Q1: K-Nearest Neighbours and Naïve Bayes

***Q1(a) Answer*:**

Overfitting refers to the statistical analysis that is corresponding too closely or exactly to a specific set of data, in which case it would probably fail to fit additional data or predict future observations. Therefore, when k = 1 in KNN classification, the label is relying on the category of the closest input vector, in which case the classification result is "controlled" by certain information which could not generally represent the local feature of the data. Therefore, it tends to overfit.

Underfitting occurs when the model isn't able to adequately capture the feature of the data's structure. When k is equal to the number of the training patterns, the classification result is determined by the global feature of the data, rather than local. Therefore, it tends to underfit.

Q1(b) Answer:

For K = 1:

$$\hat{f}(x) = \begin{cases} 4, x \in (-\infty, 3.5] \\ 25, x \in (3.5, 6] \\ 49, x \in (6, +\infty] \end{cases}$$

For K = 2:

$$\hat{f}(x) = \begin{cases} \dfrac{70 - 29x}{7 - 2x}, x \in (-\infty, 2) \\ 4, \quad x = 2 \\ 7x - 10, x \in (2, 4.5] \\ \dfrac{210 - 37x}{6 - x}, x \in (4.5, 5] \\ 25, \quad x = 5 \\ 12x - 35, x \in (5, 7) \\ 49, \quad x = 7 \\ \dfrac{37x - 210}{x - 6}, x \in (7, +\infty) \end{cases}$$

For K = 3:

$$\hat{f}(x) = \begin{cases} \dfrac{78x^2 - 616x + 980}{3x^2 - 28x + 59}, x \in (-\infty, 2) \\ \dfrac{70x^2 - 520x + 700}{x^2 - 4x - 11}, x \in (2,5) \\ \dfrac{-(20x^2 - 70x)}{x^2 - 14x + 39}, x \in (5,7) \\ \dfrac{78x^2 - 616x + 980}{3x^2 - 28x + 59}, x \in (7, +\infty) \end{cases}$$

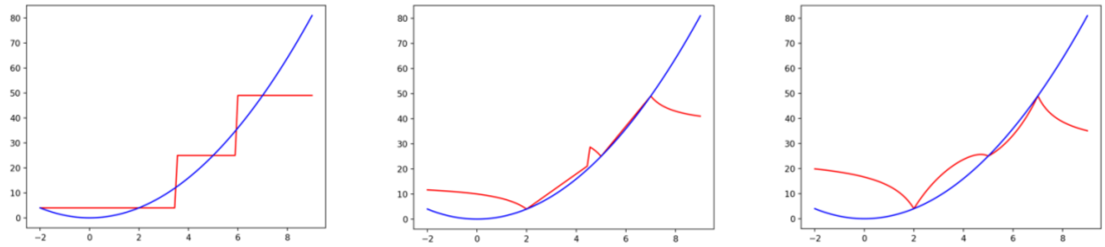

Figure 1: K = 1 (left); K = 2 (middle); K = 3 (right)

Since the **Root Mean Square Error (RMSE)** is

$$RMSE = \begin{cases} 6.56, K = 1, \\ 2.14, K = 2 \quad where\ x \in (2,7) \\ 4.99, K = 3 \end{cases}$$

And the **Coefficient of Determination (R^2)** is

$$R^2 = \begin{cases} 0.75, K = 1 \\ 0.97\ K = 2 \quad where\ x \in (2,7) \\ 0.86, K = 3 \end{cases}$$

Therefore, ***K = 2 is the optimal value since both metric values are the smallest***.

