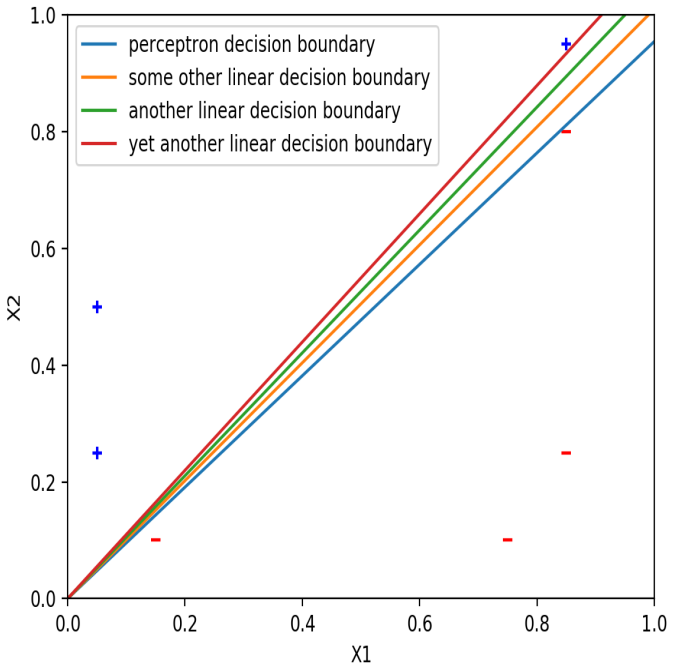
Homework 2

Xixiang Chen

1 Classifiers for Basketball Courts

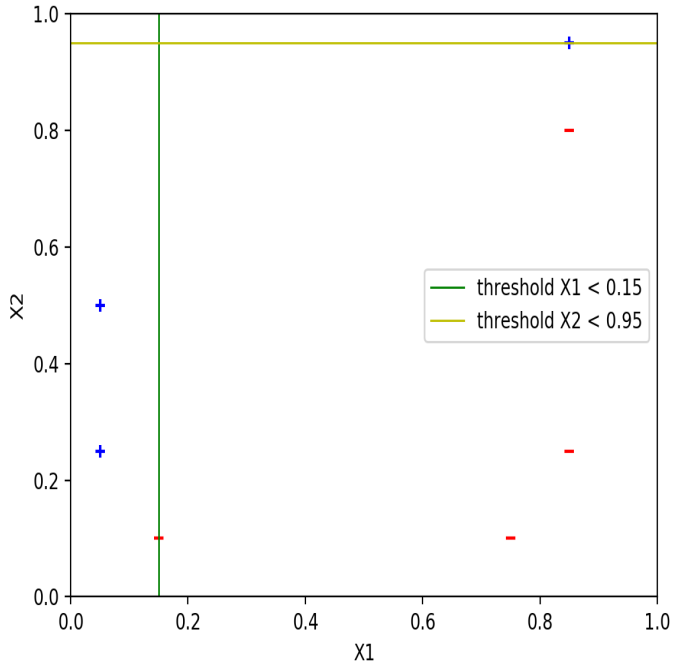
(a)

The Perceptron algorithm takes three iterations to converge. There is no error from the classifier. There are infinite many other linear classifiers that give the same result. Some of them are illustrated in the figure below. In general, any line $X_2 = kX_1$ where $\frac{0.80}{0.85} < k < \frac{0.95}{0.85}$ will work.

