VIGNESH PRABHAKAR
4024890667

Homework : 4

6.8.3

a) (iv) steadily decreases. When the value of s (budget term) increases; the value of λ decreases. This implies that upon increasing the value of s; all the β's increase from zero to their least square estimate values. therefore training RSS steadily decreases.

(b) (ii) Decreases initially and then eventually starts increasing in a U shape. When $s=0$; the model has high test RSS as all the β_s are ~~zero~~ zero. Upon increasing s; the model fits the test data leading to non zero estimates for the βs. So test RSS initially decreases. Upon further increasing the value of s; we end up overfitting and therefore the test RSS rises again leading to the U shape characteristic.

(c) (iii) steadily increases. When $s=0$ we have a high λ. So; all β_s are zero & the prediction is very different from the actual value. As we slowly increase s; λ decreases and we obtain non zero values for our βs. As we continue to further increase our δ; we overfit our model.

therefore the variance continuously increases with increasing values of δ.

(d) (iv) steadily decreases. When $\delta = 0$; our coefficient estimates (β_s) are zero and the predicted values are far from the actual values. This implies that at smaller values of β_s; the bias is very high and the model is underfitted. As we gradually increase the value of δ; the β_s begin to become non zero and the model starts to fit the training data well. So the bias continuously decreases with increasing values of δ.

(e) (v) Remains constant. Irreducible error is always independent of our model and choice of any associated parameters (in this case δ)

6.8.5

(a) Ridge regression optimization problem

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \beta_j x_j)^2 + \lambda \sum_{i=1}^{p} \hat{\beta}_i^2$$

$\hat{\beta}_0 = 0$, $n = p = 2$

Minimize

$$(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2$$
$$+ \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2) \rightarrow ①$$

(b) Differentiate ① wrt $\hat{\beta}_1$ & $\hat{\beta}_2$

with $x_{11} = x_{12} = x_1$ & ~~x_{21}~~ $x_{21} = x_{22} = x_2$

we get:

* diff wrt β_1:

$$(y_1 - (\hat{\beta}_1 + \hat{\beta}_2) x_1)^2 + (y_2 - (\hat{\beta}_1 + \hat{\beta}_2) x_2)^2$$
$$+ \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$$

$$2(y_1 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_1)(-x_1) + 2(y_2 - \hat{\beta}_1 x_2 - \hat{\beta}_2 x_2)(-x_2)$$
$$+ 2\lambda\hat{\beta}_1 = 0$$

$$-x_1 y_1 + x_1^2(\hat{\beta_1} + \hat{\beta_2}) - x_2 y_2$$
$$+ x_2^2(\hat{\beta_1} + \hat{\beta_2}) + \lambda \beta_1 = 0$$

$$(\hat{\beta_1} + \hat{\beta_2})(x_1^2 + x_2^2) - x_1 y_1 - x_2 y_2 + \lambda \beta_1 = 0$$

$$\beta_2 (x_1^2 + x_2^2) - x_1 y_1 - x_2 y_2$$
$$+ (\lambda + x_1^2 + x_2^2)\beta_1 = 0$$

$$(\lambda + x_1^2 + x_2^2)\beta_1 = x_1 y_1 + x_2 y_2 - \beta_2 (x_1^2 + x_2^2)$$

$$\boxed{\beta_1 = \frac{x_1 y_1 + x_2 y_2 - \beta_2 (x_1^2 + x_2^2)}{\lambda + x_1^2 + x_2^2}}$$

Similarly diff ① w.r.t β_2:

$$\cancel{2}(y_1 - \hat{\beta_1} x_1 - \hat{\beta_2} x_1)(-x_1) + \cancel{2}(y_2 - \hat{\beta_1} x_2 - \hat{\beta_2} x_2)(-x_2)$$
$$+ \cancel{2}\lambda \hat{\beta_2} = 0$$

$$-x_1 y_1 + x_1^2(\hat{\beta_1} + \hat{\beta_2}) - x_2 y_2 + x_2^2(\hat{\beta_1} + \hat{\beta_2})$$
$$+ \lambda \hat{\beta_2} = 0$$

$$-x_1 y_1 - x_2 y_2 + \hat{\beta_1}(x_1^2 + x_2^2) + \hat{\beta_2}(x_1^2 + x_2^2$$
$$+ \lambda) = 0$$

$\cancel{- x_1 y_1 - x_2 y_2}$

$$\hat{\beta}_2 (x_1^2 + x_2^2 + \lambda) = x_1 y_1 + x_2 y_2 - \hat{\beta}_1 (x_1^2 + x_2^2)$$

$$\boxed{\hat{\beta}_2 = \frac{x_1 y_1 + x_2 y_2 - \hat{\beta}_1 (x_1^2 + x_2^2)}{x_1^2 + x_2^2 + \lambda}}$$

the symmetry in these expressions implies that $\hat{\beta}_1 = \hat{\beta}_2$.