Q1(c) Answer:

Let $P(c_1) = p_1, P(c_1) = 1 - p_1$. The data provides some evidences E corresponding to N$_1$ examples of C$_1$ and N$_2$ = N - N$_1$ examples of C$_2$. By Bayes Theorem:

$$f(p_1|E) = \frac{P(E|p_1)f(p_1)}{\int_0^1 P(E|p_1)f(p_1)dp_1}$$

where f represents uniform distribution,

$$f(p_1|E) = \frac{C_N^{N_1} {p_1}^{N_1}(1-p_1)^{N_2}}{\int_0^1 C_N^{N_1} {p_1}^{N_1}(1-p_1)^{N_2}}$$

$$= \frac{{p_1}^{N_1}(1-p_1)^{N_2}}{\int_0^1 {p_1}^{N_1}(1-p_1)^{N_2}}$$

$$E(p_1|E) = \int_0^1 p_1 \cdot f(p_1|E)\, dp_1$$

$$= \frac{\int_0^1 {p_1}^{N_1+1}(1-p_1)^{N_2} dp_1}{\int_0^1 {p_1}^{N_1}(1-p_1)^{N_2}\, dp_1}$$

Assume that for x and y natural numbers,

$$\int_0^1 p^x(1-p)^y \, dp = \frac{x!\,y!}{(x+y+1)!}$$

Then

$$E(p_1|E) = \frac{(n_1+1)!\,n_2!}{(n_1+n_2+2)!} \Big/ \frac{n_1!\,n_2!}{(n_1+n_2+1)!}$$

$$= \frac{(n_1+1)!\,n_2!}{(N+2)!} \Big/ \frac{n_1!\,n_2!}{(N+1)!}$$

Therefore,

$$\boldsymbol{E(p_1|E) = P(c_1) = \frac{n_1+1}{N+2}}$$

Q2: Clustering and Distance

Q2(a) Answer:

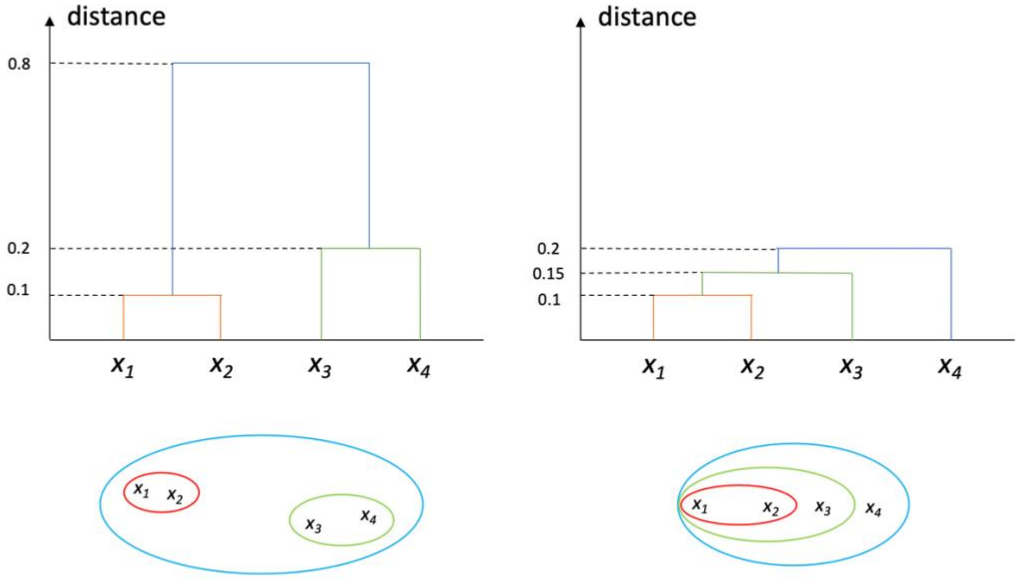

Figure 2. Complete Linkage Clustering and natural clusters (Left), Single Linkage Clustering and natural clusters (Right)

In this case, natural clusters using max() are $\{x_1, x_2\}, \{x_3, x_4\}$, using min() are $\{x_1, x_2, x_3\}, \{x_4\}$

Q2(b) Answer:

Given that $d(C_1, C_2) \leq d(C_1, C_3)$ and $d(C_1, C_2) \leq d(C_2, C_3)$
By the complete-linkage algorithm,

$$d(C_1, C_2) = \max\{d(x, y), x \in C_1, y \in C_2\}$$

$$d(C_3, C_1 \cup C_2) = \max\{d(x, y), x \in C_3, y \in C_1 \cup C_2\}$$

$$= \max\{d(C_1, C_3), d(C_2, C_3)\}$$

Therefore,

$$\boldsymbol{d(C_3, C_1 \cup C_2) \geq d(C_1, C_2)}$$

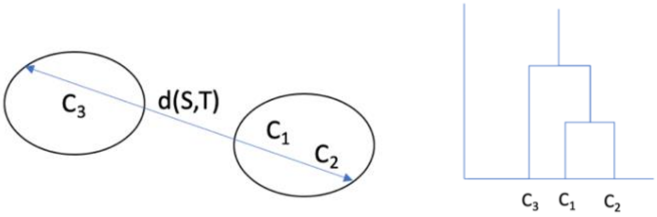

Figure 3. Complete Linkage Clustering Algorithm

Comment: In the hierarchical clustering algorithm, a **dendrogram** shows the hierarchical relationship between data points. The vertical axis of dendrogram represents the level of dissimilarity based on different types of distance metrics.

In this case, it is necessary to dig out the hierarchical relations. Initially, all data points are regarded as single clusters, then calculate the distance between points in each pair, a proximity matrix can be obtained. Find the two closest clusters and merge, then re-calculate the proximity matrix according to the new clusters. Stop the process when all points are within one cluster. The process and the result could be visualised a dendrogram. ***Therefore, the distance metric is important for digging out the relationships among the data points, simultaneously keep merging the clusters and visualising the process via dendrogram.***

Q3: Linear Regression and Decision Tree

Q3(a) Answer:

Given that $y = ax^2 + b$. Therefore, E for error is

$$E = \frac{1}{N}\sum_{(x,y)}(y - ax^2 - b)^2$$

The partial derivatives with respect to a and b are

$$\frac{\partial E}{\partial a} = \frac{1}{N}\sum_{(x,y)} 2(y - ax^2 - b)(-x^2) = 0$$

$$\frac{\partial E}{\partial b} = \frac{1}{N}\sum_{(x,y)} 2(y - ax^2 - b)(-1) = 0$$

Therefore, for b

$$\frac{1}{N}\sum_{(x,y)}(y - ax^2 - b) = 0$$

$$\boldsymbol{b = \overline{y} - a\overline{x^2}}$$

For a,

$$\frac{1}{N}\sum_{(x,y)} x^2(y - ax^2 - b) = 0$$

$$\overline{yx^2} - a\overline{x^4} - b\overline{x^2} = 0$$

Then substitute b,

$$\boldsymbol{a} = \frac{\overline{\boldsymbol{yx^2}} - \overline{\boldsymbol{y}}\,\overline{\boldsymbol{x^2}}}{\overline{\boldsymbol{x^4}} - \overline{\boldsymbol{x^2}}^2}$$

Q3(b) Answer:

$$\bar{y} = \frac{1}{4}(1 + 23 + 40 + 90) = 38.5$$

$$\overline{x^2} = \frac{1}{4}(0^2 + 5^2 + 7^2 + 10^2) = 43.5$$

$$\overline{yx^2} = \frac{1}{4}(23 \cdot 5^2 + 40 \cdot 7^2 + 90 \cdot 10^2) = 2883.75$$

$$\overline{x^4} = \frac{1}{4}(0^4 + 5^4 + 7^4 + 10^4) = 3256.5$$

Therefore,

$$a = \frac{2883.75 - 38.5 \times 43.5}{3256.5 - 43.5^2} \approx 0.886$$

$$b = 38.5 - a \times 43.5 = -0.0497 \approx -0.05$$

$$\boldsymbol{y = 0.886x^2 - 0.05}$$

Q3(c) Answer:

$$G(D_1) = 1 - \sum_{i=1}^{k} p_{i,1}^2$$
$$G(D_2) = 1 - \sum_{i=1}^{k} p_{i,2}^2$$