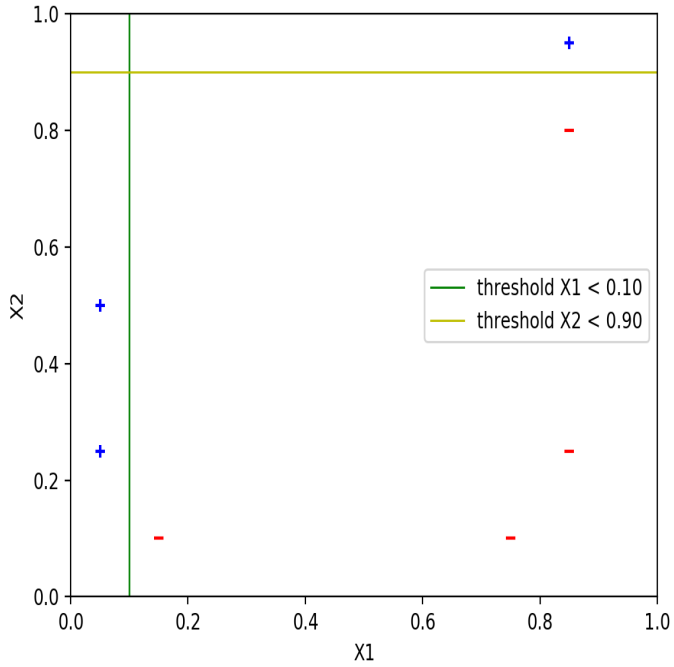


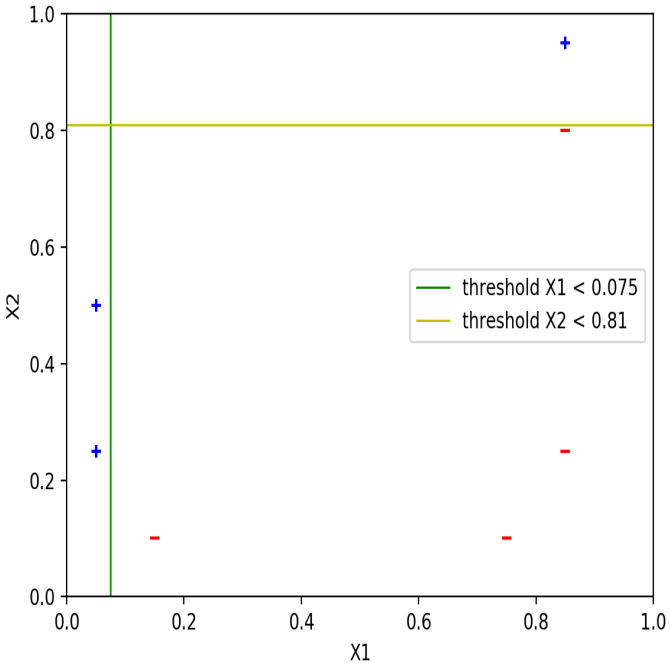
(b)

The decision boundaries of the decision tree are illustrated below.



There is no error in these decision boundaries. Some other decision boundaries giving the same errors are shown below. In general any thresholds (s_1, s_2) where $0.05 < s_1 < 0.15$ and $0.80 < s_2 < 0.95$ will work.



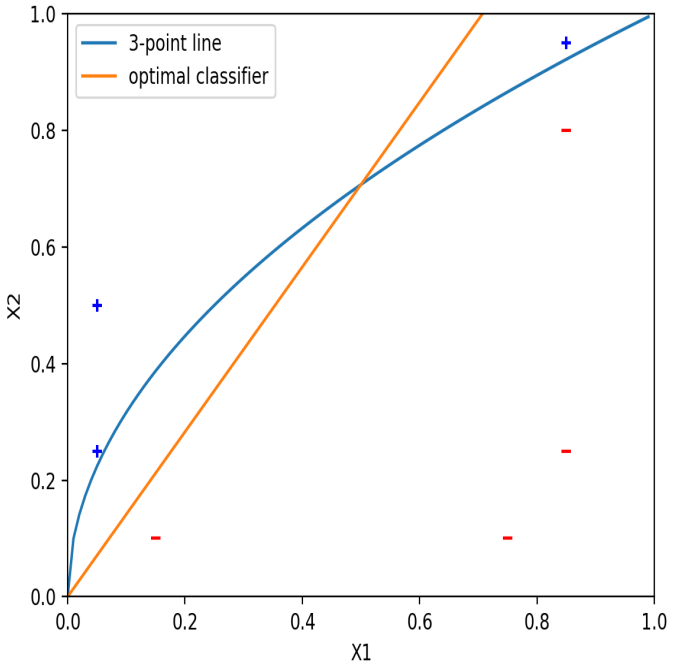


(c)

