Exercise 6.1:

Part a:

From textbook we know:

CosSim(a,b) = $\frac{a*b}{||a||*||b||}$

-1 <= CosSim(a,b) <= 1, and larger CosSim(a,b) means higher similarity.

d(a,b) = ||a - b||

d(a,b) >= 0, smaller d(a,b) means higher similarity.

For high cosine similarity and low Euclidean distance similarity:

a = (1,1) b = (10, 10)

CosSim(a,b) = $\frac{a*b}{||a||*||b||}$ = $\frac{20}{\sqrt{2}*\sqrt{200}}$ = 1

d(a,b) = $\sqrt{162}$

For low cosine similarity and high Euclidean distance similarity:

a = (1,1) b = (-1, -1)

CosSim(a,b) = $\frac{-2}{2}$ = -1

d(a,b) = $\sqrt{8}$

Part b:

If the origin changes, Cosine similarity will be affected, but Euclidean distance similarity won't be affected as the distance between 2 points won't change. Therefore, we should choose Euclidean distance similarity

Exercise 6.2:

When $\pi(x) >= 0.5$, f(x) = 1

e(f(x)) = P|f(x) != y| = P|x = 1|P|y = -1| = 1 - $\pi(x)$ = min($\pi(x)$, 1 - $\pi(x)$)

When $\pi(x) < 0.5$, f(x) = -1

e(f(x)) = P|f(x) != y| = P|x = -1|P|y = 1| = $\pi(x)$ = min($\pi(x)$, 1 - $\pi(x)$)

e(h(x)) = P|h(x) = 1|*P|y = -1| + P|h(x) = -1|*P|y = 1|

= P|h(x) = 1|*(1 - $\pi(x)$) + P|h(x) = -1|*$\pi(x)$

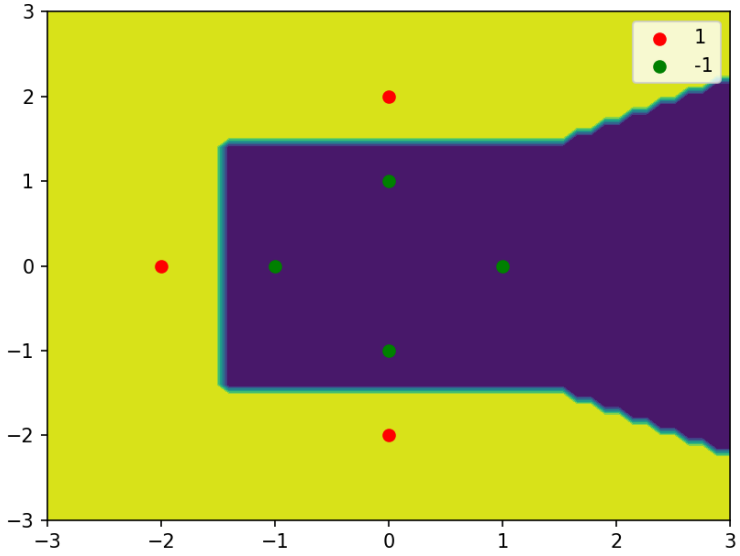
>=P|h(x) = 1|*min($\pi(x)$, 1 - $\pi(x)$) + P|h(x) = -1|*min($\pi(x)$, 1 - $\pi(x)$)

So e(h(x)) >= e(f(x))

