2X^T X w + 2\frac{\eta}{2} w$
$-X^T y + X^T X w + \frac{\eta}{2} w = 0 X^T X w + \frac{\eta}{2} w = X^T y$
$w(X^T X + \frac{\eta}{2} I) = X^T y$
$w = X^T y (X^T X + \frac{\eta}{2} I)^{-1}$

3.4

3.4.1

1) If there are more columns than rows, the matrix is not perfectly invertible, since it is not a square matrix. Regularization can still be applied in terms of ridge regression
2) If the features are identical, the matrix will not have a unique inverse so a solution cannot be computed. Again, ridge regression can use regularization to stabilize results.

3.4.2

η **ensures invertibility for the matrix** $X^T X$**. This also means that our results are more stable than in linear regression. That coefficient is also used for regularization, a feature linear regression does not have. Ridge regression allows large coefficients to shrink towards**

zero which reduces overfitting. This could also potentially reduce any noise that may infringe data analysis.

4 Problem 4: Locality Sensitive hashing

4.1

1) If there exists a point within r from point q, oracle returns point x such that distance of x and q is max cr
2) No point within cr distance, oracle returns nothing
3) If there is a point with cr but not within r, unstable so can be inconsistent.

If we want to find the exact nearest neighbor in X of query point q:
We can first set r to 1, if the oracle returns something, we are within distance r so set c to 1.
We double r until the oracle returns us with nothing (there is no nearby point within cr)
Once we find the range of r, we can perform a binary search to search through the data and find the exact nearest point.

4.2

p1= prob single hash function hashes two close points
p2= prob that single hash function hashes two far points
upper bound $p_2 \leq$1-(cr$\frac{}{m)}$
Lower bound $p_1 \geq$1- (r$\frac{}{m)}$

4.3

Lower bound $p_1 \geq p_1^k$
Upper bound $p_2 \leq p_2^k$

4.4

We need at least one instance of g function to map. Probability that no maps:
Upper bound= $1-(1-p_2^k)$
Lower bound= $1-(1-p_1)^l$ **taking its complement to give us the lower bound Upper bound is the same since we are looking for at least one match for all g instances.**

4.5

Event B:
$P(X \geq a) \leq \frac{E[X]}{a}$
$E[X_b] = n_p p_2^k$
$P(X_b \geq 4l) \leq \frac{n_p p_2^k}{4l}$
$P(X_b \geq 4l) \leq \frac{1}{4l}$
$]EventAhappeningis1\text{- e}^{-1}\ e^{-1} = (\frac{1-1}{k})^k$
Let $p_1 \geq \frac{1}{2}^r and p-2 \leq 1-\frac{1}{2}^r$
$\rho = \frac{ln(p_1}{ln(p_2)}$
$\text{l= n}_p^\rho k = \frac{lnn_p}{ln1/p_1}$
Probability that q does not match in all k is $1-p_1)^k$
Since $k = \frac{lnn_p}{ln1/p_1}\ p_1^k = \frac{1}{n_p}$
Thus, $(1-p_1)^k \approx e^{-1}$
Hence, probability q matches:
$1-(1-p_1)^k l \approx 1-e^{-kl}$
$= 1-e^{-1}$

Event B:
$P(X \geq a) \leq \frac{E[X]}{a}$
$E[X_b] = n_p p_2^k$
$P(X_b \geq 4l) \leq \frac{n_p p_2^k}{4l}$
$P(X_b \geq 4l) \leq \frac{1}{4l}$

4.6

We should check if all points within the l buckets that share the same hash value as q for each hash function. It is not guaranteed to find a point at most cr of q due to the probabilistic nature of hashing. If we run multiple instances, the probability of missing all the close points becomes negligible.

5 Problem 5: Linear Regression (code)

5.1

Top 3 features mostly related to MEDV:
LSTAT -0.737663
RM 0.695360
PTRATIO -0.507787

5.2

Top 3 features mostly linearly related to MEDV (correlation matrix):
LSTAT', 'RM', 'PTRATIO'

These correspond to the results from part 1

5.3

See code for the computations. The eta values 0.1, 1.0, 10.0 and 100.0 were used for ridge regression. I settled on 10 for the analysis parts

5.4

Training RMSE - Linear Regression: 4.820626531838223
Training RMSE - Ridge Regression: 4.829777333975097
Test RMSE - Linear Regression: 5.209217510530916
Test RMSE - Ridge Regression: 5.189347305423606

Training RMSE analysis:
Linear regression here tends to have a slightly lower RMSE for training datasets. Linear tends to perform better for training datasets since it aims to minimize error between its predictions and actual values. In contrast, ridge regression tends to correct for overfitting, which would be a reason why the RMSE is slightly higher

Test RMSE analysis:
We see here that ridge regression performs slightly better than linear on test datasets a.k.a on unseen datasets Ridge aims to use regularization on datasets. Predictors remain but it shrinks the coefficients rather than equaling them to zero. This is meant to control overfitting, where the algorithm provides near perfect results for seen data but cannot do the same for unseen. This is why we can see a slightly better result for ridge than linear, which is more sensitive to outliers as it is more prone to underfitting by failing to find a pattern or overfitting which renders the model unprepared to tackle unknown data

5.5

RMSE using top-3 features:
Training RMSE - Linear Regression: 5.273361751695365
Training RMSE - Ridge Regression: 5.276310228536868

Test RMSE - Linear Regression: 5.494723646664543
Test RMSE - Ridge Regression: 5.47757344311873

New RMSE analysis:
Using only the top 3 features increases our RMSE overall, but the overall trend is still similar Linear performs better in training, and ridge in test data. The overall trend is also similar in that error increases with the test data This could be caused by loss of information, since the algorithm has less data to work with and would be harder to find a general pattern This could also indicate that taking all 13 features into account provides valuable information to the trend, regardless that the top 3 features are the most linear. Intuitively, the top 3 features are not the deciding factor in the Boston housing prices, which would be why the RMSE is slightly higher, since there is lower accuracy to the actual pricing of housing.