HW 6:

3. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j|\le s$$

for a particular value of s. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

(a) As we increase s from 0, the training RSS will:
   i. Increase initially, and then eventually start decreasing in an inverted U shape.
   ii. Decrease initially, and then eventually start increasing in a U shape.
   iii. Steadily increase.
   iv. Steadily decrease.
   v. Remain constant.

(b) Repeat (a) for test RSS.

(c) Repeat (a) for variance.

(d) Repeat (a) for (squared) bias.

(e) Repeat (a) for the irreducible error.

(a) Training RSS: Steadily decrease
Because the model is more flexible when s increase.

(b) Test RSS: ii ↓↗
The model has a better fitting on test set first, then may overfitting.

(c) Var: Steadily increase
The model is more flexible, the var is higher.

(d) Bias: Steadily decrease.
More flexible, less bias.

(e) irr error: Stay constant.
Irr error is caused by uncertain noize. It will not change with model.

5. It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the lasso may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting.

Suppose that $n = 2$, $p = 2$, $x_{11} = x_{12}$, $x_{21} = x_{22}$. Furthermore, suppose that $y_1 + y_2 = 0$ and $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, so that the estimate for the intercept in a least squares, ridge regression, or lasso model is zero: $\hat{\beta}_0 = 0$.

(a) Write out the ridge regression optimization problem in this setting.

(b) Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$.

(c) Write out the lasso optimization problem in this setting.

(d) Argue that in this setting, the lasso coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are not unique—in other words, there are many possible solutions to the optimization problem in (c). Describe these solutions.

(a) minimize $_{\beta_1, \beta_2}$ $\left[y_1 - (\beta_1 x_{11} + \beta_2 x_{12})\right]^2 + \left[y_2 - (\beta_1 x_{21} + \beta_2 x_{22})\right]^2 + \lambda(\beta_1^2 + \beta_2^2)$

$\Rightarrow \min_{\beta_1, \beta_2} \left[y_1 - (\beta_1 + \beta_2) x_{11}\right]^2 + \left[y_2 - (\beta_1 + \beta_2) x_{21}\right]^2 + \lambda(\beta_1^2 + \beta_2^2)$

$\Rightarrow \min_{\beta_1, \beta_2} y_1^2 - 2y_1(\beta_1 + \beta_2) x_{11} + (\beta_1 + \beta_2)^2 x_{11}^2 + y_2^2 - 2y_2(\beta_1 + \beta_2) x_{21} + (\beta_1 + \beta_2)^2 x_{21}^2$

$= y_1^2 - 2y_1(\beta_1 + \beta_2) x_{11} + (\beta_1 + \beta_2)^2 x_{11}^2$
$+ y_1^2 - 2y_1(\beta_1 + \beta_2) x_{11} + (\beta_1 + \beta_2)^2 x_{11}^2 + \lambda(\beta_1^2 + \beta_2)^2$

$\Rightarrow \min \; -4y_1(\beta_1 + \beta_2) x_{11} + 2(\beta_1^2 + 2\beta_1\beta_2 + \beta_2^2) x_{11}^2 + \lambda(\beta_1^2 + \beta_2^2)$

(b) $f(\beta_1, \beta_2) = -4y_1(\beta_1 + \beta_2) x_{11} + 2(\beta_1^2 + 2\beta_1\beta_2 + \beta_2^2) x_{11}^2 + \lambda(\beta_1^2 + \beta_2^2)$

$\frac{\partial f(\beta_1, \beta_2)}{\partial \beta_1} = -4y_1 x_{11} + 2x_{11}^2 \times 2\beta_1 + 2x_{11}^2 \times 2\beta_2 + 2\lambda\beta_1 = 0$

$\Rightarrow \beta_1 = \frac{-4y_1 x_{11} + 4x_{11}^2 \beta_2}{4x_{11}^2 + 2\lambda} = \frac{2y_1 x_{11} - 2x_{11}^2 \beta_2}{2x_{11}^2 + \lambda} = A + B\beta_2$

Same, $\beta_2 = \frac{2y_1 x_{11} - 2x_{11}^2 \beta_1}{2x_{11}^2 + \lambda} = A + B\beta_1$

