CSCI 4100 Fall 2018
Assignment 10 Answers

Damin Xu
661679187

November 19, 2018

Exercise 6.1

(a) 1. Two vectors with very high cosine similarity but very low Euclidean distance similarity:
vector1: [1, 1], vector2: [1000, 1010],
Eucildean Distance: 1419.888, Cosine Similarity: 0.9999876

2. Two vectors with very low cosine similarity but very high Euclidean distance similarity:
vector1: [0.1, 0.1], vector2: [-0.1, -0.1],
Eucildean Distance: 0.2828, Cosine Similarity: -1

(b) When the origin changes, cosine similarity changes only a little bit, but Euclidean distance similarity keeps the same. This may not affect y choice of features.

Exercise 6.2 If f(x) $\geq \frac{1}{2}$ and $y = -1$,

$$\begin{aligned} e(f(x)) &= P[f(x) \neq y] \\ &= P[f(x) = +1, y = -1] + P[f(x) = -1, y = +1] \\ &= 1 \times (1 - \pi(x)) + 0 \times \pi(x) \\ &= 1 - \pi(x) \end{aligned}$$

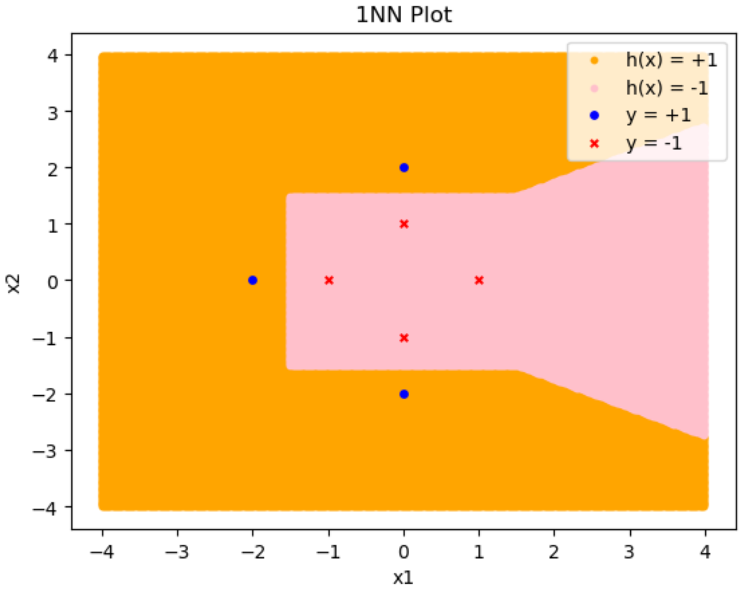
If f(x) $\leq \frac{1}{2}$ and $y = +1$,

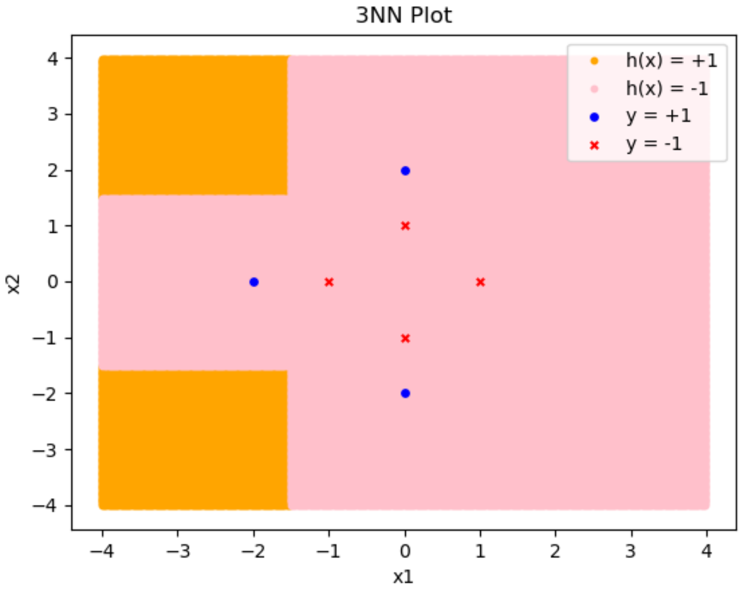
$$\begin{aligned} e(f(x)) &= P[f(x) \neq y] \\ &= P[f(x) = +1, y = -1] + P[f(x) = -1, y = +1] \\ &= 0 \times (1 - \pi(x)) + 1 \times \pi(x) \\ &= \pi(x) \end{aligned}$$

