Tejaswini Srinivas

FG43775

Collaborated with –

Harish Ramamoorthy

1. **Support Vector Machines**

*We have seen that in p = 2 dimensions, a linear decision boundary takes the form β0+β1X1+β2X2 =0. We now investigate a non-linear decision boundary.*

1. *Sketch the curve (1 + X1)2 + (2 - X2)2 = 5*

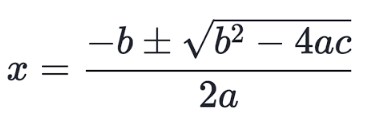
*When considering X1 =0 , we get the following equation –*

*(1+0) 2+(2-X2)2=5*

*1++4-4 X2+ X2* 2 *=5*

*-4X2+X2* 2=0

The above equation can be applied in quadration formula



Substituting values b=-4 ,a=1 ,c=0

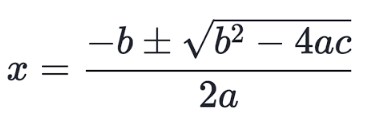
We get, X11=4 and X12=0

*When considering X2 =0 , we get the following equation –*

*(1+X1)2+(2-0)2=5*

*2X1+ X1* 2*=0*

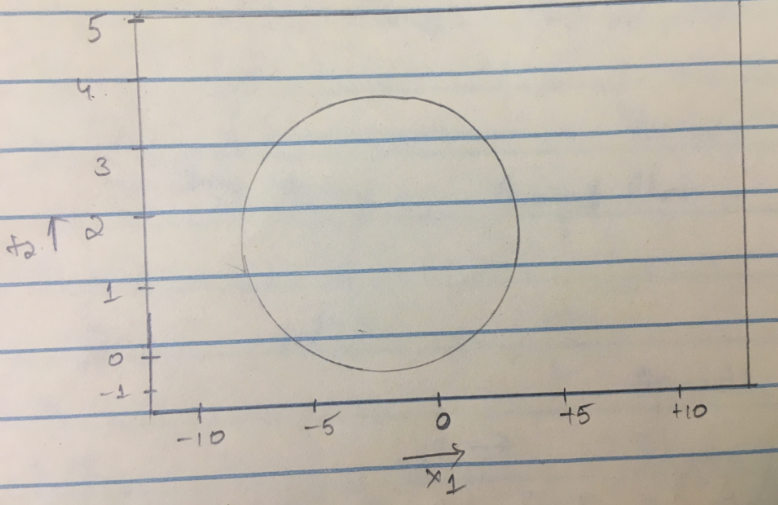
above equation can be applied in quadration formula



Substituting values b=-2 ,a=1 ,c=0

We get, X21=0 and X22=-2

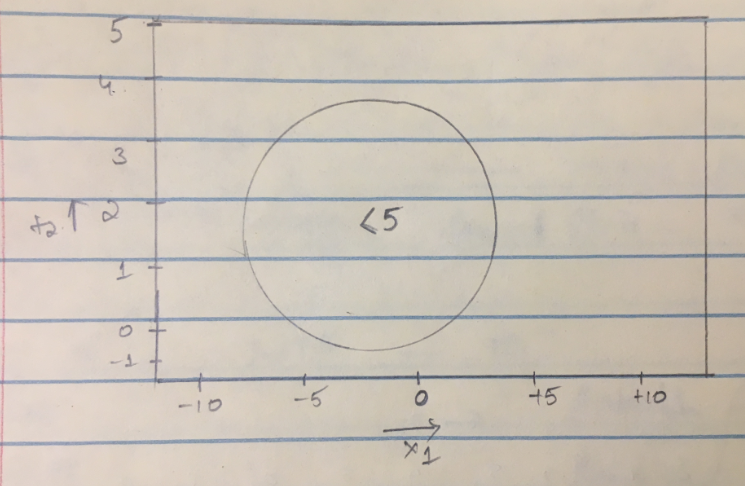
Below is the curve drawn for the given values.