(c) Lasso ~~minimization~~ optimization problem:

$$(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda (|\hat{\beta}_1| + |\hat{\beta}_2|)$$

↳ need to minimize this

(d) minimize

$$(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda (|\hat{\beta}_1| + |\hat{\beta}_2|)$$

subject to $x_{11} = x_{12} = x_1$ & $x_{21} = x_{22} = x_2$ &

$x_{12} + x_{22} = 0$ & $y_1 + y_2 = 0$

the expression to be minimized simplifies to:

$$2\left(y_1 - (\hat{\beta_1} + \hat{\beta_2})x_{11}\right)^2$$

$\hat{\beta_1} + \hat{\beta_2} = \frac{y_1}{x_{11}}$ is a solution to this optimization problem

the above eqn is parallel to the edge of the lasso diamond $\hat{\beta_1} + \hat{\beta_2} = s$.

But since $\hat{\beta_1}$ & $\hat{\beta_2}$ vary along the line $\hat{\beta_1} + \hat{\beta_2} = \frac{y_1}{x_{11}}$; we can say that the contours touches the diamond at several points and therefore $\hat{\beta_1} + \hat{\beta_2} = s$ is one of the potential solutions.

Similarly we can make the case for the other part of the diamond $\hat{\beta_1} + \hat{\beta_2} = -s$ as well

therefore the lasso problem does not have a unique solution & general solution to this problem are the 2 line segments $\hat{\beta_1} + \hat{\beta_2} = s$ and $\hat{\beta_1} + \hat{\beta_2} = -s$.

8.4.5:

Majority vote approach

Find sum of those bootstrap samples where

$P(\text{class is red} \mid x) > 0.5$

$\Rightarrow 0.55 + 0.6 + 0.6 + 0.65 + 0.7 + 0.75$

$= 3.85$

Find sum of those bootstrap samples where

$P(\text{class is red} \mid x) < 0.5$

$\Rightarrow 0.1 + 0.15 + 0.2 + 0.2$

$= 0.65$

Since; sum of $P(\text{class is red} / x) > 0.5$ is greater than sum of $P(\text{class is red} / x) < 0.5$ we assign the final classification decision as RED for this approach.

Average Probability approach

Compute the average of all the probabilities which has been given for each bootstrap sample.

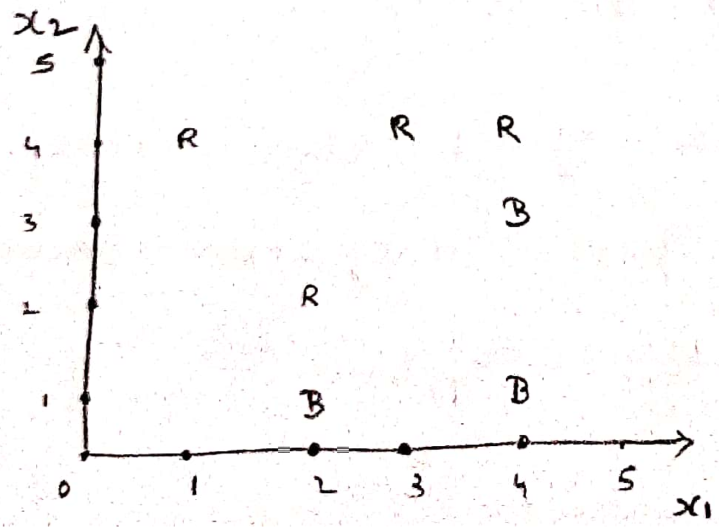
$$\Rightarrow \frac{0.1 + 0.15 + 0.2 + 0.2 + 0.55 + 0.6 + 0.6 + 0.65 + 0.7 + 0.75}{10}$$

