Homework 10

Karan Sarkar
sarkak2@rpi.edu

November 18, 2019

Exercise 6.1.

(a) The vectors $(1, 1)$ and $(10, 10)$ have high cosine similarity but low Euclidean similarity. The vectors $(0.5, -0.5)$ and $(0.5, 0.5)$ have low cosine similarity but high Euclidean similarity.

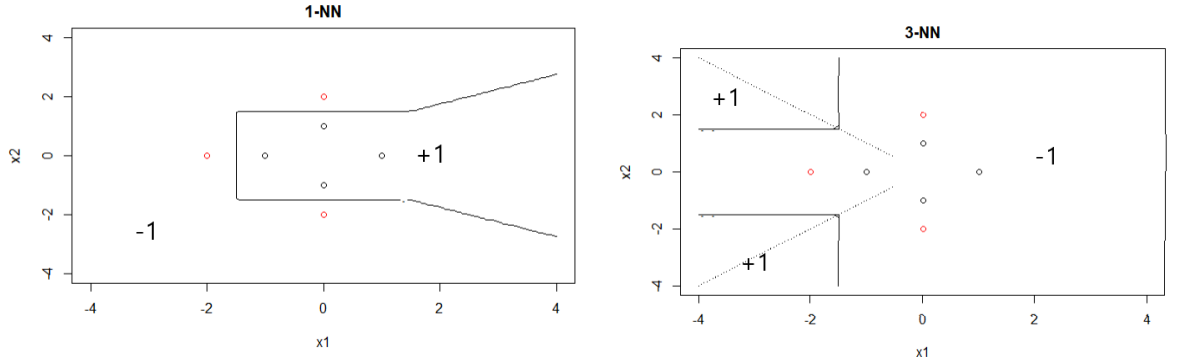
(b) When the origin changes, the cosine similarity changes but not the Euclidean similarity.This means your features have to be chosen such that the origin is the center of the distribution. In particular, the choice of origin which represents the most common or default data point must be purposeful.

Exercise 6.2. We will first find $e(f(x))$. If $\pi(x) > 0.5$, we have $e(f(x)) = P[y == -1] = 1 - \pi(x)$. If $\pi(x) < 0.5$, we have $e(f(x)) = P[y == +1] = \pi(x)$. Note that we are always choosing the smaller of $1 - \pi(x)$ and $\pi(x)$. Thus, $e(f(x)) = \min(\pi(x), 1 - \pi(x))$

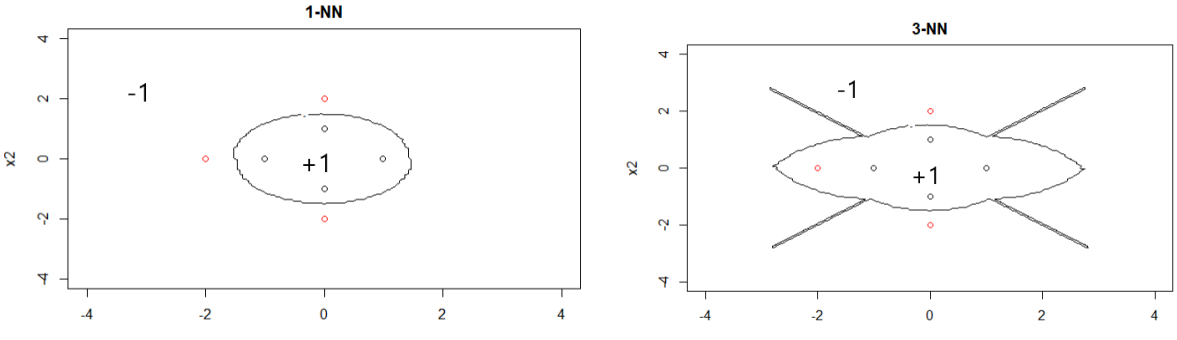
Note that the probability of error equals $\pi(x)$ when $x = -1$ and $1 - \pi(x)$ when $x = 1$. Therefore, $e(h(x)) \geq \min(\pi(x), 1 - \pi(x)) = e(f(x))$

Problem 6.1.

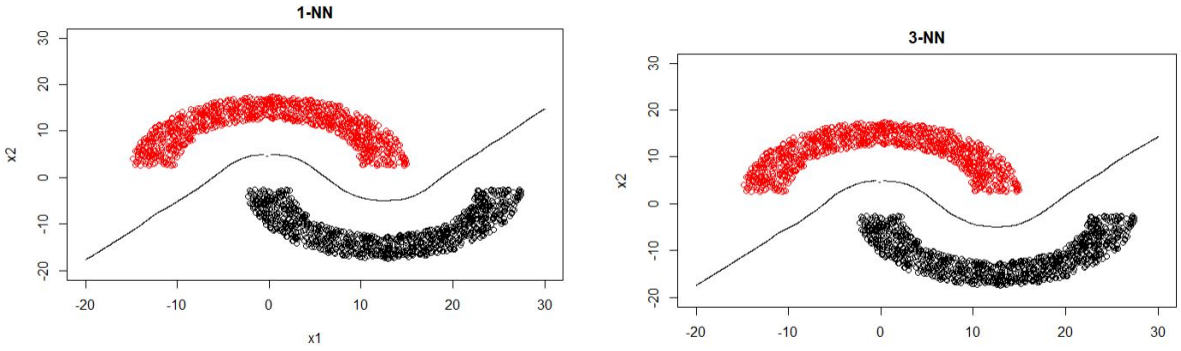
(a)



(b)



Problem 6.4.



Problem 6.16.

(a) Brute force takes 35.7 seconds. Branch and bound takes 7.6 seconds.

(b) Brute force takes 33.5 seconds. Branch and bound takes 3.4 seconds.

(c) Branch is more effective if the data has natural clusters. The more random the data the poorer branch and bound performs.

(d) Branch and bound has overhead from using more complex reasoning. Thus, it has much greater time savings when the number of data points is large. From a small amount of data, it may not be desirable to run branch and bound.