From above,

$$\frac{|\boldsymbol{D_1}|}{|\boldsymbol{D}|} \times \boldsymbol{G}(\boldsymbol{D_1}) + \frac{|\boldsymbol{D_2}|}{|\boldsymbol{D}|} \times \boldsymbol{G}(\boldsymbol{D_1}) = \frac{|D_1|}{|D|} \times \left(1 - \sum_{i=1}^{k} p_{i,1}^2\right) + \frac{|D_2|}{|D|} \times \left(1 - \sum_{i=1}^{k} p_{i,2}^2\right)$$
$$= \boldsymbol{1} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \left(\frac{|\boldsymbol{D_1}|}{|\boldsymbol{D}|} \times \boldsymbol{p_{i,1}^2} + \frac{|\boldsymbol{D_2}|}{|\boldsymbol{D}|} \times \boldsymbol{p_{i,2}^2}\right)$$

Given that,

$$p_i \times |D| = p_{i,1} \times |D_1| + p_{i,2} \times |D_2|$$

Therefore,

$$p_i = \frac{|D_1|}{|D|} \times p_{i,1} + \frac{|D_2|}{|D|} \times p_{i,2}$$

And,

$$\boldsymbol{G}(\boldsymbol{D}) = 1 - \sum_{i=1}^{k} p_i^2 = \boldsymbol{1} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \left(\frac{|\boldsymbol{D_1}|}{|\boldsymbol{D}|} \times \boldsymbol{p_{i,1}} + \frac{|\boldsymbol{D_2}|}{|\boldsymbol{D}|} \times \boldsymbol{p_{i,2}}\right)^2$$

Since given that $\forall a, b \in R^{+}, x, y \in R, a + b = 1,$

$$\boldsymbol{ax^2 + by^2 \geq (ax + by)^2}$$

Then

$$1 - \sum_{i=1}^{k} \left(\frac{|D_1|}{|D|} \times p_{i,1} + \frac{|D_2|}{|D|} \times p_{i,2} \right)^2 \geq 1 - \sum_{i=1}^{k} \left(\frac{|D_1|}{|D|} \times p_{i,1}^2 + \frac{|D_2|}{|D|} \times p_{i,2}^2 \right)$$

Therefore,

$$\boldsymbol{G(D) \geq \frac{|D_1|}{|D|} \times G(D_1) + \frac{|D_2|}{|D|} \times G(D_2)}$$

Why is this property important when using Gini Impurity?
The above relationship indicates that "the Gini impurity is greater or equal to the weighted sum of its children nodes' Gini impurity". In this case, as the level of decision tree goes deeper, the classification would be more accurate in general, since the values of Gini Impurity have a decreasing trend. Otherwise, the decision tree algorithm is not reasonable.

Q4: Markov Decision Processes

Q4(a) Answer:

- **A policy**, denoted by π, refers to a solution that describe what the agent does in every state. $\pi(s)$ for an individual state describes which action is taken in state s.
- The utility or value function is the sum of the received rewards, which depends on a sequence of states. **The expected utility/value of a policy** refers the execution of π starting in s is given by

$$U^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t)\right]$$

  where $\gamma^t R(s_t)$ represents the discounted rewards, γ is the discount factor within [0, 1].
- **An optimal policy** refers to the solution for an MDP which describes the best action in a state, which is the one obtaining higher expected utilities than all the others, denoted by π^*.

$$\pi_s^* = \arg max\, U^{\pi}(s)$$

Q4(b) Answer:

Use **value iteration** because there are many states and only 4 possible actions per state. Therefore, the value iteration is cheaper than policy iteration.

Q4(c) Answer:

Use **policy iteration** because teleporting the robot from one state to another cause a number of actions per state instead of 4. Therefore, the policy iteration is better than value iteration when there is a lot of actions.

Q4(d) Answer:

Table 1. Utilities/Values after 1 iteration

wall	-100	-100	-100	-100	-100	wall
1	-17.28	-18.0	-18.0	-18.0	-10.8	10
wall	-100	-100	-100	-100	-100	wall

Table 2. Utilities/Values after 2 iteration

wall	-100	-100	-100	-100	-100	wall
1	-17.28	-30.44	-36.96	-25.78	-10.8	10
wall	-100	-100	-100	-100	-100	wall

Q4(e) Answer:

The increment of the discount value wouldn't change the failure of crossing the bridge, since the trending of the grids' utilities/values remains the same. But if the discount value increases, the convergence needs more iterations to complete.