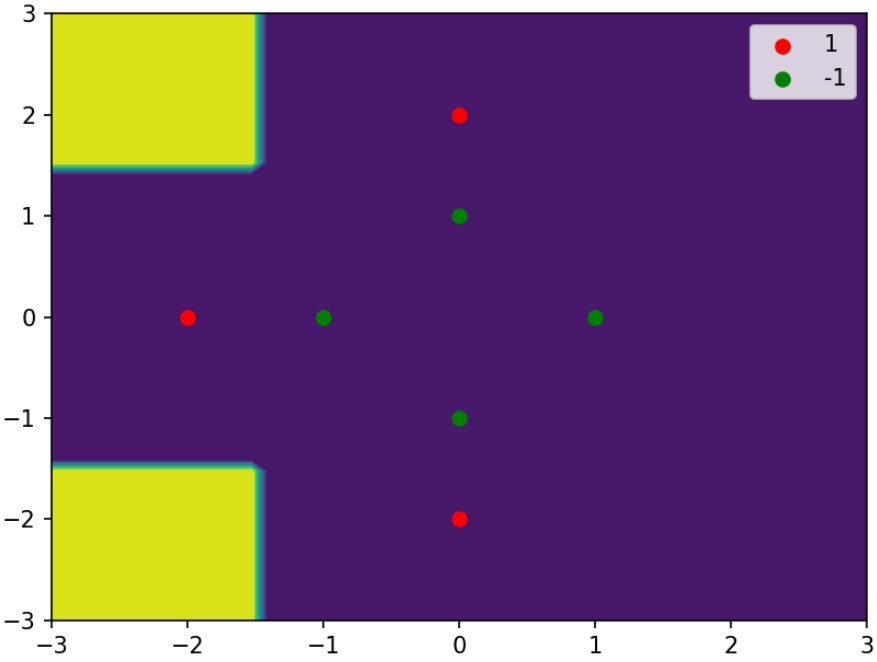
Problem 6.1

Part a:

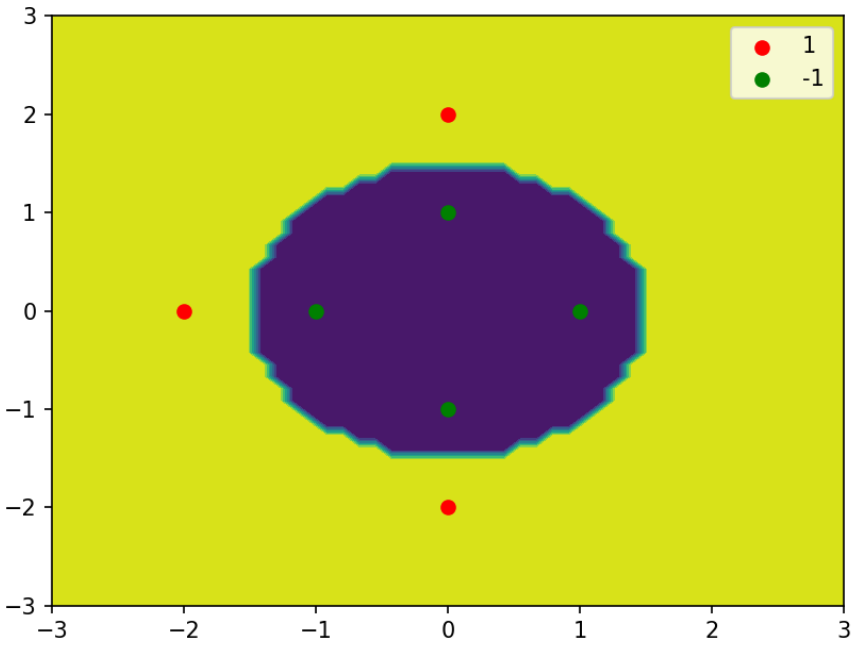
1-NN:



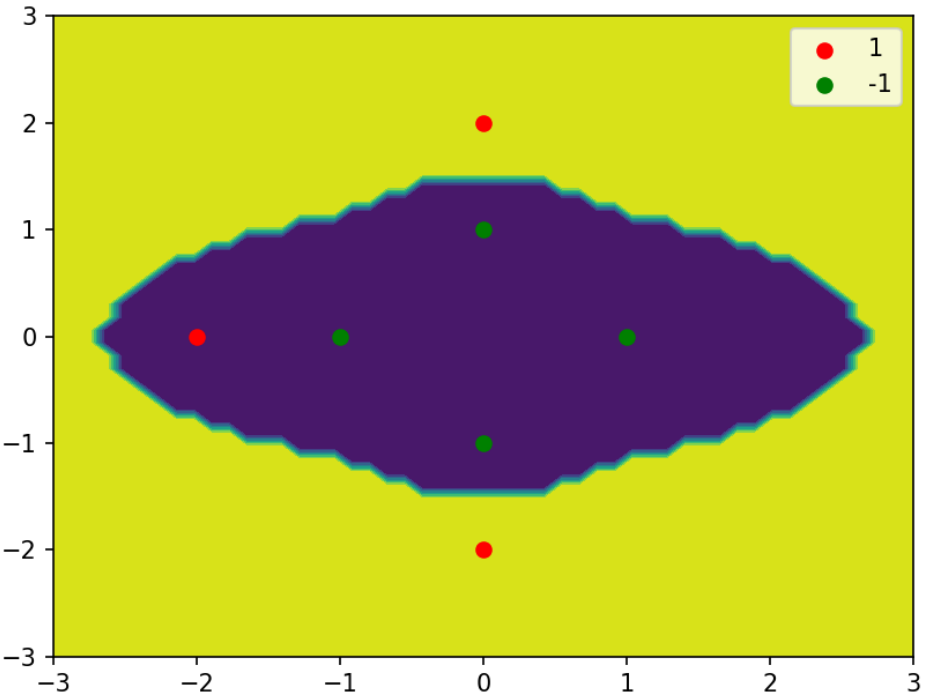
3-NN:



