1.

Q1.

(1)

Euclidean = ((1-4)^2+0+(3-2)^2)^(1/2) = 3.16

Minkowski = 3.04

(2)

A

(3)

N

(4)

when K>min(count(A),count(B))

Q2.

(1)

K = 3

(2)

K = 3

2/11

(3)

all

Q3.

(1)

Kernel projects the data to a high dimension and make it linear seperable, then return it to the lower dimension and get the decision boundary.

(2)

(3)

b

(4)

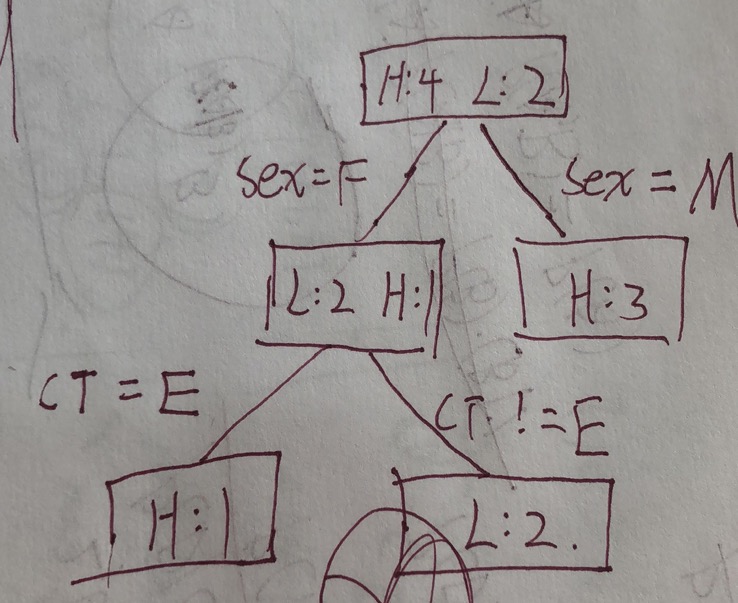
2. Week4

Q1

(1) 0.667

H(High|car\_type) = 1/3\*(1)+1/3\*(1) = 2/3

(2)



Q2.

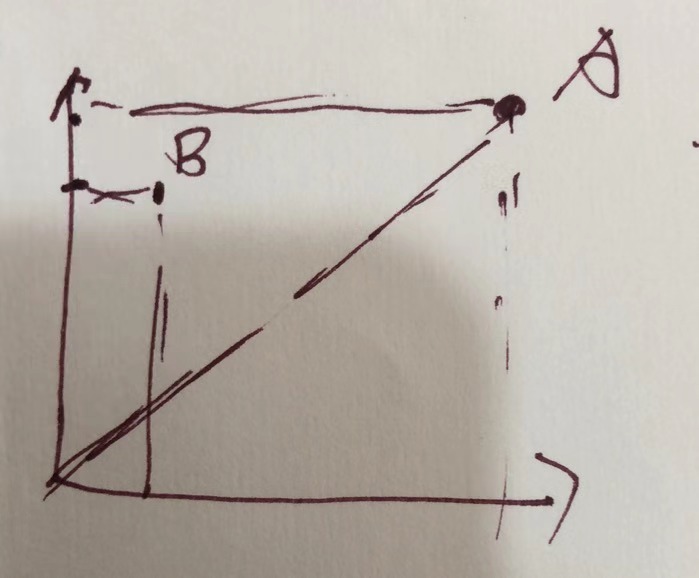
(1)

No, because in different situation, you need to get different accuracy.

(2)

sensitivity = TP/P = 450/450 = 1

specificity = TN/(FP+TN) = 0/50 = 0



(3)B classifier is better because A always predict everyone to positive.

3 Week 5

Q1.

(1)

The fundamental difference between bagging and random forest is that in Random forests, only a subset of features are selected at random out of the toal and the best split feature from the subset is used to split each node in a tree, unlike in bagging where all features are considered for splitting a node."

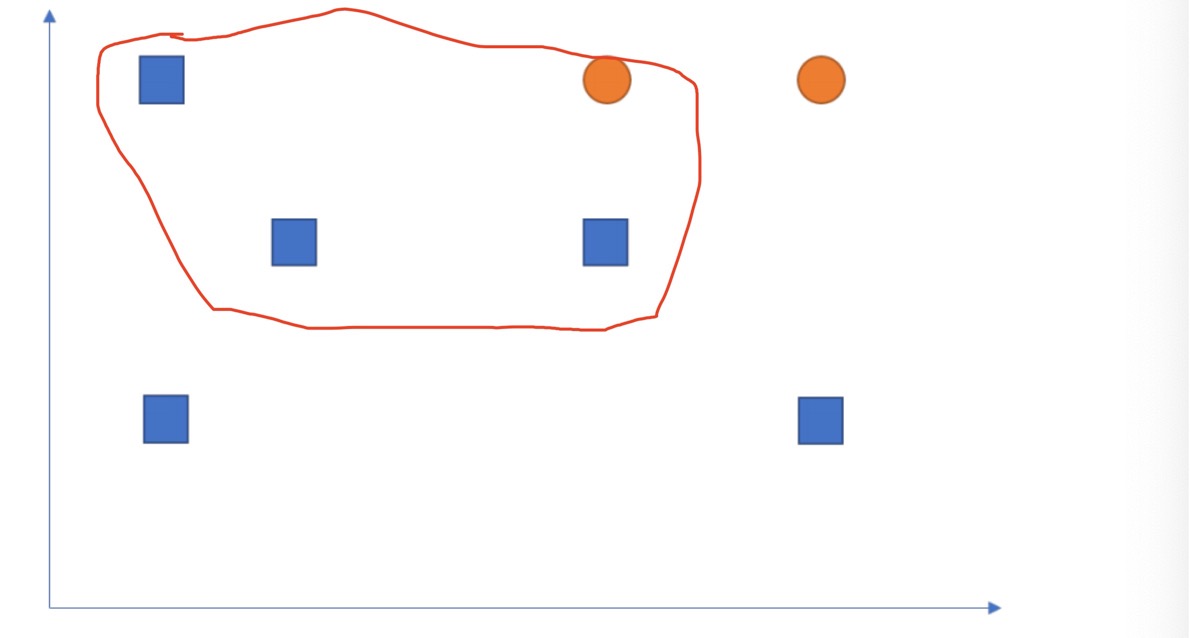
(2)

Random forest choose feature randomly

boosting improves the weight of the last iteration.

Q2

(1)



(2)

E1 0.12

E2 0.48

E3 0.16

E4 0.24

Q3.

(1)

sometime there are thousands of feature, and if you use them all, it is so hard to compute.

(2)

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables.

use pearson's correlation coefficient to measure the relationship between label and feature.

(3)

cc = sum((X-X\_)(Y-Y\_))/(sqrt(sum(X-X\_)^2)\*sqrt(sum(Y-Y\_)^2))