**

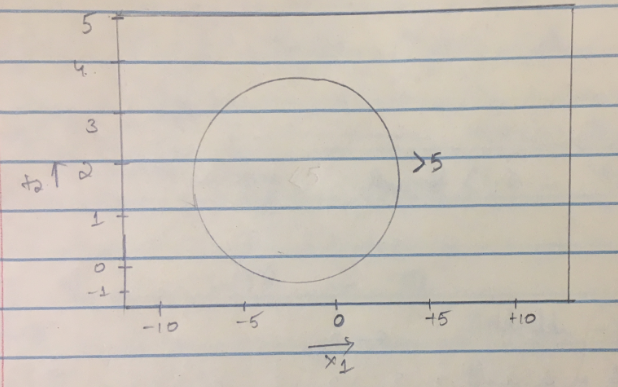
1. *On your sketch, indicate the set of points for which (1 + X1)2 + (2 - X2)2 > 5*

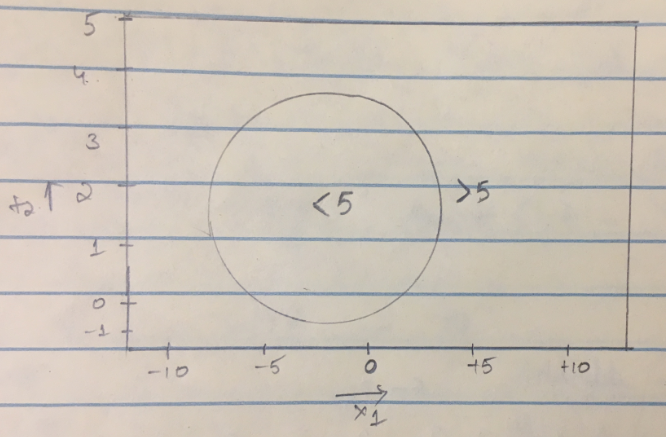
*as well as the set of points for which (1 + X1)2 + (2 - X2)2 <=5*

*Below is the plot for (1 + X1)2 + (2 - X2)2 < 5*

**

Below is the plot for (1 + X1)2 + (2 - X2)2 >5

**



1. *Suppose that a classifier assigns an observation to the blue class if*

*(1 + X1)2 + (2 - X2)2 >5and to the red class otherwise. To what class is the observation (0; 0) classified? (-1; 1),(2; 2),(3; 8)*

By substituting (-1, 1),(2, 2),(3, 8) values in the given equation we get below values and the specified class.

f(−1,1) = 1 ≤5, red class.  
f(2,2) = 9 >5, blue class.  
f(3,8) = 52 >5, blue class.

1. *Argue that while the decision boundary in (c) is not linear in terms of X1 and*

*X2, it is linear in terms of X1;X21 ;X2;X22 .*

ƒ(X) = (1+X1)2 + (2 - X2)2 -5

= 1+ X12 + 2X12+4+ X22 -4X22-5

ƒ(X) = X12 + X22 + 2X1 - 4X2

Given equation (1+X1)2 + (2 - X2)2 > 5 can be written in quadratic form

1+ X12 + 2X12+4+ X22 -4X22>5

X12 + 2X12+ X22 -4X22+5>5

 The above equation can be transformed into form of higher dimension:



Where, β0 = 0, β1=2, β2=-4, β3=1, β4=1,

we can say that ƒ(X) is non-linear with X1 and X2 and f(X) is linear with X1, X2 , X12,X22

*.*

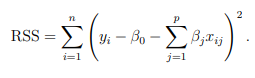
*.*

*2*

1. **Model Selection**
2. *Why does Ridge Regression improve over least squares? Explain.*

Ridge regression’s advantage over least squares is embedded in the *bias-variance trade-off*. As *λ i*ncreases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.

The least square fitting procedure estimates *β0, β1, β2… βp* using the values that minimizes

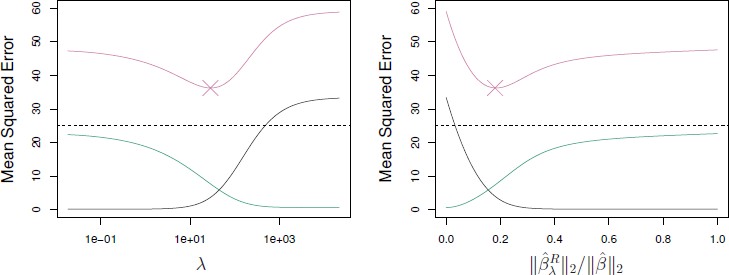


The Ridge Regression co-efficient estimates the *β̂R* values that minimize

where, λ>=0 is a tuning parameter. As with least square methods, ridge regression finds the coefficient that best fits the data and also finds the unbiased coefficients by making the RSS small. Unbiased means that the OLS (Ordinary Least Square) will not consider which independent variable is more important than the others; rather it simply finds the coefficients for the given data set.

