Exercise 6.1

a)

$v1 = (5, 1)^T$

$v2 = (25, 5)^T$

$CosSim(v1, v2) = 1$ since the angle between them is 90 degrees. This pair of vectors have high similarity under cosine similarity.

$d(v1, v2) = ||v1 - v2|| \approx 20.396$. This pair of vectors have low similarity under Euclidean distance.

$v1 = (1, 0)^T$

$v2 = (0, 1)^T$

$CosSim(v1, v2) = -1$ since the angle between them is 180 degrees. This pair of vectors have low similarity under cosine similarity.

$d(v1, v2) = ||v1 - v2|| \approx 1.414$. This pair of vectors have high similarity under Euclidean distance.

b)

Assume the new origin is o.

$d(v1, v2) = ||(v1 - o) - (v2 - o)|| = ||v1 - v2||$ for euclidean distance.

$CosSim(v1, v2) = \frac{(v1-o)\bullet(v2-o)}{||(v1-o)||||(v2-o)||}$

Hence, the cosine similarity will be changed once the origin of the coordinate system changes. If I am using Euclidean distance, this will not affect my choice of feature. However, if I am using the cosine similarity, this will affect my choice of feature.

Exercise 6.2

We have $\pi(x) = P[y = 1|x]$

If $\pi(x) \geq \frac{1}{2}$, $f(x) = 1$

$P[f(x) \neq y] = P[f(x) = -1]$

$P[f(x) \neq y] = 1 - \pi(x)$

Since $\pi(x) \geq \frac{1}{2}$, $1 - \pi(x) \leq \pi(x)$

$P[f(x) \neq y] = min\{\pi(x), 1 - \pi(x)\}$

If $\pi(x) < \frac{1}{2}$, $f(x) = -1$

$P[f(x) \neq y] = P[f(x) = 1]$

$P[f(x) \neq y] = \pi(x)$

Since, $\pi(x) < \frac{1}{2}$, $\pi(x) > 1 - \pi(x)$

$P[f(x) \neq y] = min\{\pi(x), 1 - \pi(x)\}$

Hence, we can conclude that $e(f(x)) = P[f(x) \neq y] = min\{\pi(x), 1 - \pi(x)\}$

Assume for $e(h(x))$, $P[h(x) = 1] = p$ and $P[h(x) = -1] = 1 - p$

$e(h(x)) = P[h(x) = 1]P[y = -1|x] + P[h(x) = -1]P[y = 1|x]$

$= p(1 - \pi(x)) + (1 - p)\pi(x)$

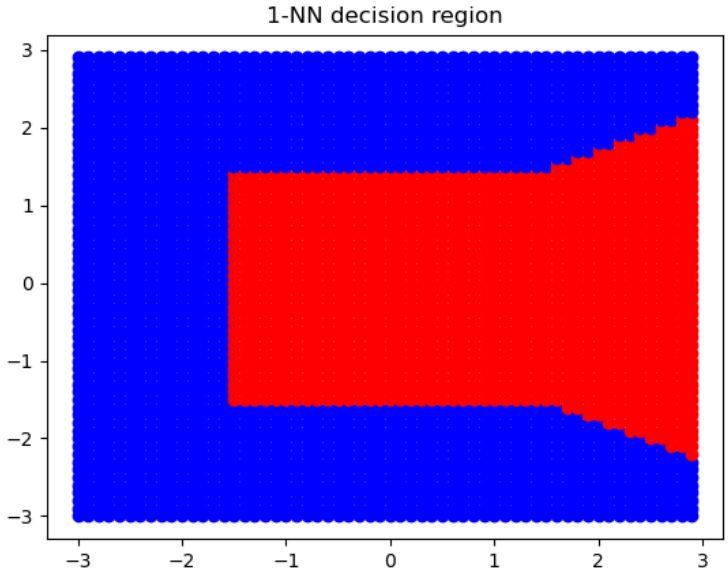
$\leq pmin\{\pi(x), 1 - \pi(x)\} + (1 - p)min\{\pi(x), 1 - \pi(x)\} = e(f(x))$

Hence, the statement is true.