To minimize the true risk, we want to minimize the area between the two function $f_1(x) = kx$ and $f_2(x) = \sqrt{x}$.

$$\begin{aligned} Area &= \int_0^{\frac{1}{k^2}} (\sqrt{x} - kx)dx + \int_{\frac{1}{k^2}}^1 (kx - \sqrt{x})dx - \frac{1}{2}(1 - \frac{1}{k})(k-1) \\ &= \frac{1}{6}k^{-3} + \frac{1}{2}k - \frac{2}{3} + \frac{1}{6}k^{-3} - (k - 2 + \frac{1}{k}) \\ &= \frac{1}{3}k^{-3} - \frac{1}{2}k - \frac{1}{k} + \frac{4}{3} \end{aligned}$$

Therefore min(Area) for $k > 0$ gives $k = \sqrt{2}$.
So the theoretical optimal classifier that passes origin is $w = (-1, \sqrt{2})$ as shown in the figure below, and the error (area in between) is 0.0369709011. It does not achieve the minimum empirical loss in part (a) as it misclassifies one point.

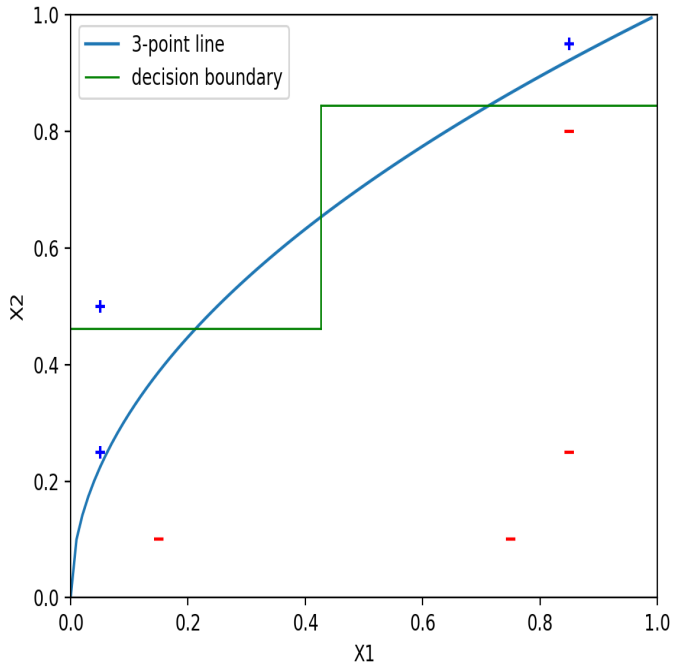


(d)

To minimize the true risk, we want to minimize the area between the decision boundary and the 3-point line where it is misclassified.
For Tree 1:

$$\begin{aligned} Area1 &= \int_0^{s_2^2} (s_2 - \sqrt{x})dx + \int_{s_2^2}^{s_1} (\sqrt{x} - s_2)dx + \int_{s_1}^{s_3^2} (s_3 - \sqrt{x})dx + \int_{s_3^2}^1 (\sqrt{x} - s_3)dx \\ &= \frac{2}{3}s_3^3 + \frac{2}{3}s_2^3 + \frac{4}{3}s_1^{\frac{3}{2}} - s_2 s_1 - s_3 s_1 - s_3 + \frac{2}{3} \end{aligned}$$

