1. Question 1

Consider one auto company that receives parts from three suppliers; assume 50% of the parts from supplier 1, 30% from supplier 2, and 20% from supplier 3. The quality of the parts could be summarized in the following table based on historically data.

	Percentage Good Parts	Percentage Bad parts
Supplier 1	98	2
Supplier 2	95	5
Supplier 3	92	8

Question: A bad part broke one of the machines (observed), what is the probability the part came from supplier 1?

ANSWER:

Let A_1 denote Supplier 1, A_2 denote Supplier 2, and A_3 denote Supplier 3

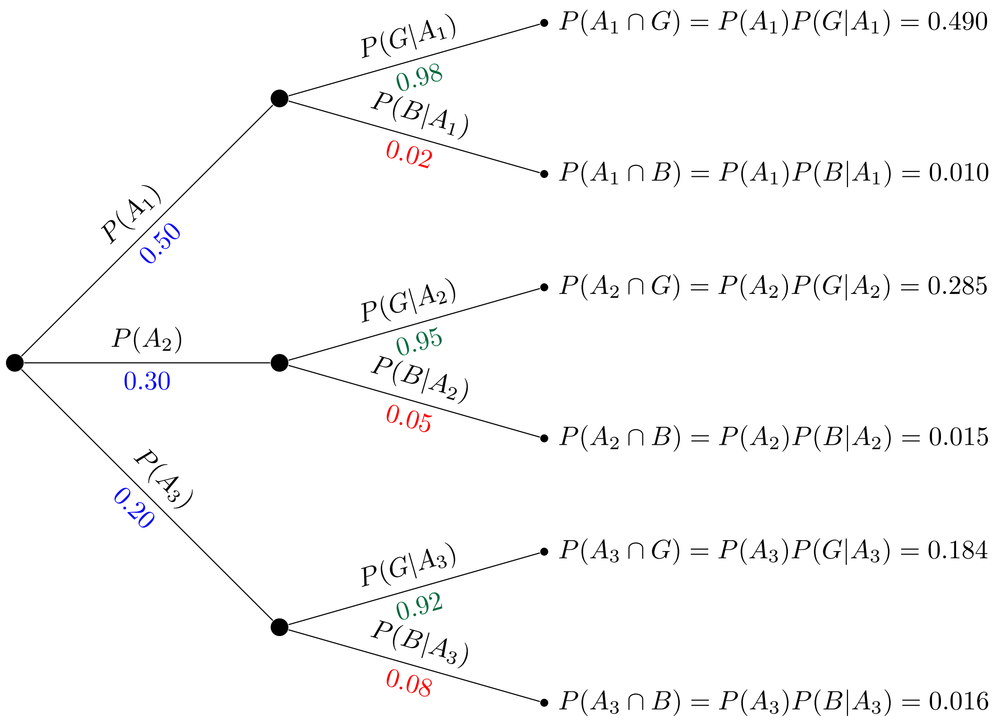
$$\begin{aligned} P(A_1) &= 0.50 \\ P(A_2) &= 0.30 \\ P(A_3) &= 0.20 \end{aligned} \tag{1.1}$$

Let G denote that a part is good and B denote that a part is bad.

$$\begin{aligned} P(G|A_1) &= 0.98 \\ P(G|A_2) &= 0.95 \\ P(G|A_3) &= 0.92 \end{aligned} \tag{1.2}$$

$$\begin{aligned} P(B|A_1) &= 0.02 \\ P(B|A_2) &= 0.05 \\ P(B|A_3) &= 0.08 \end{aligned} \tag{1.3}$$

Here is the Probability Tree for Three-Suppliers



Law of Conditional Probability gives us the following equation

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} \tag{1.4}$$

We know from are probability tree that

$$P(A_1 \cap B) = P(A_1)P(B|A_1) = 0.010 \tag{1.5}$$

We also know that

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) \tag{1.6}$$

Combining are equations togther we obtain Bayes' Theorem

$$\begin{gathered} 1 \leq k \leq n \\ P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)} \end{gathered} \tag{1.7}$$

$$\begin{aligned} P(A_1|B) &= \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3)} \\ P(A_2|B) &= \frac{P(A_2)P(B|A_2)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3)} \\ P(A_3|B) &= \frac{P(A_3)P(B|A_3)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3)} \end{aligned} \tag{1.8}$$