Problem 6.1

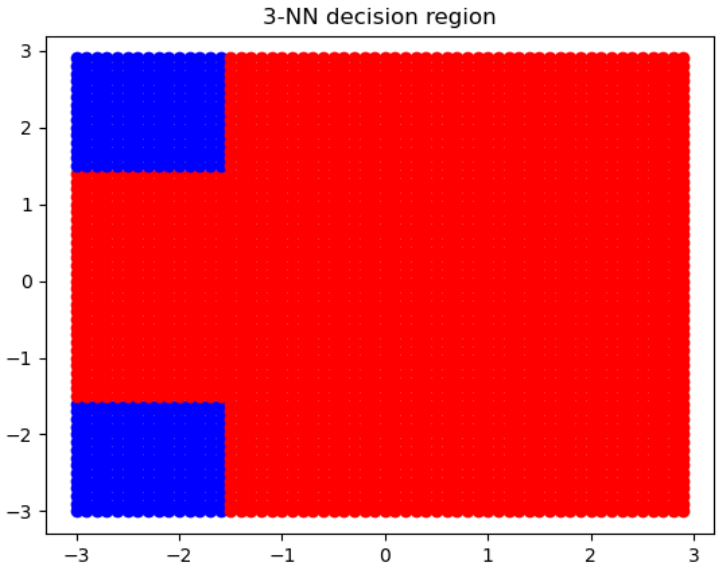
a)

1-NN



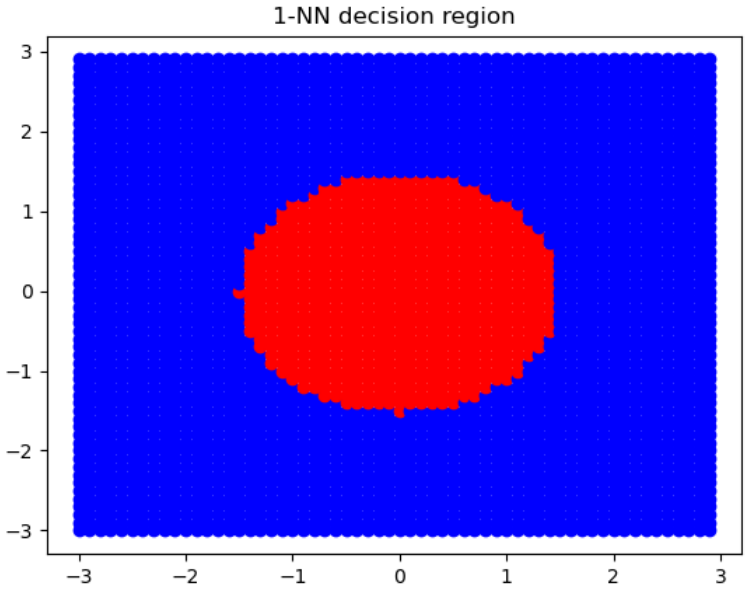

The blue region is the +1 region, while the red region is the -1 region.

3-NN



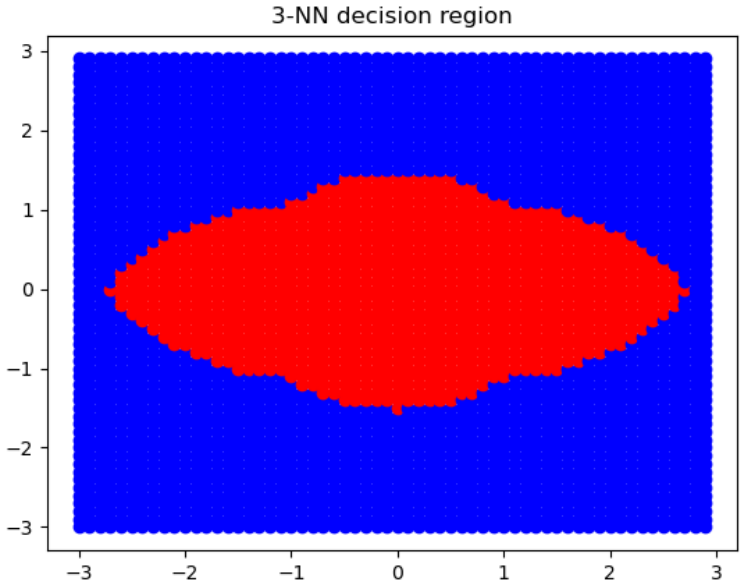

The blue region is the +1 region, while the red region is the -1 region.

