Text book

5.4 ROC

Diagonal in ROC space
→ random classification.

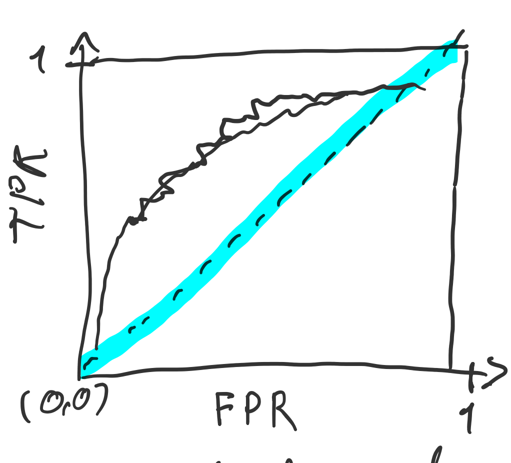
		Predicted class NO	Predicted class YES	
actual class	NO	True Negative TN	False Positive FP	①
actual class	YES	False Negative FN	True Positive TP	②

From this, we can compute

① false positive rate = FPR
$= \frac{FP}{FP+TN}$

and ② true positive rate = TPR
$= \frac{TP}{TP+FN}$

The ROC curve is then



Show that diagonal corresponds to a random classifier.

Let $Pr(Yes) = \alpha$ and $Pr(No) = 1-\alpha$, and $\alpha \in [0,1]$.

Further-more assume that the data consists of N_{NO} objects and N_{Yes} objects.
↖ # class 'NO'
↖ # class 'Yes'.

And in total: $N = N_{NO} + N_{Yes}$.

Compute:

$TN = Pr(NO) \cdot N_{NO} = (1-\alpha) N_{NO}$

$FP = Pr(Yes) \cdot N_{NO} = \alpha N_{NO}$

$FN = (1-\alpha) N_{Yes}$

$TP = \alpha N_{Yes}$

Then

$$FPR = \frac{FP}{FP+TN} = \frac{\alpha \cancel{N_{NO}}}{\alpha \cancel{N_{NO}} + (1-\alpha)\cancel{N_{NO}}} = \frac{\alpha}{\cancel{\alpha} + 1 - \cancel{\alpha}} = \alpha$$

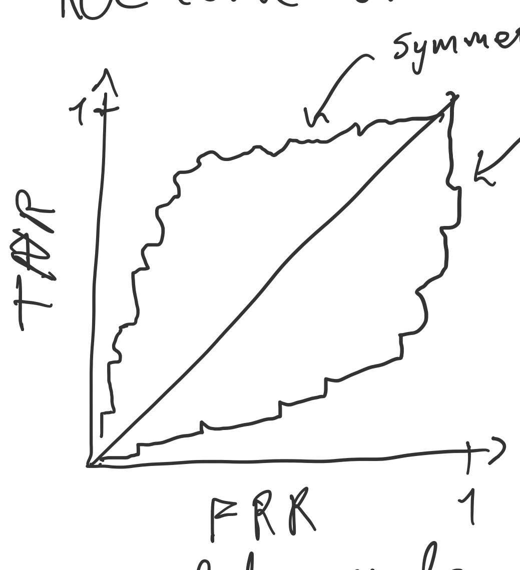
$$TPR = \frac{TP}{TP+FN} = \frac{\alpha N_{Yes}}{\alpha N_{Yes} + (1-\alpha) N_{Yes}} = \alpha$$

So, in particular, we have that a classifier based on coin-flipping will yield the point/coordinate (0.5, 0.5) in ROC-space.