$$= 0.45$$

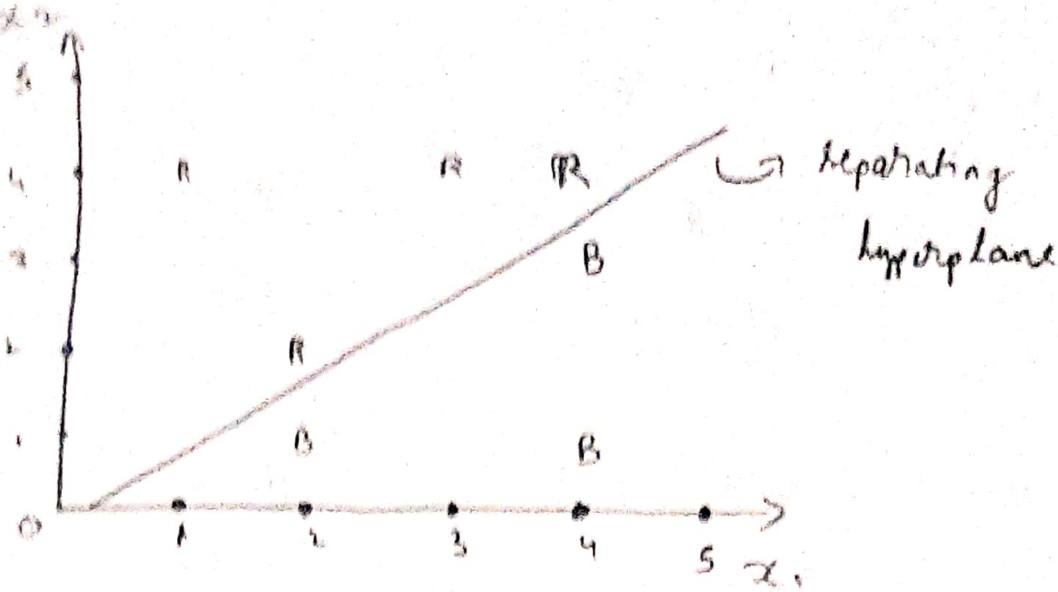
Since average of the probabilities is < 0.5, the final classification decision made using this approach is GREEN.

9.7.3 :

(a)



(b)



consider any given point on this hyperplane by eyeballing say : (2, 1.5) and (4, 3.5)

then eqn of line in 2 point form is given by :

$$x_2 - 1.5 = \frac{\cancel{2}}{\cancel{2}}(x_1 - 2)$$

$$x_2 - 1.5 = x_1 - 2$$

$$x_2 - x_1 = 2 - 1.5$$

$\boxed{x_2 - x_1 = 0.5}$ is the eqn of the hyperplane.

(c) Maximal margin classifier :

* $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$ then classify to red

else classify to blue

so comparing the above eqn with eqn of hyperplane we get :