b)
1-NN



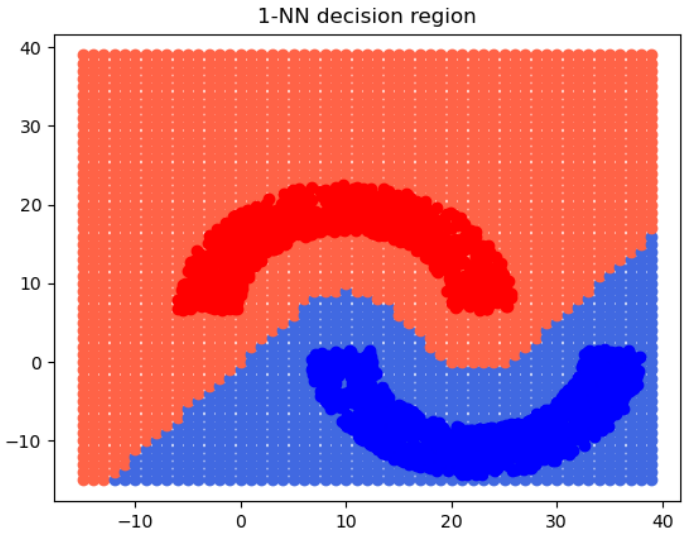

The blue region is the +1 region, while the red region is the -1 region.

3-NN



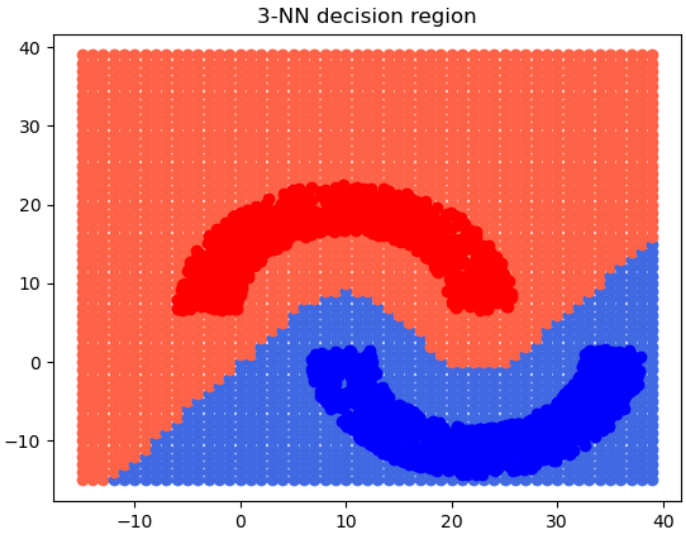

The blue region is the +1 region, while the red region is the -1 region.

Problem 6.4
1-NN




The lighter blue is the +1 region and the darker blue is the points generated.
The lighter red is the -1 region and the darker red is the points generated.

3-NN




The lighter blue is the +1 region and the darker blue is the points generated.
The lighter red is the -1 region and the darker red is the points generated.

Problem 6.16
a)
i) included in the code problem6.16a.py
ii)
Brute force time: 507.191s
Branch and bound time: 194.644s

b)
problem6.16b.py
Brute force time: 508.394s
Branch and bound time: 123.159s

c)
As we can see, for brute force, a)'s running time and b)'s running time are almost the same since all of them will cost $O(n^2)$. For each query point, we need to go through all data in the data set in order to compute the nearest neighbor.

For branch and bound, b) perform slightly better than a). The cause of this difference is due to the bound condition stated in the book. In a uniform distributed data set, each cluster has relatively large radii. In a gaussian distributed data set, since we set each of the gaussian bumps to have standard deviation equal to 0.1, we will gain a smaller radius for each cluster. As a result, the gaussian distribution is more well separated than the uniform distribution, which results in better computational saving.

d)
No.
If we want to let branch and bound perform better, we need the clusters to be well separated and have small radii in order to satisfy the bound condition.

Hence, the decision to use the branch and bound does not depend on the size of the test point to evaluate; it should depend on how well the existing clusters are separated.