The shrinkage penalty (λ ∑ *βj* 2) is small when *β0, β1, β2… βp*  are near to zero and services the co-efficient estimates to zero. The tuning parameter controls the impact on the coefficient estimates. Here ridge will have the smallest MSE when compared to Least square model as the majority of the coefficients are expected to be shrinked to zero value. The Ridge has better prediction accuracy when *p > n* in order to control the variance and model interoperability as by removing the irrelevant features from the model .

By considering the observation n = 50 and the value of p = 45 predictors having non zero co- efficient with the squared bias, variance, and test mean squared error horizontal dashed line indicates the minimum MSE and test mean squared error (purple) crosses indicate the ridge regression where MSE is smallest



1. *What are the disadvantages of Ridge Regression? Explain.*

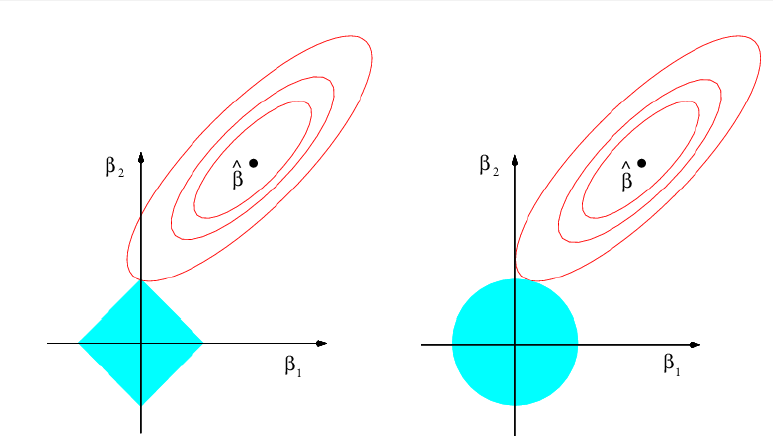
* The ridge regression includes all the predictors ‘p’ in the final model, unlike the subset selection which selects the model the specific subset of variables.
* Ridge regression includes all predictors because of the model interoperability which shrinks the coefficients for least important predictors, very close to zero. But it will not make them exactly zero.
* Since, ridge cannot shrink coefficient to zero it cannot perform variable selection.

1. *Why is it that the LASSO, unlike Ridge Regression, results in coefficient estimates that are exactly equal to zero? Explain.*

Unlike Ridge, Lasso regression performs well with variable selection. LASSO can produce a model that has high predictive power and it is simple to interpret.

In ridge regression, it shrinks only the parametric values where as in LASSO, it selects the parameters first and then it shrinks so that they are in the corners of the constraint. As the LASSO performs well for variable selection and it shrinks the coefficients to be 0 when the *λ is large.*

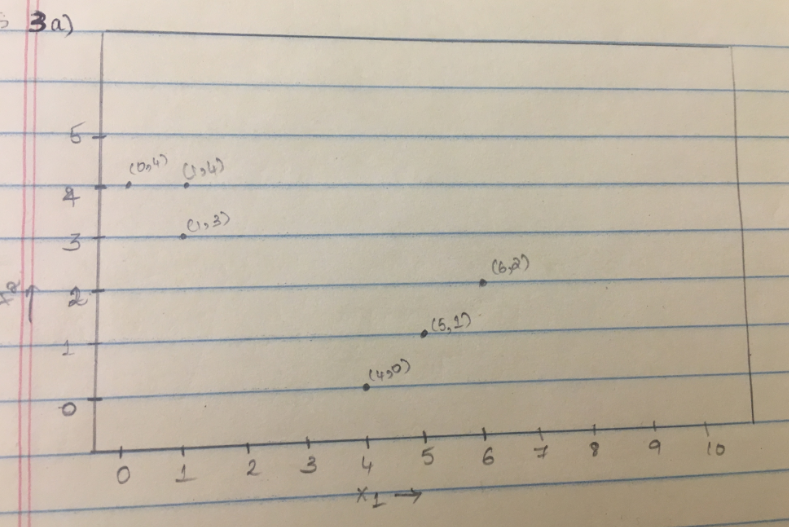
LASSO can select the value of λ easily as it uses cross-validation method which yields to the smallest prediction error,where as in Ridge it is difficult the value of it. The main difference between them is the is the shape of constraints, when the value of *p* = 2 theshape for Ridge is circle as it uses *l1* norms for constraints and Lasso is diamond so does *l2* norms for constraints. When the value of *p* increases the diamond increase in number of corners of the coeffients and so will be set to zero. From this we can say that lasso performs shrinkage successfully than ridge regression.