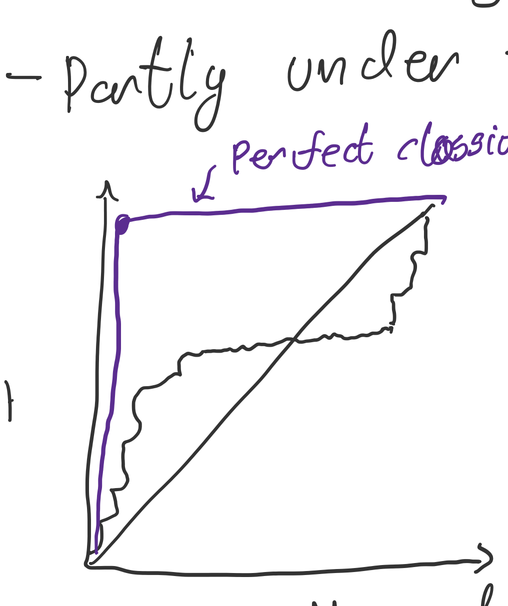
∎

5.5

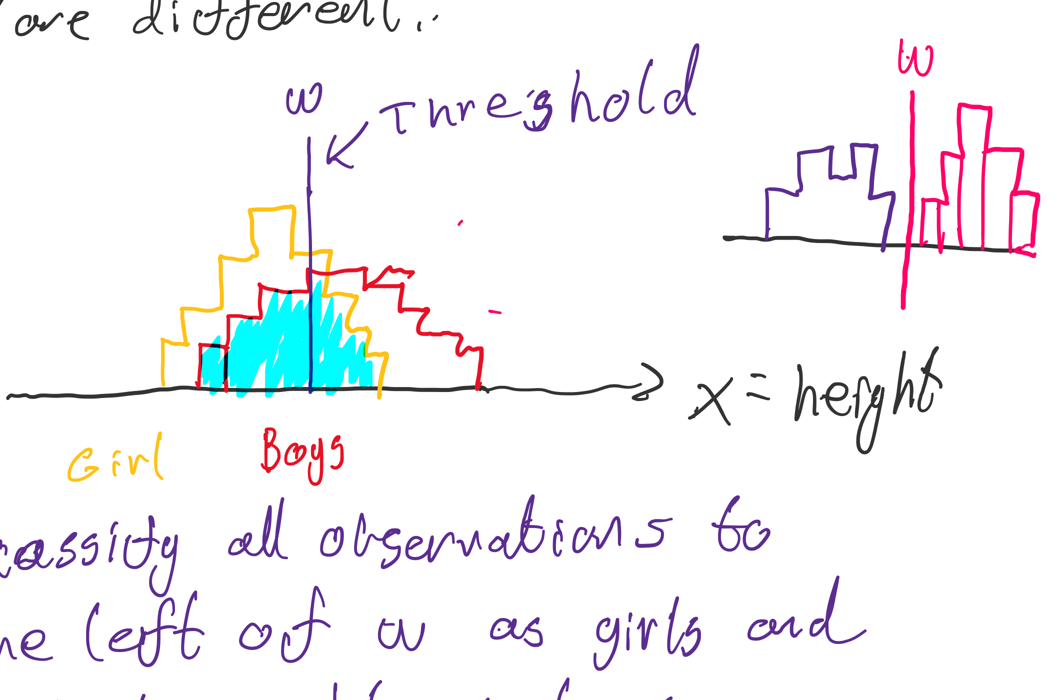
ROC curve under diagonal.



completely under the line.
- Classifier does actually a good job in distinguishing the two classes, but give them the wrong label.
  ↳ Solution: Just flip the predicted labels and then we have a good classifier.
- Partly under the line.



This can happen, and does so when the spread of the instances in class No and class Yes overlap / are different.



classify all observations to the left of w as girls and all to the right as boys.

See ROC.R code.

5.7

Prove (6.3), where

$$\eta_r(x_i) = \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_{jr},$$

using the fact that (5.2) also holds with $r=0$, by setting $\beta_{j0} = 0$, for $j = 0, \dots, p$.

Let $\sum_{k=0}^{K-1} \pi_k(x) = 1$, this means that $\boxed{\sum_{k=1}^{K-1} \pi_k(x) = 1 - \pi_0(x).}$ (*)

Assume that $\log\left(\frac{\pi_k(x)}{\pi_0(x)}\right) = \eta_k(x)$.

