Question 1

We have a classification problem with two classes "+" and "−", three trained classifiers h_1, h_2, and h_3, with the following probabilities of the classifiers, given the training data D:

$$\Pr(h_1|D) = 0.5$$
$$\Pr(h_2|D) = 0.3$$
$$\Pr(h_3|D) = 0.2$$

For the three test instances o_1, o_2, o_3, the classifiers give the following class probabilities:

$$\begin{aligned}
o_1 : &\Pr(+|h_1) = 0.6 && \Pr(-|h_1) = 0.4 \\
&\Pr(+|h_2) = 0.2 && \Pr(-|h_2) = 0.8 \\
&\Pr(+|h_3) = 0.9 && \Pr(-|h_3) = 0.1 \\
o_2 : &\Pr(+|h_1) = 0.6 && \Pr(-|h_1) = 0.4 \\
&\Pr(+|h_2) = 0.6 && \Pr(-|h_2) = 0.4 \\
&\Pr(+|h_3) = 1 && \Pr(-|h_3) = 0 \\
o_3 : &\Pr(+|h_1) = 0.6 && \Pr(-|h_1) = 0.4 \\
&\Pr(+|h_2) = 0.6 && \Pr(-|h_2) = 0.4 \\
&\Pr(+|h_3) = 0 && \Pr(-|h_3) = 1
\end{aligned}$$

We combine the three classifiers to get a Bayes optimal classifier. Which of the following class probabilities will we get from this Bayes optimal classifier?

1. $\underline{o_1 : \Pr(+|\text{Bayes optimal}) = 0.54}$
2. $o_1 : \Pr(-|\text{Bayes optimal}) = 0.42$
3. $o_2 : \Pr(+|\text{Bayes optimal}) = 0.58$
4. $\underline{o_2 : \Pr(-|\text{Bayes optimal}) = 0.32}$
5. $o_3 : \Pr(+|\text{Bayes optimal}) = 0.38$

6. $\underline{o_3 : \Pr(-|\text{Bayes optimal}) = 0.52}$

Solution:
The Bayes optimal classifier adds the conditional class probabilities given the classifier, weighted with the conditional classifier probabilities given the data:

$$\Pr(c_j|D) = \sum_{h_i \in H} \Pr(c_j|h_i) \Pr(h_i|D)$$

The resulting probabilities are:

$$\begin{aligned}
o_1 : &\quad \Pr(+|\text{Bayes optimal}) = 0.6 \cdot 0.5 + 0.2 \cdot 0.3 + 0.9 \cdot 0.2 = 0.54 \\
o_1 : &\quad \Pr(-|\text{Bayes optimal}) = 0.4 \cdot 0.5 + 0.8 \cdot 0.3 + 0.1 \cdot 0.2 = 0.46 \\
o_2 : &\quad \Pr(+|\text{Bayes optimal}) = 0.6 \cdot 0.5 + 0.6 \cdot 0.3 + 1 \cdot 0.2 = 0.68 \\
o_2 : &\quad \Pr(-|\text{Bayes optimal}) = 0.4 \cdot 0.5 + 0.4 \cdot 0.3 + 0 \cdot 0.2 = 0.32 \\
o_3 : &\quad \Pr(+|\text{Bayes optimal}) = 0.6 \cdot 0.5 + 0.6 \cdot 0.3 + 0 \cdot 0.2 = 0.48 \\
o_3 : &\quad \Pr(-|\text{Bayes optimal}) = 0.4 \cdot 0.5 + 0.4 \cdot 0.3 + 1 \cdot 0.2 = 0.52
\end{aligned}$$

So the correct probabilities are 1, 4 and 6.

Question 2

A fortune teller specialized on palm reading wants to improve his prediction accuracy. As training data, he has observations on some students, characteristics of their palms, and whether or not these students passed their exam (i.e., the class to predict).

Note: the answers use a short notation for tests in the tree:

- A test on attribute A is denoted by "A"?
- A possible answer is given after ":"
- The consequence of an answer is given after "$\rightarrow$"
- The consequence can be a new test (e.g., "B?", i.e., the root of the subtree below this branch) or a class (leave node, e.g., "yes").

student	little finger	head line	fate line	passed exam
1	straight	long	invisible	yes
2	bent	long	invisible	yes
3	straight	long	deep	no
4	straight	long	invisible	no
5	straight	long	invisible	yes
6	bent	long	invisible	yes
7	bent	short	deep	no
8	straight	short	deep	no
9	bent	short	invisible	no
10	straight	short	deep	yes

Now the fortune teller wants to learn a decision tree for these training data. As root of the tree, the "head line" was already selected.

Using the gini index, which attributes are used as test nodes at the next level?