$\beta_0 + \beta_1 x_1 + \beta_2 x_2 > 0 \Rightarrow$ Red

otherwise $\Rightarrow$ blue

$-0.5 - x_1 + x_2 > 0 \Rightarrow$ Red

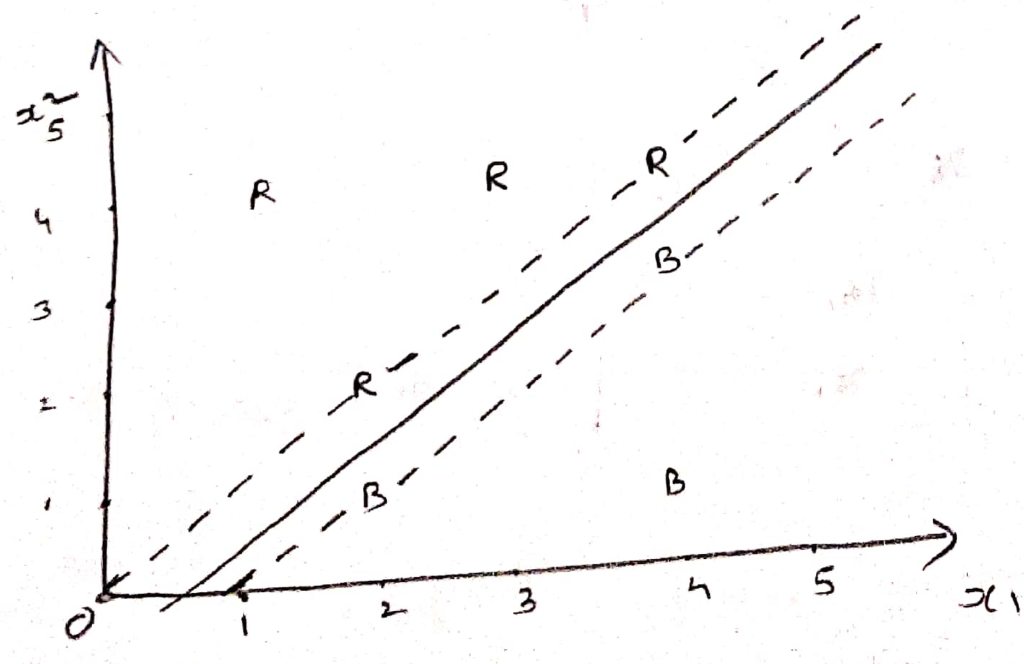
otherwise $\Rightarrow$ Blue

$\therefore \beta_0 = -0.5$

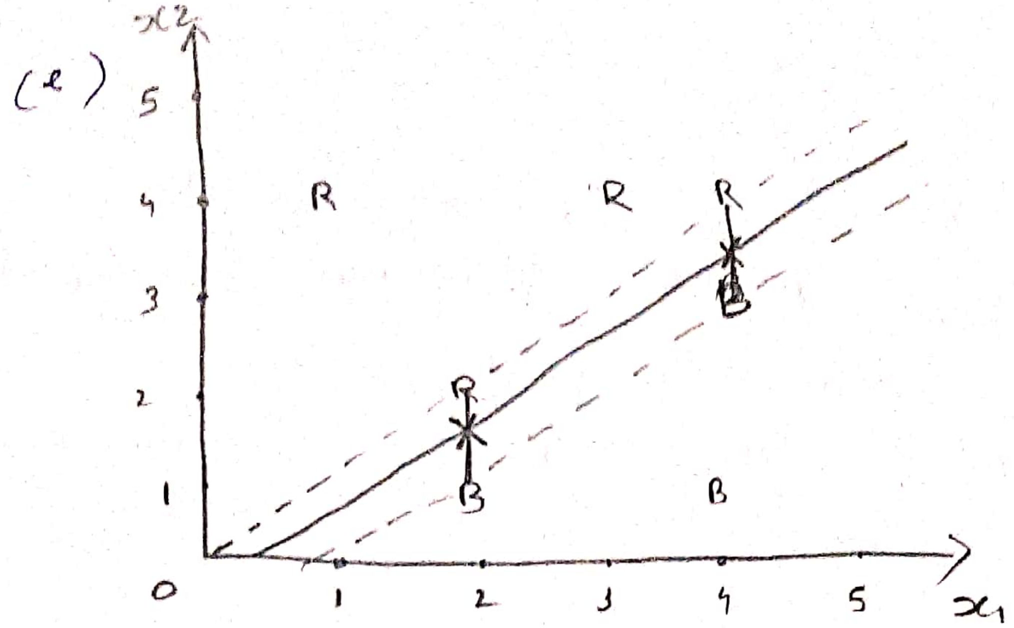
$\beta_1 = -1$

$\beta_2 = 1$

(d)



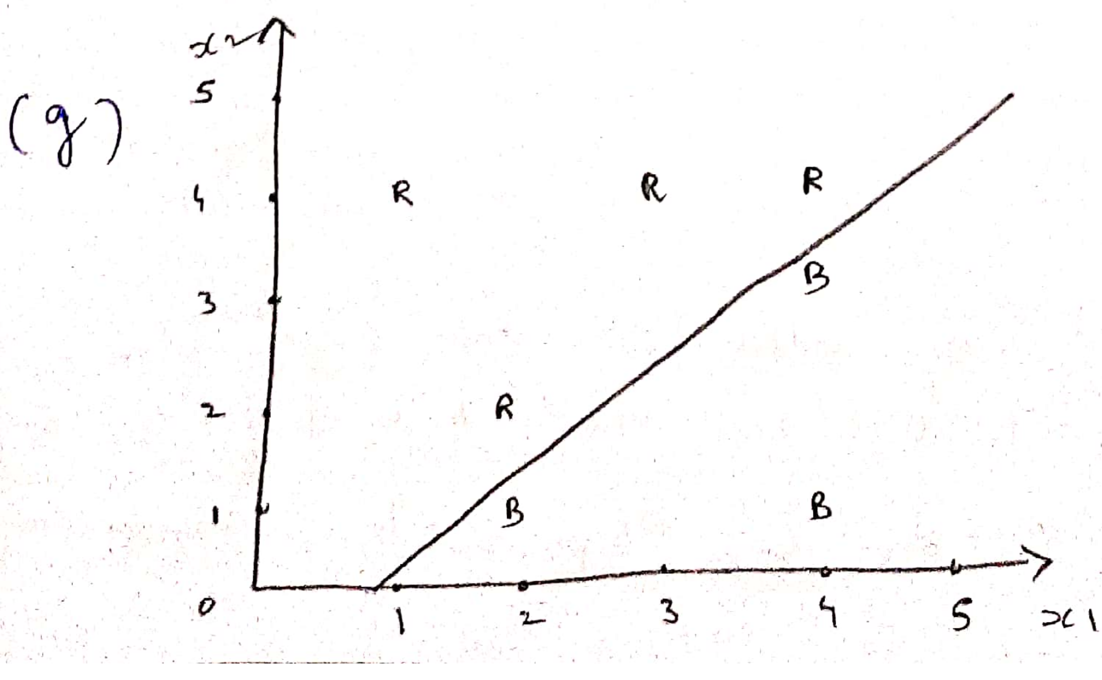
(e)