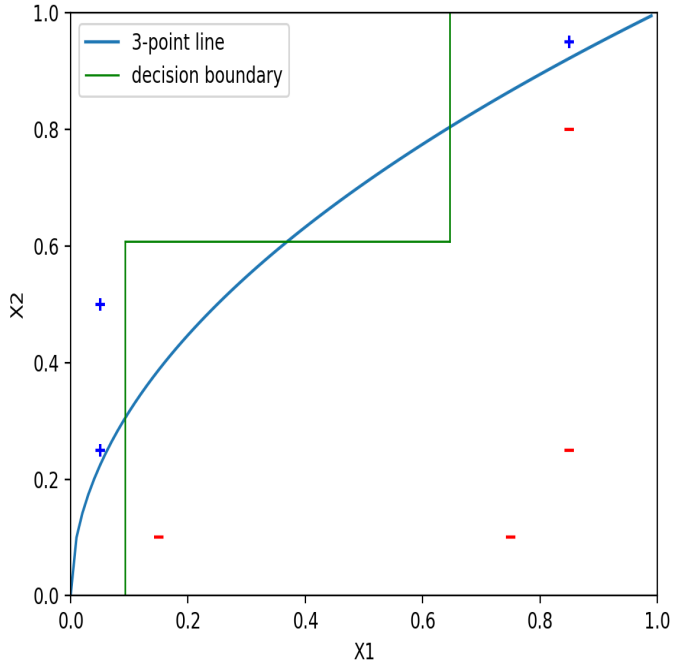
Therefore argmin(Area1) for $s_1, s_2, s_3 > 0$ gives ($s_1 = 0.426777, s_2 = 0.46194, s_3 = 0.844623$). The error is 0.1036. It does not achieve the minimum empirical loss in part (b). One point is misclassified.



For Tree 2:

$$\begin{aligned} Area2 &= \int_0^{s_2} \sqrt{x}dx + \int_{s_2}^{s_1^2} (s_1 - \sqrt{x})dx + \int_{s_1^2}^{s_3} (\sqrt{x} - s_1)dx + \int_{s_3}^{1} (1 - \sqrt{x})dx \\ &= \frac{2}{3}s_1^3 + \frac{4}{3}s_2^{\frac{3}{2}} + \frac{4}{3}s_3^{\frac{3}{2}} - s_1 s_2 - s_1 s_3 - s_3 - \frac{1}{3} \end{aligned}$$

Therefore argmin(Area2) for $s_1, s_2, s_3 > 0$ gives $(s_1 = 0.607625, s_2 = 0.0923021, s_3 = 0.646115)$. The error is 0.1180. It does not achieve the minimum empirical loss in part (b). One point is misclassified.



(e)

Let $f(X_1, X_2) = (X_1, X_2^2)$ be the tranformation function. Then the optimal classifier $w = (-1, 1)$ that minimizes the true risk can achieve no error.

(f)

No, because the decision boundary is always vertical or horizontal and will intercept the 3-point line (which will always pass through origin) no matter the transformation thus difference between the transformed 3-point line and the decision boundary.

(g) [This part was never included]

(h)

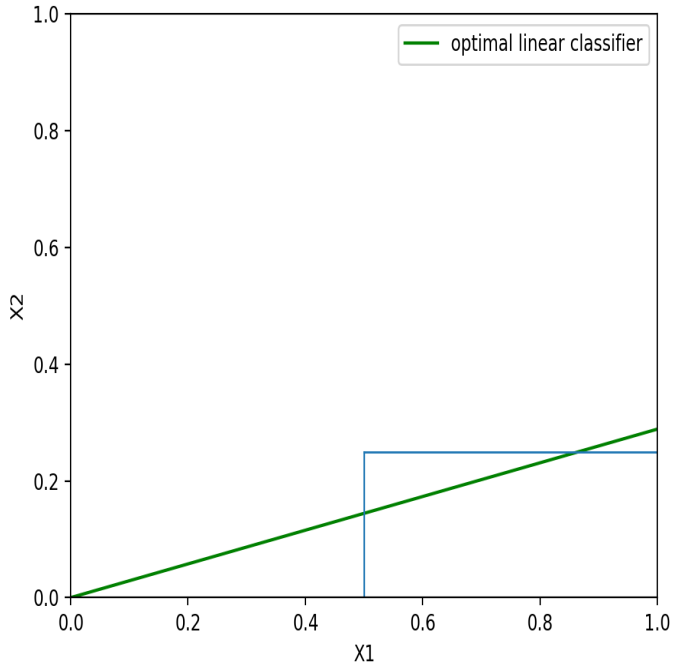
Let $f(x) = kx$,
Case 1: $0 < k \leq 0.25$

$$\begin{aligned} Area &= \frac{1}{2}\frac{1}{2}\frac{k}{2} + \frac{1}{2}\frac{1}{2}(\frac{1}{4} - \frac{k}{2} + \frac{1}{4} - k) \\ &= -\frac{k}{4} + \frac{1}{8} \end{aligned}$$