1. **Unsupervised learning**

*In this problem, you will perform K-means clustering manually, with K = 2, on a small example with n = 6 observations and p = 2 features. The observations are as follows.*

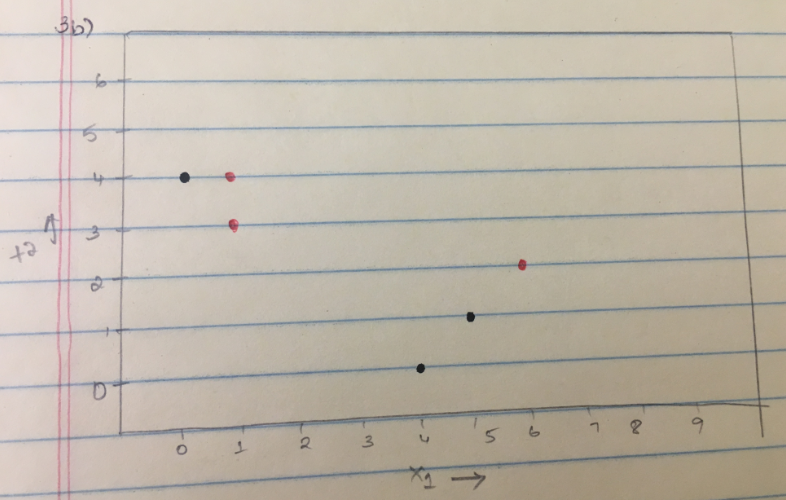
1. *Plot the observations*



1. *Randomly assign a cluster label to each observation.*

Cluster 1: Black

Cluster 2: Red

****

1. *Compute the centroid for each cluster.*

Centroid for Black cluster-

X1= 1/3(0+4+5) =3

X2 = 1/3(4+0+1) =1.66

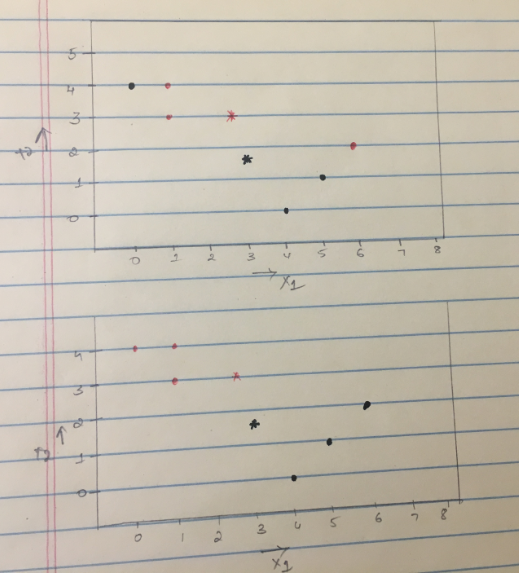
Centroid for Red cluster –

X1 = 1/3(1+1+6) =2.66

X2 = 1/3(2+4+3) =3

1. *Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X1 | X2 | Euclidean distance for Black cluster | Euclidean distance for Red cluster | Cluster Label |
| 1 | 4 | 3.07 | 1.94 | Red |
| 1 | 3 | 2.40 | 1.66 | Red |
| 0 | 4 | 3.80 | 2.847 | Red |
| 5 | 1 | 2.10 | 3.07 | Black |
| 6 | 2 | 3.01 | 3.48 | Black |
| 4 | 0 | 1.94 | 3.28 | Black |