Thus, $e(f(x)) < \pi(x)$ for $\pi(x) \geq \frac{1}{2}$, and $e(f(x)) < 1 - \pi(x)$ for $\pi(x) \leq \frac{1}{2}$. So at any time, $e(f(x)) = min(\pi(x), 1 - \pi(x))$, and this indicates $e(f(x)) \leq \frac{1}{2}$. Therefore error of any other hypothesis function $e(h(x)) \geq min(\pi(x), 1 - \pi(x))$.

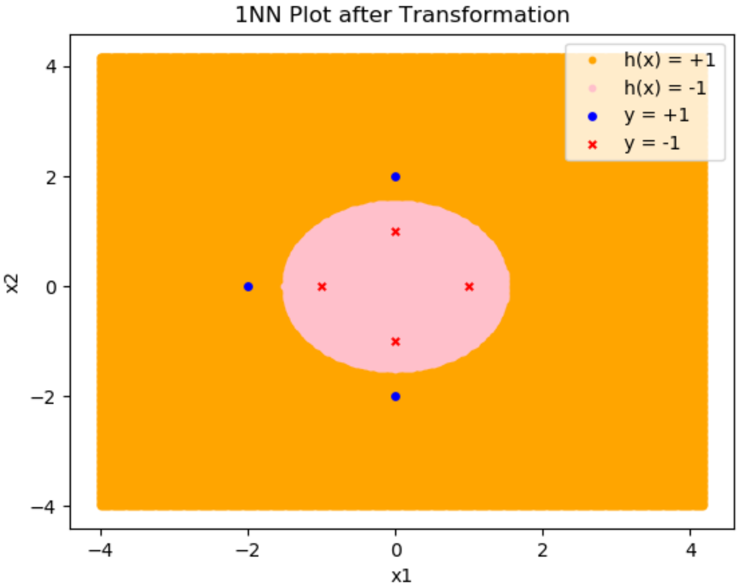
Problem 6.1

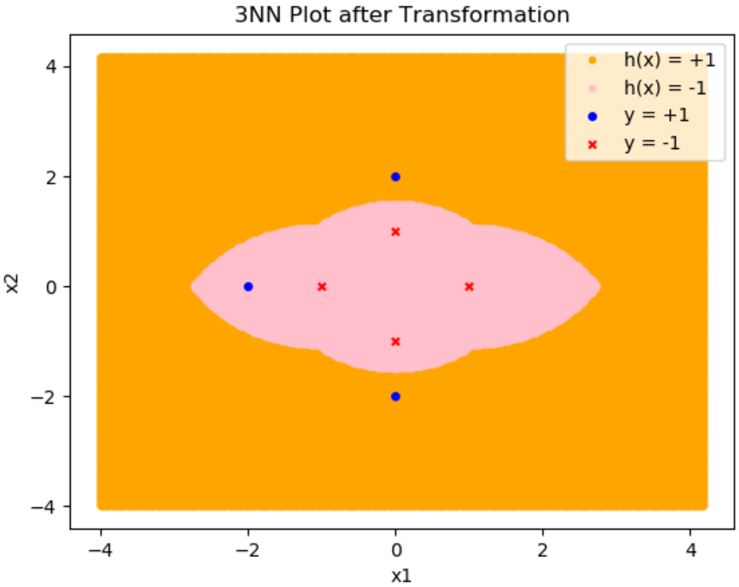
(a)


1NN Plot
h(x) = +1
h(x) = -1
y = +1
y = -1
x2
x1
4
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3
4