$\min(\text{Area}) = 0.0625$ when $k = 0.25$.
Case 2: $0.25 < k \leq 0.5$:

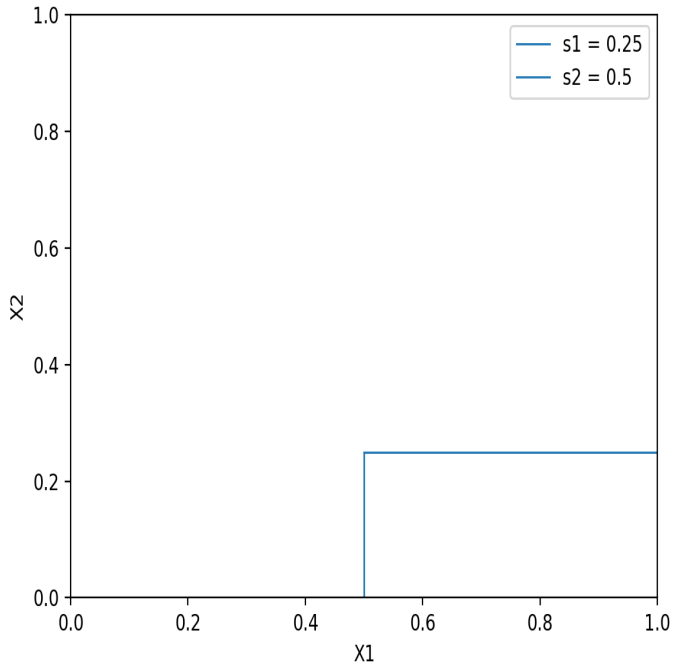
$$\begin{aligned} Area &= \frac{1}{2}\frac{1}{2}\frac{k}{2} + \frac{1}{2}(\frac{1}{4} - \frac{k}{2})(\frac{1}{4k} - \frac{1}{2}) + \frac{1}{2}(1 - \frac{1}{4k})(k - \frac{1}{4}) \\ &= \frac{3}{4}k + \frac{1}{16k} - \frac{3}{8} \end{aligned}$$

$\min(\text{Area}) = 0.0580127$ when $k = 0.288675$.
Therefore the optimal classifier that passes through the origin is of the form $f(x) = 0.288675x$.



(i)

The optimal decision tree will just have its decision boundary the same as the boundary of the paint. It has 0 true risk.
$X_2 \geq 0.25 \implies +1$,
$X_2 < 0.25, X_1 \geq 0.5 \implies -1$
$X_2 < 0.25, X_1 < 0.5 \implies +1$



2 Variable Importance for trees and random forests

(a)

Program output:

```
X1 == 1 , sum of children's Gini Index=0.2290334502
X2 == 1 , sum of children's Gini Index=0.3937897702
X3 == 1 , sum of children's Gini Index=0.4987555321
X4 == 1 , sum of children's Gini Index=0.4987969689
X5 == 1 , sum of children's Gini Index=0.4992357851
Split on X1 == 1 with sum of children's Gini Index = 0.2290334502
p-Left: 0.87549
p-right: 0.13992
Best split Delta Gini Index = 0.2703185498
surrogate split on X2
X2 == 1 , sum of children's Gini Index=0.3937897702
Surrogate split Delta Gini Index = 0.1055622298
Least-square error for best split: 0.1000000000
Least-square error for surrogate split: 0.2700000000
```

(i)

Best split decision stump: $X_1 = 1$ *predicts* 1; $X_1 = 0$ *predict* 0.

Best surrogate split decision stump: $X_2 = 1$ *predicts* 1; $X_2 = 0$ *predict* 0.