Q4(f) Answer:

The new value for the utility of the goal state is ***179*** or ***other bigger values***.

Table 3: Utility of the grid when Goal Value is **179**

wall	-100	-100	-100	-100	-100	wall
1	-17.21 >>	1.09 >>	26.52 >>	61.83 >>	110.88 >>	179
wall	-100	-100	-100	-100	-100	wall

Note that the reason of choosing the utility of the goal state is to ensure that the utility of the second grid is bigger than that of the starting grid, 1, in which case when the robot moves to the first grid, it would continue to move right instead of returning to the starting point.

Q5: Bayes Networks and Knowledge Representation

Q5(a) Answer:

Let the probability that a patient has disease A be denoted by $P(A)$, then

$$\boldsymbol{P(A)} = P(A|G)P(G) = 1 \times 0.1 + 0.9 \times 0.1 = \mathbf{0.19}$$

$$\boldsymbol{P(A)} = P(a = 1) = \sum_{s}\sum_{b}\sum_{g} P(s|a = 1, b)P(a = 1|g)P(g)P(b)$$

$$= \sum_{b} \sum_{g} P(a = 1|g) P(g) P(b) \sum_{s} P(s|a = 1, b)$$

$$= \sum_{b} \sum_{g} P(a = 1|g) P(g) P(b) \times 1 = \mathbf{0.19}$$

Q5(b) Answer:

$$P(A|B) = P(a = 1|b = 1) = \frac{P(a = 1, b = 1)}{P(b = 1)}$$

Since,

$$P(b = 1) = \sum_{s} \sum_{a} \sum_{g} P(s|a, b = 1) P(a|g) P(g) P(b = 1) = 0.4$$

$$P(a = 1, b = 1) = \sum_{s} \sum_{g} P(s|a = 1, b = 1) P(a = 1|g) P(g) P(b = 1) = 0.076$$

Therefore,

$$\boldsymbol{P(A|B) = 0.19}$$

Q5(c) Answer:

$$P(A|B, S) = P(a = 1|b = 1, s = 1) = \frac{P(a = 1, b = 1, s = 1)}{P(b = 1, s = 1)}$$

Since,

$$P(a = 1, b = 1, s = 1) = 0.076$$
$$P(b = 1, s = 1) = 0.3352$$

Therefore,

$$\boldsymbol{P(A|B, S) = 0.2267}$$

Q6: Search

Q6(a) Answer:

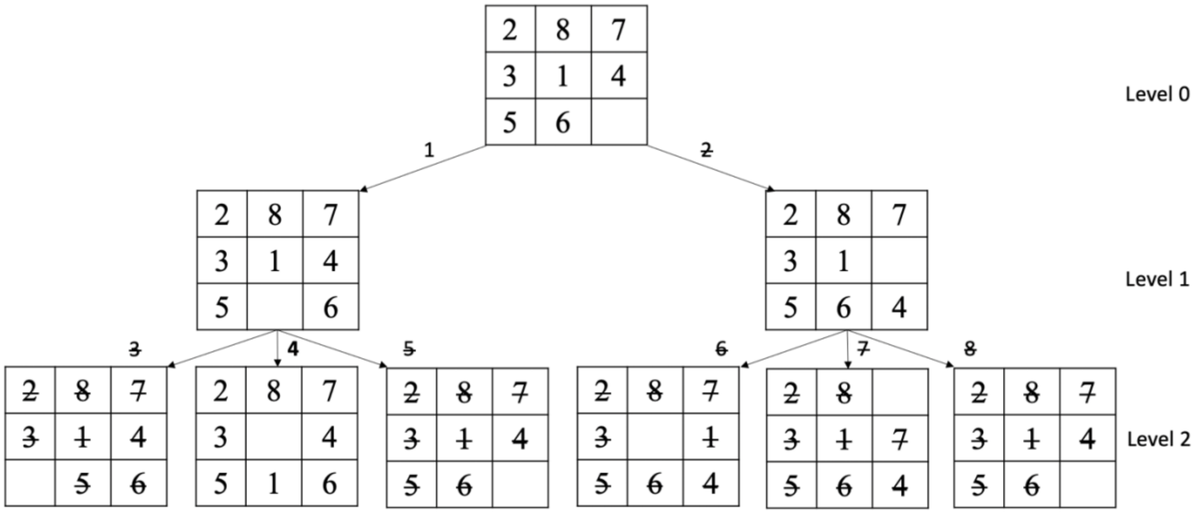

Figure 4. Breadth-first searching algorithm

Since there is a requirement about moving preference, **some branches and nodes are pruned**, only branch 1 and 4 are remained, as Figure 4 illustrates.

Q6(b) Answer:

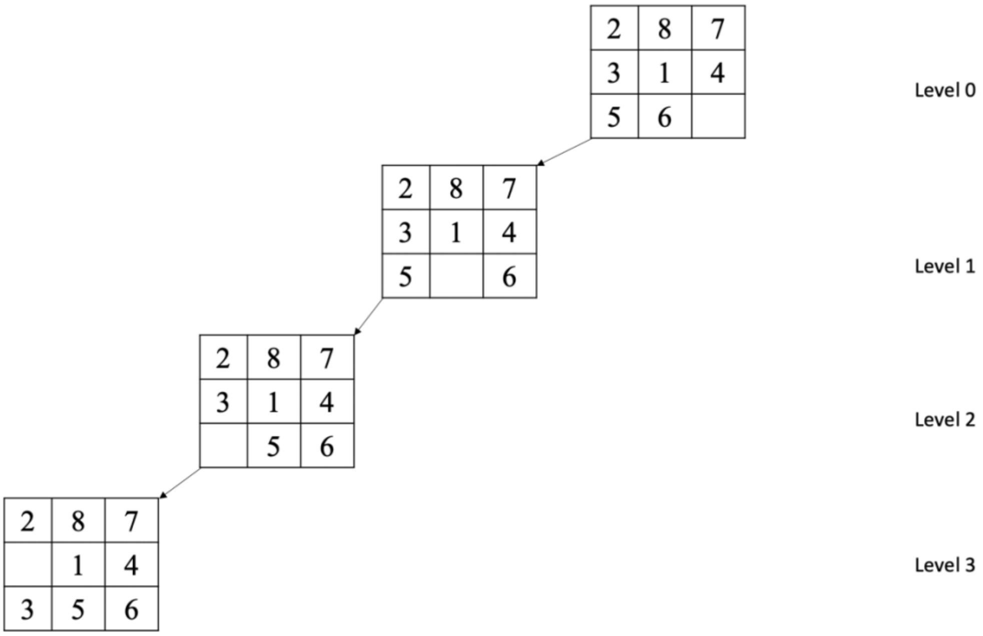

Figure 5. Depth-first searching algorithm

Q6(c) Answer:

Let the two heuristic functions be:

$$h_1 = \text{Number of Misplaced Tiles}$$
$$h_2 = \text{Total Manhattan Distance}$$

Therefore,

$$h_1 = 6$$
$$h_2 = 1 + 2 + 4 + 3 + 2 + 2 = 14$$

Q6(d) Answer:

Table 4. Advantages and Disadvantages

	Advantages	***Disadvantages***
Breadth-first Search	• Guaranteed to find a/an (optimal) solution; • Never trapped by unwanted nodes.	• Cost more time; • Consume more memory when searching.
Depth-first Search	• Consume less memory when searching; • Efficient when confronting an abundance of solutions.	• Not guaranteed to find a solution; • Cost much time when the solution is hard to find.
Heuristic search	• Guaranteed to find a reasonable solution in an acceptable time; • Efficient with the cost of incompleteness.	• Might not find the optimal solution.