Take exp on both sides

$$\frac{\pi_k(x)}{\pi_0(x)} = e^{\eta_k(x)} \quad (*)$$

Take the sums over $k=1$ to $K-1$ on both sides.

$$\sum_{k=1}^{K-1} \frac{\pi_k(x)}{\pi_0(x)} = \sum_{k=1}^{K-1} e^{\eta_k(x)}$$

$$\frac{1}{\pi_0(x)} \boxed{\sum_{k=1}^{K-1} \pi_k(x)} = \sum e^{\eta_k(x)}$$

$$\frac{1-\pi_0(x)}{\pi_0(x)} \overset{(*)}{=} \text{— '' —}$$

$$\frac{1}{\pi_0(x)} - 1 = \text{— '' —}$$

$$\frac{1}{\pi_0(x)} = 1 + \sum_{k=1}^{K-1} e^{\eta_k(x)}$$

$$\pi_0(x) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\eta_k(x)}} \quad (*)$$

Done with first part.
Now we look at $\pi_k(x)$, from (*) we get

$$\frac{\pi_k(x)}{\pi_0(x)} = e^{\eta_k(x)}$$

$$\pi_k(x) = \pi_0(x) e^{\eta_k(x)}$$

$$\pi_k(x) \overset{(*)}{=} \frac{e^{\eta_k(x)}}{1 + \sum_{j=1}^{K-1} e^{\eta_j(x)}}.$$

This is (5.3), so we are done.

If we set $k=0$, we must have that $e^{\eta_0(x)} = 1$, take log

$$\eta_0(x) = \log(1) = 0$$

$$\beta_0 + \sum_{j=1}^{p} x_{ij} \beta_{j0} = 0$$

which is what the exercise states.

∎

ISLR

Ex. 7.5

Consider two curves $\hat{g}_1$ and $\hat{g}_2$ defined by

$$\hat{g}_1 = \arg\min_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int [g^{(3)}(x)]^2 dx \right)$$

$$\hat{g}_2 = \text{—————— '' ——} + \lambda \int [g^{(4)}(x)]^2 dx)$$

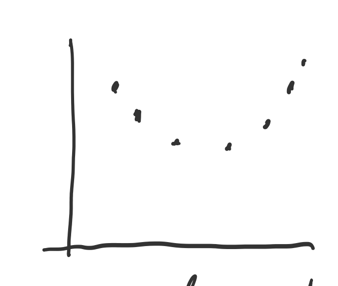
a) as $\lambda \to \infty$, will $\hat{g}_1$ or $\hat{g}_2$ have smaller training RSS?

when $\lambda \to \infty$, $\hat{g}_1 = ax^2 + bx + c$ (quadratic)
$\boxed{\hat{g}_1^{(3)} = 0}$

$g_1 = ax^3 + bx^2 + cx + d$
$g_1' = 3ax^2 + 2bx + c$
$g_1'' = 6ax + 2b$
$g_1''' = 6a$

while $\hat{g}_2 = ax^3 + bx^2 + cx + d$ (cubic)

g_2 more flexible, ⇒ fit data better → lower training RSS.



b) what about test RSS?

We don't know.
Depends on if the true relationship between x and y is best approximated by a cubic or quadratic function, and depends on the noise level of data.
It is possible that $\hat{g}_2$ could over fit and yield a large test RSS.
$\hat{g}_1$ could also underfit the data → large test RSS.

c) Let $\lambda = 0$. what happen with training and test RSS
↳ No penalty and this means that $\hat{g}_1 = \hat{g}_2$ → same train and test RSS.

We get that

$$\hat{g} = \hat{g}_1 = \hat{g}_2 = \arg\min_g \sum_{i=1}^{n} (y_i - g(x_i))^2$$

We can get a loss of by using a high enough order on the polynomial

We have no restrictions on $\hat{g}$, so it can be any polynomial.