(ii)

For equation 2:

$$Imp^T(X_1) = 0.2703185498$$
$$Imp^T(X_2) = 0$$
$$Imp^T(X_3) = 0$$
$$Imp^T(X_4) = 0$$
$$Imp^T(X_5) = 0$$

For equation 3:

$$Imp_s^T(X_1) = 0.2703185498$$
$$Imp_s^T(X_2) = 0.1055622298$$
$$Imp_s^T(X_3) = 0$$
$$Imp_s^T(X_4) = 0$$
$$Imp_s^T(X_5) = 0$$

The result suggests that X_1 and X_2 are more important than the others.
(iii)
Least square error for best split is 0.1
Least square error for surrogate split is 0.27

(b)

(i)

	X_1	X_2	X_3	X_4	X_5
best split count	2961	1035	375	361	268
surrogate split count	0	1976	705	440	879

The result suggests that X_1 and X_2 are more important than the others.
The counts on variable X_3, X_4, X_5 are mainly due to the cases where K=1. When K increase, the more important variables are more likely to be chosen and split on.
(ii)

	X_1	X_2	X_3	X_4	X_5
variable importance(5)	0.27036753	0.10609478	0.00129859	0.00120952	0.0006835
variable importance(6)	0.36266464	0.22315942	0.00381333	-0.00149584	0.00347015

Again the result suggests that X_1 and X_2 are more important than the others. The impact of masking is lessen because every time a variable X_j is chosen as one of the K sample features, the variable that would have masked X_j might not appear in those K sample features in this round, thus X_j is not masked by that variable (unless when K = 5 where every feature is chosen).
(iii)
Using first method: loss = 0.1
Using second method: loss = 0.217834
The first method is correct because the random forest uses majority vote of the trees to predict.

(c)

(i)

	X_1	X_2	X_3	X_4	X_5
B = 200					
equation(5)	0.27345069	0.10667681	0.00461637	0.00506484	0.00414359
equation(6)	0.36496683	0.23098361	0.0086532	-0.01915152	0.0027381
B = 250					
equation(5)	0.27119652	0.10734508	0.00333603	0.00358556	0.00310326
equation(6)	0.36746	0.22726073	0.00448696	-0.00185263	0.00289655
B = 300					
equation(5)	0.27142288	0.10513424	0.00292945	0.00313381	0.00217645
equation(6)	0.36371703	0.2293	0.00428571	-0.00199074	0.00042857
B = 350					
equation(5)	0.27110697	0.1062295	0.00186875	0.00200137	0.00129111
equation(6)	0.36133676	0.22735849	0.00086207	0.00148148	-0.0026087
B = 400					
equation(5)	0.27101965	0.1057246	0.00161503	0.00135715	0.00116307
equation(6)	0.36131188	0.22752412	-0.00594595	-0.00367521	-0.02912281

The result suggests that changing the bootstrap sample size does not affect the variable importance significantly because we sample the data uniformly at random.
(ii)

	X_1	X_2	X_3	X_4	X_5
B = 200					
equation(5)	0.02808278	0.02148381	0.0042129	0.00504086	0.0038634
equation(6)	0.03248441	0.03350418	0.03379452	0.0359445	0.03606344
B = 250					
equation(5)	0.02136291	0.01806184	0.00326645	0.0029634	0.00342651
equation(6)	0.0366064	0.03712411	0.0420125	0.038181	0.04305437
B = 300					
equation(5)	0.01916232	0.01515004	0.00297868	0.00277809	0.00213636
equation(6)	0.04012686	0.0466513	0.04549426	0.05152732	0.050212
B = 350					
equation(5)	0.01454407	0.01176199	0.00164236	0.00201735	0.00150091
equation(6)	0.04574604	0.05095534	0.05373791	0.05457907	0.06225207
B = 400					
equation(5)	0.01159907	0.00971624	0.00136445	0.00113526	0.00132709
equation(6)	0.06024711	0.0649064	0.05928065	0.06306418	0.07498025