Now all we have to do is plug in the numbers and solve the equations

$$
\begin{aligned}
P(A_1|B) &= \frac{(0.50)(0.02)}{(0.50)(0.02)+(0.30)(0.05)+(0.20)(0.08)} = \frac{0.010}{0.041} = 0.2439024390 \\
P(A_2|B) &= \frac{(0.30)(0.05)}{(0.50)(0.02)+(0.30)(0.05)+(0.20)(0.08)} = \frac{0.015}{0.041} = 0.3658536585 \\
P(A_3|B) &= \frac{(0.20)(0.08)}{(0.50)(0.02)+(0.30)(0.05)+(0.20)(0.08)} = \frac{0.016}{0.041} = 0.3902439024
\end{aligned}
\tag{1.9}
$$

Therefore the probability the part came from supplier 1: 0.243902439

2. Question 2

For the play tennis data set shown below:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

2.1 Part 1

Please use Nave Bayes to help make decision on playing tennis or not when Outlook is rain, Temperature is mild, Humidity is normal and Wind is weak.

ANSWER:

OUTLOOK	Play=Yes	Play=No	Total
Sunny	2/9	3/5	5/14
Overcast	4/9	0/5	4/14
Rain	3/9	2/5	5/14

TEMPERATURE	Play=Yes	Play=No	Total
Hot	2/9	2/5	4/14
Mild	4/9	2/5	6/14
Cool	3/9	1/5	4/14

HUMIDITY	Play=Yes	Play=No	Total
High	3/9	4/5	7/14
Normal	6/9	1/5	7/14

WIND	Play=Yes	Play=No	Total
Strong	3/9	3/5	6/14
Weak	6/9	2/5	8/14

$$P(Play = Yes) = 9/14$$
$$P(Play = No) = 5/14$$

Let x' denote the conditions of playing tennis or not
$x' = (Outlook = Rain, Temperature = Mild, Humidity = Normal, Wind = Weak)$
The probability that tennis is played lookup table

$$\begin{aligned}
P(Outlook = Rain|Play = Yes) &= 3/9 \\
P(Temperature = Mild|Play = Yes) &= 4/9 \\
P(Humidity = Normal|Play = Yes) &= 6/9 \\
P(Wind = Weak|Play = Yes) &= 6/9
\end{aligned}$$

$$\begin{aligned}
P(Outlook = Rain|Play = No) &= 2/5 \\
P(Temperature = Mild|Play = No) &= 2/5 \\
P(Humidity = Normal|Play = No) &= 1/5 \\
P(Wind = Weak|Play = No) &= 2/5
\end{aligned}$$

Now we can construnct the following equation

$$\begin{aligned}
P(Play = Yes|x') &= P(x'|Play = Yes)P(Play = Yes) \\
&= [P(Rain|Yes)P(Mild|Yes)P(Normal|Yes)P(Weak|Yes)]P(Play = Yes) \\
&= (3/9 * 4/9 * 6/9 * 6/9)(9/14) \\
&= 0.0423280423
\end{aligned}$$

$$\begin{aligned}
P(Play = No|x') &= P(x'|Play = No)P(Play = No) \\
&= [P(Rain|No)P(Mild|No)P(Normal|No)P(Weak|No)]P(Play = No) \\
&= (2/5 * 2/5 * 1/5 * 2/5)(5/14) \\
&= 0.0045714286
\end{aligned}$$

Given The Fact that $P(Play = Yes|x') > P(Play = No|x')$ we would label x' to be Yes

2.2 Part 2

The humidity value could be continuous practically. In the above data set, if the humidity value is as follows per original data set order:

$$\begin{aligned}
Yes :& 65.7, 20.7, 5.1, 6.9, 4.8, 6.9, 8.7, 10.4, 15.3, \\
No :& 58.1, 66.4, 6.5, 10.5, 12.8
\end{aligned}$$

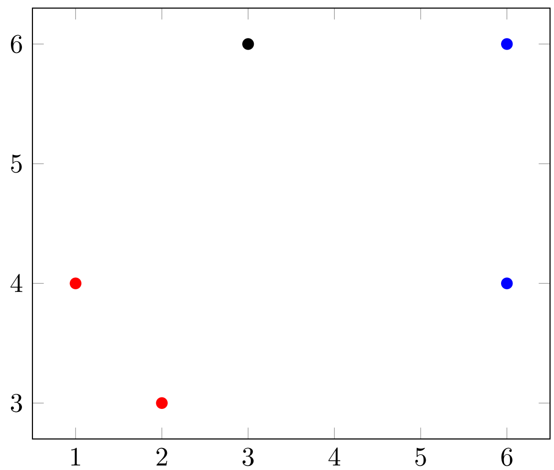
Please use Nave Bayes to help make decision on playing tennis or not when the Outlook is Overcast, Temperature is mild, Humidity is 8.8, and Wind is weak.