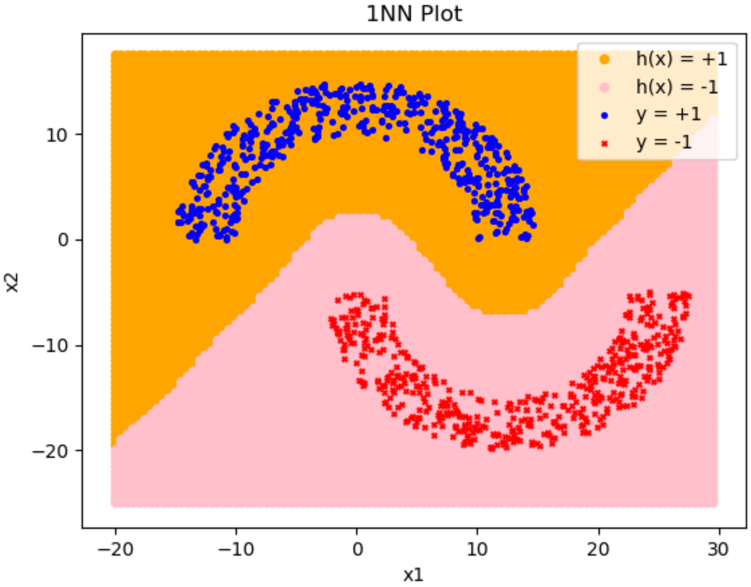

3NN Plot
h(x) = +1
h(x) = -1
y = +1
y = -1
x2
x1
4
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3
4

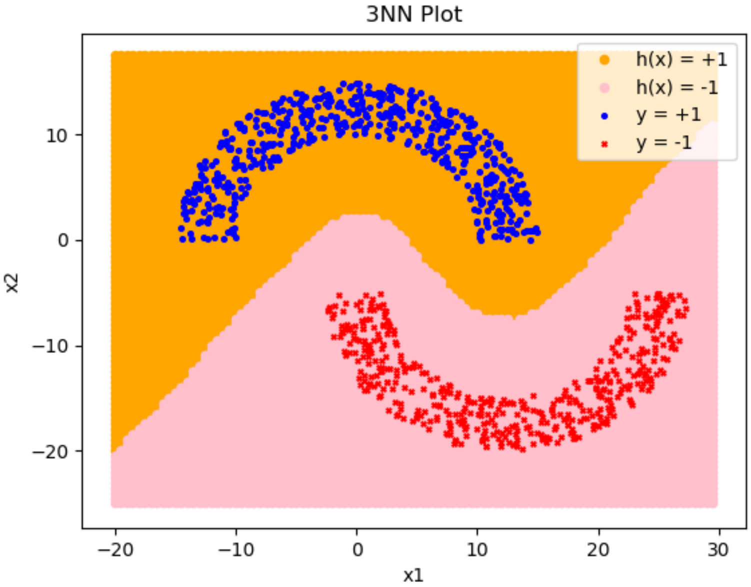
(b)


1NN Plot after Transformation
h(x) = +1
h(x) = -1
y = +1
y = -1
x2
4
2
0
−2
−4
x1
−4
−2
0
2
4


3NN Plot after Transformation
h(x) = +1
h(x) = -1
y = +1
y = -1
x2
4
2
0
−2
−4
x1
−4
−2
0
2
4

Problem 6.4


1NN Plot
h(x) = +1
h(x) = -1
y = +1
y = -1
x2
x1
10
0
−10
−20
−20
−10
0
10
20
30


3NN Plot
h(x) = +1
h(x) = -1
y = +1
y = -1
x2
x1
10
0
−10
−20
−20
−10
0
10
20
30

Problem 6.16

(a) Time cost for finding the nearest neighbor with partition: 20.435 seconds.
Time cost for finding the nearest neighbor with brute force: 153.322 seconds.

(b) Time cost for finding the nearest neighbor with partition: 28.723 seconds.
Time cost for finding the nearest neighbor with brute force: 158.261 seconds.

(c) Finding the nearest neighbor by using the partition with branch and bound cost much less time than the brute force approach, because the brute force approach go through all 10000 training point for every test point, whereas by using partition, for each point, only distances from about $\frac{1}{10}$ of points need to be compared with. Thus using partition shortens a lot of time.

Also, Finding nearest neighbor using data of gaussians distributions will cost much more time than using uniform data when running partition algorithm. The reason is that for gaussians distribution, most points located in the center part of coordination, and huge amount of points located in a few regions, and remaining regions only have relativly a few points. So if the test point is located at the center regions, the program needs to go through a mountain of points to compare the distance.

(d) Yes, because for small amount of data, partition have little effect on running time, and if the data set is small enough, the partition might cause even more time.