$\Rightarrow \beta_1 = A + B(A + B\beta_1) = A + AB + B^2\beta_1 \quad \Rightarrow \beta_1 = \frac{A + AB}{1 - B^2} = \beta_2$

$\Rightarrow \beta_1 = \beta_2$

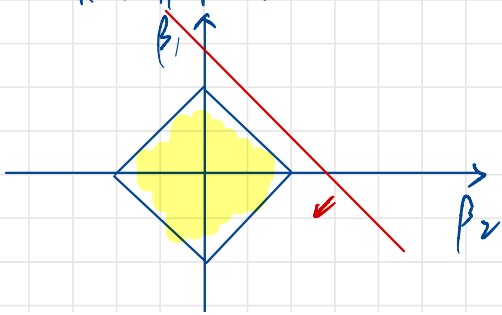
(c) minimize $_{\beta_1, \beta_2}$ $\left[y_1 - (\beta_1 x_{11} + \beta_2 x_{12})\right]^2 + \left[y_2 - (\beta_1 x_{21} + \beta_2 x_{22})\right]^2 + \lambda(|\beta_1| + |\beta_2|)$

(d) $\min \; 2[y_1 - (\beta_1+\beta_2)x_{11}]^2 \Rightarrow y_1 = (\beta_1+\beta_2)x_{11}$

s.t. $|\beta_1|+|\beta_2| \le s$

$\Rightarrow \beta_1 = -\beta_2 + \frac{y_1}{x_{11}}$

The line of β_1, β_2 will touch the s.t. area on infinite points



5. Suppose we produce ten bootstrapped samples from a data set containing red and green classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of X, produce 10 estimates of $P(\text{Class is Red}|X)$:

$$0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, \text{ and } 0.75.$$

There are two common ways to combine these results together into a single class prediction. One is the majority vote approach discussed in this chapter. The second approach is to classify based on the average probability. In this example, what is the final classification under each of these two approaches?

① Class is Red (60%, >0.5)

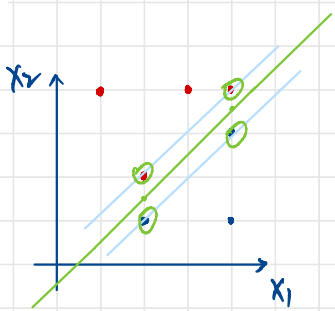
② Class is NOT Red (0.485)

3. Here we explore the maximal margin classifier on a toy data set.

(a) We are given $n = 7$ observations in $p = 2$ dimensions. For each observation, there is an associated class label.

Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

Sketch the observations.



(a) $X_2 = X_1 - 0.5$

(b) $X_2 - X_1 + 0.5 = 0$

(c) X_2 classify to red if $X_2 - X_1 + 0.5 > 0$, and classify to Blue otherwise.

(b) Sketch the optimal separating hyperplane, and provide the equation for this hyperplane (of the form (9.1)).

(c) Describe the classification rule for the maximal margin classifier. It should be something along the lines of "Classify to Red if $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$, and classify to Blue otherwise." Provide the values for β_0, β_1, and β_2.

(d) On your sketch, indicate the margin for the maximal margin hyperplane.

(e) Indicate the support vectors for the maximal margin classifier.

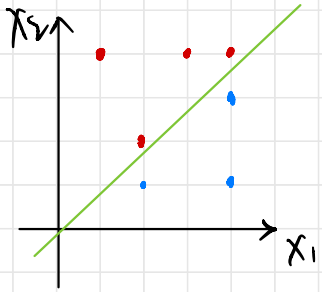
(f) Argue that a slight movement of the seventh observation would not affect the maximal margin hyperplane.

(f) Because 7th point is not on the margin or inside margin, so it will not affect the result.

(g) Sketch a hyperplane that is *not* the optimal separating hyperplane, and provide the equation for this hyperplane.

(h) Draw an additional observation on the plot so that the two classes are no longer separable by a hyperplane.

(g)



$x_2 = \frac{3}{4}x_1$

$\Rightarrow x_2 - \frac{3}{4}x_1 = 0$

(h)