1. head line?: short $\rightarrow$ little finger?
2. head line?: short $\rightarrow$ fate line?
3. head line?: long $\rightarrow$ little finger?
4. head line?: long $\rightarrow$ fate line?

Solution:
We choose the attribute and the split that minimizes the Gini index.

For head line=short:
$|T| = 4$, 3 "no and 1 "yes".

- G(little finger)
  - straight: T_1 = persons 8, 10

$$p(PE = no) = \frac{1}{2}$$
$$p(PE = yes) = \frac{1}{2}$$
$$G(T_1) = 1 - \left(\frac{1^2}{2^2} + \frac{1^2}{2^2}\right) = \frac{1}{2}$$

  - bent: T_2 = persons 7, 9

$$p(PE = no) = \frac{2}{2}$$
$$p(PE = yes) = \frac{0}{2}$$
$$G(T_2) = 1 - \left(\frac{2^2}{2^2} + \frac{0^2}{2^2}\right) = 0$$

  We can now calculate the Gini index

$$G(\text{little finger}) = \frac{2}{4} \cdot \frac{1}{2} + \frac{2}{4} \cdot 0 = \frac{1}{4} = 0.25$$

- G(fate line)
  - deep: T_1 = persons 7, 8, 10

$$p(PE = no) = \frac{2}{3}$$
$$p(PE = yes) = \frac{1}{3}$$
$$G(T_1) = 1 - \left(\frac{2^2}{3^2} + \frac{1^2}{3^2}\right) = \frac{4}{9}$$

- invisible: T_2 = persons 9

$$p(PE = no) = \frac{1}{1}$$
$$p(PE = yes) = \frac{0}{1}$$
$$G(T_2) = 1 - \left(\frac{1^2}{1^2} + \frac{0^2}{1^2}\right) = 0$$

We can now calculate the Gini index

$$G(\text{fate line}) = \frac{3}{4} \cdot \frac{4}{9} + \frac{1}{4} \cdot 0 = \frac{1}{3} \approx 0.333$$

- Since $G(\text{little finger}) < G(\text{fate line})$, we choose to split on little finger.

For head line=long:
$|T| = 6$, 2 "no and 4 "yes".

- G(little finger)
  - straight: T_1 = persons 1, 3, 4, 5

$$p(PE = no) = \frac{2}{4}$$
$$p(PE = yes) = \frac{2}{4}$$
$$G(T_1) = 1 - \left(\frac{2^2}{4^2} + \frac{2^2}{4^2}\right) = \frac{1}{2}$$

  - bent: T_2 = persons 2, 6

$$p(PE = no) = \frac{0}{2}$$
$$p(PE = yes) = \frac{2}{2}$$
$$G(T_2) = 1 - \left(\frac{0^2}{2^2} + \frac{2^2}{2^2}\right) = 0$$

We can now calculate the Gini index

$$G(\text{little finger}) = \frac{4}{6} \cdot \frac{1}{2} + \frac{2}{6} \cdot 0 = \frac{1}{3} \approx 0.333$$

- G(fate line)
  - deep: T_1 = person 3

$$p(PE = no) = \frac{1}{1}$$
$$p(PE = yes) = \frac{0}{1}$$
$$G(T_1) = 1 - \left(\frac{1^2}{1^2} + \frac{0^2}{1^2}\right) = 0$$

  - invisible: T_2 = persons 1, 2, 4, 5, 6

$$p(PE = no) = \frac{1}{5}$$
$$p(PE = yes) = \frac{4}{5}$$
$$G(T_2) = 1 - \left(\frac{1^2}{5^2} + \frac{4^2}{5^2}\right) = \frac{8}{25} \approx 0.32$$

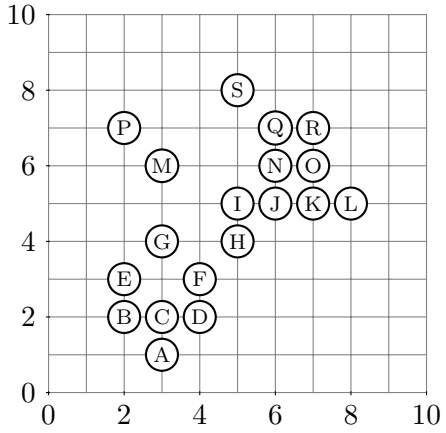
  We can now calculate the Gini index

$$G(\text{fate line}) = \frac{1}{6} \cdot 0 + \frac{5}{6} \cdot \frac{8}{25} = \frac{4}{15} \approx 0.27$$

- Since G(fate line) $<$ G(little finger), we choose to split on fate line.

Question 3

Given the following dataset and Manhattan distance as distance function:



Given the definitions of DBSCAN and counting the query object as a member of its neighborhood: which of the following statements are correct?