1. *Repeat (c) and (d) until the answers obtained stop changing.*

**Centroid for Black cluster-**

**X1= 1/3(4+5+6) =5**

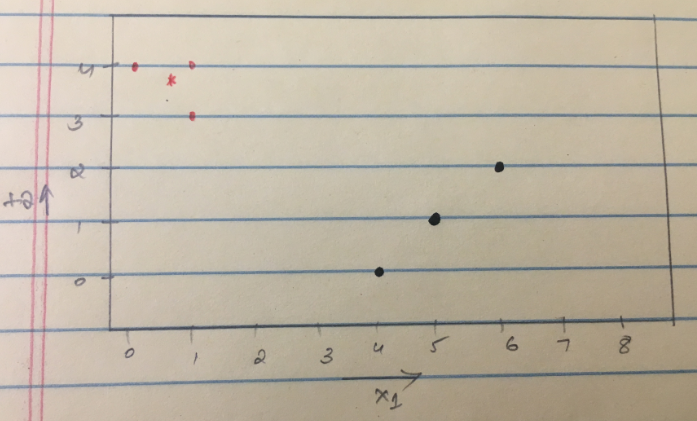
**X2 =1/3(0+1+2) =1**

**Centroid for Red cluster-**

**X1 =1/3(0+1+1) =0.66**

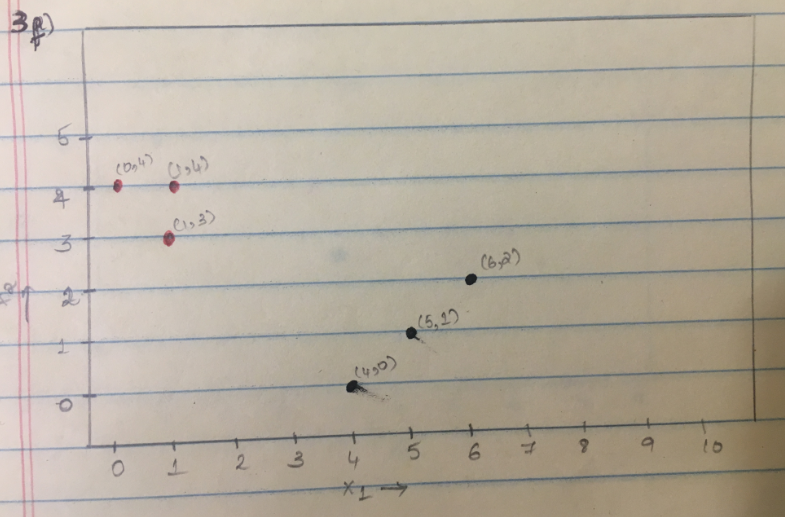
**X2 =1/3(3+4+4) =3.66**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X1 | X2 | Euclidean distance for Black cluster | Euclidean distance for Red cluster | Cluster Label |
| 1 | 4 | 5 | 0.47 | Red |
| 1 | 3 | 4.4 | 0.74 | Red |
| 0 | 4 | 5.83 | 0.74 | Red |
| 5 | 1 | 0 | 5.08 | Black |
| 6 | 2 | 1.41 | 5.58 | Black |
| 4 | 0 | 1.41 | 4.96 | Black |

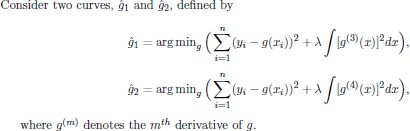
****

**f.** *In your plot from (a), color the observations according to the cluster labels*

*obtained.*



1. **Non-linearity**

****

1. *As λ->* ***∞****, will ĝ1 or ĝ2 have the smaller training RSS?*

When λ-> **∞**, *ĝ2* will have smaller training RSS when compared to *ĝ1* that is by considering the higher order of derivative on the penalty function as it decreases the bias.

1. *As λ->* ***∞****, will ĝ1 or ĝ2 have the smaller test RSS?*

When λ-> **∞**, *ĝ1* will have smaller test RSS than *ĝ2* as it increases the variance in the higher order of derivative which is more than decrease in the bias, this might be because of the over fit in the degree of the freedom.

1. *As λ = 0, will ĝ1 or ĝ2 have the smaller training and test RSS?*

When the λ tends to 0, then *ĝ1* and *ĝ2* will have almost the same training and test RSS i.e. **(*ĝ1* = *ĝ2*)**.

**References:**

ISLR textbook

Slides

<https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db>