(f) Slight movement of the 7th observation (4,1) will not affect the hyperplane since it is located outside the margin.

Only the points located within the margin will have any affect on the hyperplane

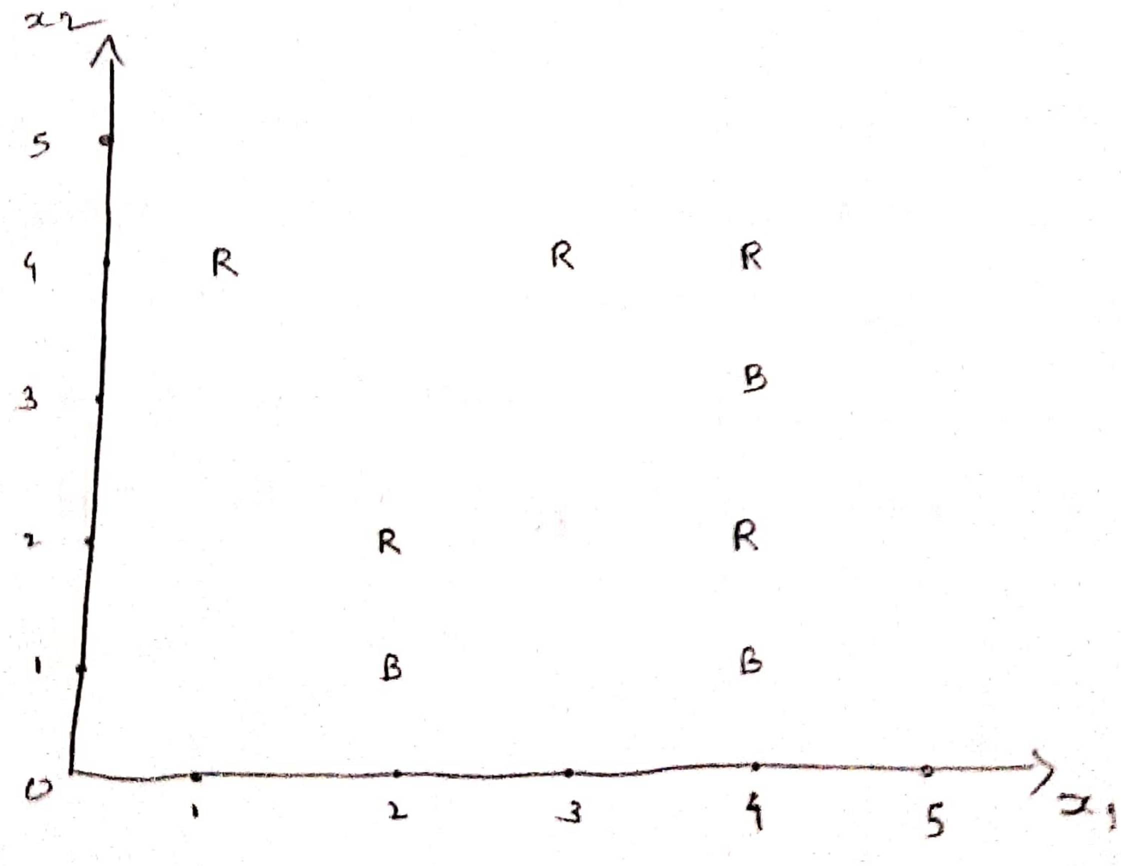
(g)



$-0.8 - x_1 + x_2 > 0$

As we see the separating hyperplane is more closely inclined towards the Blue data points region than the red data points region; we can say that this is an example of a non optimal hyperplane.

(h)



Added an additional data point (4,2) belonging to the red class such that the given classes are no longer separable by a hyperplane