1. A is core point with $\varepsilon = 2$ and MinPts $= 4$.
2. B is core point with $\varepsilon = 2$ and MinPts $= 4$.
3. G is core point with $\varepsilon = 2$ and MinPts $= 4$.
4. J is core point with $\varepsilon = 2$ and MinPts $= 9$.
5. M is core point with $\varepsilon = 2$ and MinPts $= 4$.
6. P is core point with $\varepsilon = 2$ and MinPts $= 1$.
7. Q is core point with $\varepsilon = 1$ and MinPts $= 4$.
8. R is core point with $\varepsilon = 1$ and MinPts $= 4$.
9. P is directly density-reachable from M with $\varepsilon = 2$ and MinPts $= 4$.
10. P and S are density-connected with $\varepsilon = 2$ and MinPts $= 3$.
11. A and L are density-connected with $\varepsilon = 1$ and MinPts $= 2$.

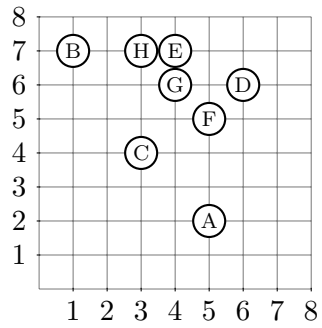
Solution:
Recall the differences between core- and border points. The core points, as the name suggests, lie usually within the interior of a cluster. A border point has fewer than MinPts within its ε-neighborhood, but it lies in the neighborhood of another core point.

Density Reachable: A point is called density reachable from another point if they are connected through a series of core points.
Density Connected: Two points are called density connected if there is a core point which is density reachable from both the points.

The statements 1, 2, 3, 6 and 10 are correct.

Question 4

Given the following dataset and Manhattan distance as distance function:



We have subsets of these points ordered by their outlier score (knn or LOF, the query point does not count to its own neighborhood here). For a correctly ordered subset of points, the top outlier among the selected points is listed first, later points follow with (not strictly) decreasing outlier score. Which of the following subsets are correctly ordered w.r.t. the given outlier method and parameter?

1. A,B,C w.r.t. LOF (MinPts = 2)
2. A,C,D w.r.t. LOF (MinPts = 2)
3. B,C,D w.r.t. LOF (MinPts = 2)
4. A,B,D w.r.t. kNN ($k = 2$)
5. C,D,E w.r.t. kNN ($k = 2$)
6. C,E,G w.r.t. kNN ($k = 2$)

Solution:
We calculate LOF_2 for the three points.
For A:

$$\begin{aligned}
\text{reachdist}_2(A,C) &= \max\{k-dist(C), dist(A,C)\} \\
&= \max\{3,4\} = 4 \\
\text{reachdist}_2(A,F) &= \max\{k-dist(F), dist(A,F)\} \\
&= \max\{2,3\} = 3 \\
lrd_2(A) &= 1/\frac{4+3}{2} = \frac{2}{7} \approx 0.29 \\
\text{reachdist}_2(C,H) &= \max\{k-dist(H), dist(C,H)\} \\
&= \max\{2,3\} = 3 \\
\text{reachdist}_2(C,G) &= \max\{k-dist(G), dist(C,G)\} \\
&= \max\{2,3\} = 3 \\
\text{reachdist}_2(C,F) &= \max\{k-dist(F), dist(C,F)\} \\
&= \max\{2,3\} = 3 \\
lrd_2(C) &= 1/\frac{3+3+3}{3} = \frac{1}{3} \approx 0.333 \\
\text{reachdist}_2(F,G) &= \max\{k-dist(G), dist(F,G)\} \\
&= \max\{2,2\} = 2 \\
\text{reachdist}_2(F,D) &= \max\{k-dist(D), dist(F,D)\} \\
&= \max\{2,2\} = 2 \\
lrd_2(F) &= 1/\frac{2+2}{2} = \frac{1}{2} = 0.5 \\
LOF_2(A) &= \frac{\frac{0.333}{0.29} + \frac{0.5}{0.29}}{2} \approx 1.458333
\end{aligned}$$

Likewise we calculate for B and C. We get the following results.

$$\begin{aligned}
LOF_2(A) &= 1.458333 \\
LOF_2(B) &= 1.25 \\
LOF_2(C) &= 1.571428571428571
\end{aligned}$$

Neither statement 1, 2 or 3 are correct.

The kNN distances for A, B, C, D, E and G are the following.

$$kNN_2(A) = 4$$
$$kNN_2(B) = 3$$
$$kNN_2(C) = 3$$
$$kNN_2(D) = 2$$
$$kNN_2(E) = 1$$
$$kNN_2(G) = 2$$

The correct statements are statements 4, 5.