Part b:
1-NN:

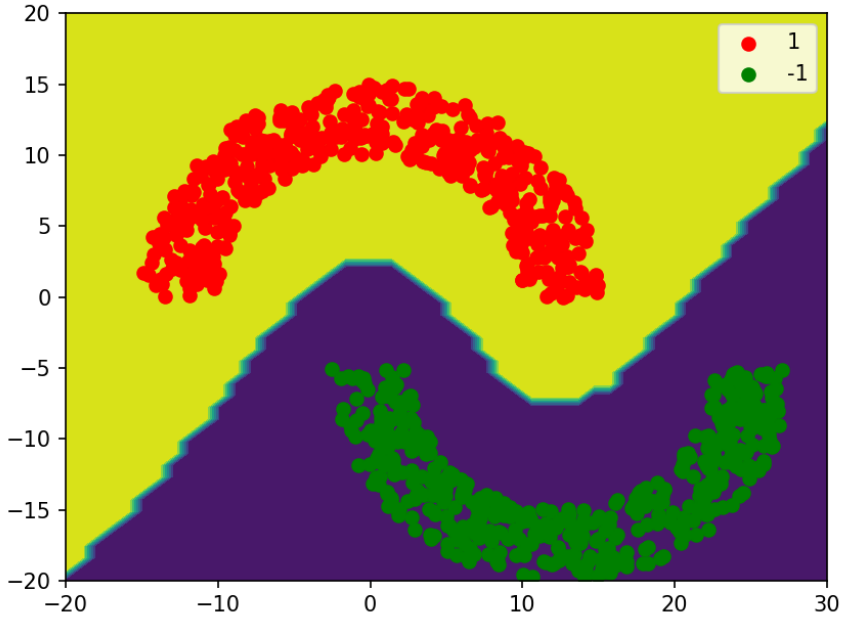


3-NN:

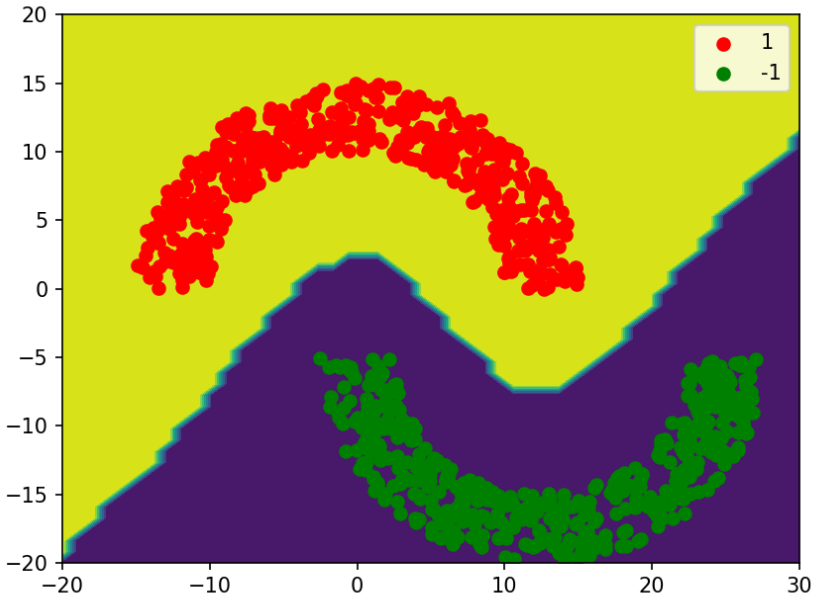


6.4:

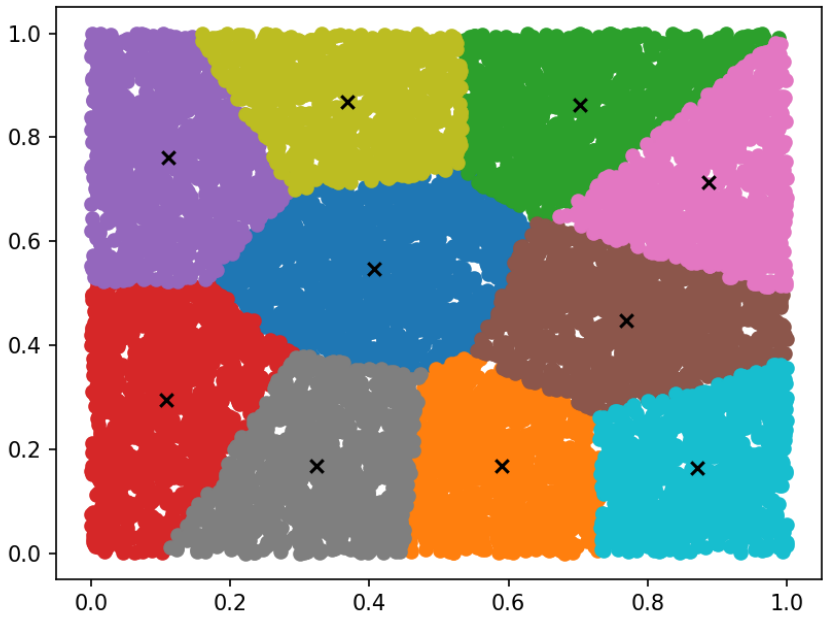
1-NN:



3-NN:

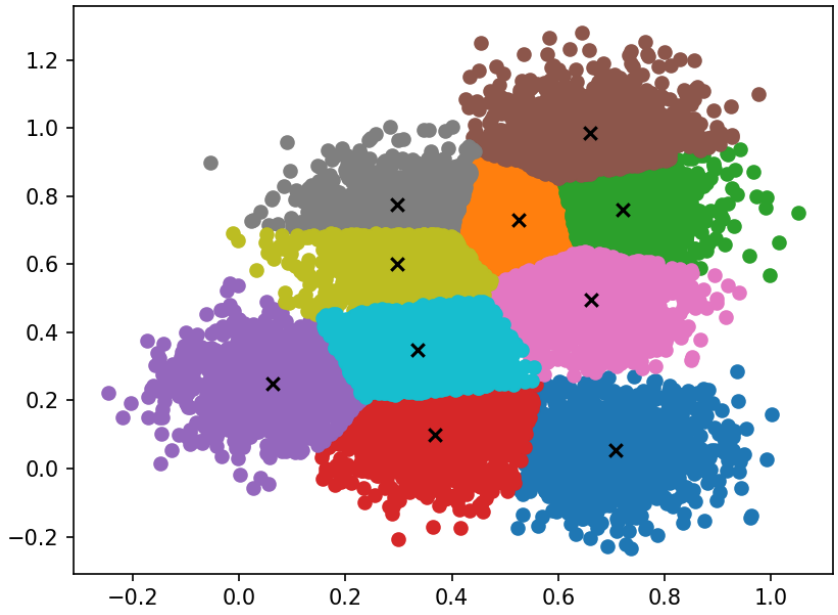


Problem 6.16:
Part a:
i:



II: Running time for brute force: 51.523218393325806
Running time using the partition with branch and bound: 5.334736108779907

Part b:



Running time for brute force: 76.08479404449463
Running time using the partition with branch and bound: 7.703598976135254

Part c:

From comparing the running time between the brute force method and the partition branch and bound method, we can see that the partition branch and bound method is approximately 10 times faster than using brute force method. This is because in the partition branch and bound method we divide points into 10 different clusters already so we only need to run 10% of data set in the brute force method.

Part d:
No, because the partition branch and bound method is always running faster than the brute force method, except when the data set is very small, because we need to cluster first., but when the data set is very small the running time won't make any difference. As the data set becomes larger, the partition branch and bound method will save more time.