3. Question 3

We have a training dataset as follows:

Feature 1	Feature 2	Label
6	6	L1
6	4	L1
2	3	L2
1	4	L2

Using K-NN algorithm to determine the label for a new data record (3, 6). The similarity measure is assumed to be Euclidean distance.



Using Euclidean Distance we can calculate the distances between the two points

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (f_r(x_i) - f_r(x_j))^2}$$

Using Euclidean Distance we get the following results

$$(6, 6, L1) : \sqrt{(6-3)^2 + (6-6)^2} = \sqrt{9} = 3$$
$$(6, 4, L1) : \sqrt{(6-3)^2 + (4-6)^2} = \sqrt{13} = 3.605551275$$
$$(2, 3, L2) : \sqrt{(2-3)^2 + (3-6)^2} = \sqrt{10} = 3.16227766$$
$$(1, 4, L2) : \sqrt{(1-3)^2 + (4-6)^2} = \sqrt{8} = 2.828427125$$

Using K=1 we can see that the new data record of (3, 6) would be classified as L2
Using K=2 the new data record is split between L1 and L2
Using K=3 we can see that the new data record of (3, 6) would be classified as L2
Using K=4 the new data record is split between L1 and L2

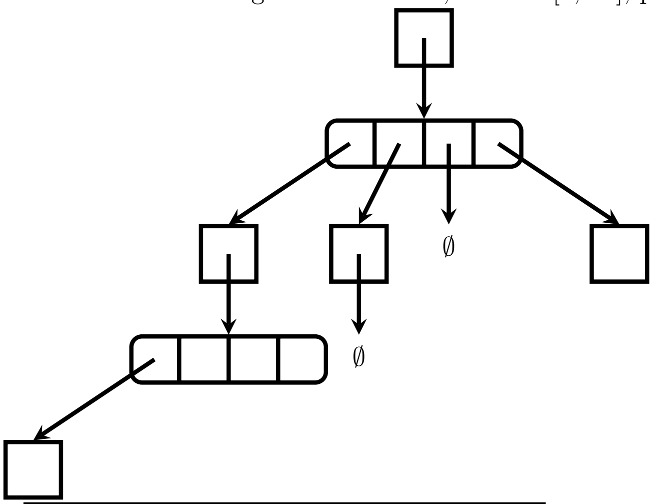
4. Question 4

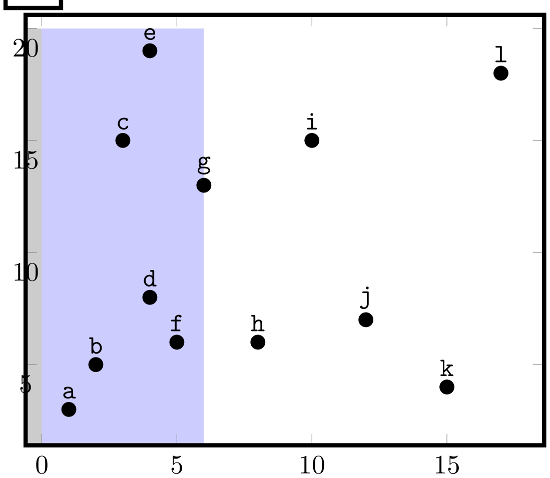
Given the following 12 data points x=(x1, x2) in the training data set.

$$(1,3),(2,5),(3,15),(4,8),(4,19),(5,6),(6,13),(8,6),(10,15),(12,7),(15,4),(17,18)$$

4.1 Part 1

If we know the value range of features x1, x2 is in [0, 20], please build the quadtree.





4.2 Part 2

Using the quadtree learnt in 4.1 to find the nearest neighbor of data point (11, 16).