Question 1 Nearest Neighbours and the Curse of Dimensionality.

(a) First, consider two independent univariate random variables X and Y sampled uniformly from the unit interval [0,1]. Determine the expectation and variance of the random variable Z, defined as the squared distance $Z = (X - Y)^2$.

I use $\mathbb{E}[Z]$ to represent the expectation of random variable Z.
From the properties of expectation, we have

$$\mathbb{E}[Z] = \mathbb{E}[(X-Y)^2] = \mathbb{E}[X^2 - 2XY + Y^2] = \mathbb{E}[X^2] - 2\mathbb{E}[XY] + \mathbb{E}[Y^2]$$

For random variables X sampled uniformly from the interval [a,b], its expectation

$$\mathbb{E}[X] = \int_a^b xf(x)dx = \int_a^b x\frac{1}{b-a}dx$$

According to the definition of expectation, the expectation of X^2 is

$$\mathbb{E}[X^2] = \int_a^b x^2\frac{1}{b-a}dx$$

Put in $a = 0$ and $b = 1$, we get

$$\mathbb{E}[X] = \int_0^1 xdx = \frac{1}{2}$$
$$\mathbb{E}[X^2] = \int_0^1 x^2dx = \frac{1}{3}$$

Y has the same distribution as X, hence,

$$\mathbb{E}[Y] = \int_0^1 ydy = \frac{1}{2}$$
$$\mathbb{E}[Y^2] = \int_0^1 y^2dy = \frac{1}{3}$$

According the properties of covariance,

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] + Cov[X,Y]$$

Due to the independence between X and Y,

$$Cov[X,Y] = 0$$

Hence,

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] = \frac{1}{4}$$

Then,

$$\mathbb{E}[Z] = \mathbb{E}[X^2] - 2\mathbb{E}[XY] + \mathbb{E}[Y^2] = \frac{1}{3} - 2\times\frac{1}{4} + \frac{1}{3} = \frac{1}{6}$$

Question 1 [(a)] continued on next page...

I use $Var[Z]$ to represent the variance of random variable Z.

$$\begin{aligned}
Var[Z] &= Var[(X-Y)^2] = \mathbb{E}[(X-Y)^4] - [\mathbb{E}[(X-Y)^2]]^2 = \mathbb{E}[(X-Y)^4] - \frac{1}{36} \\
&= \mathbb{E}[X^4 - 4X^3Y + 6X^2Y^2 - 4XY^3 + Y^4] - \frac{1}{36} \\
&= \mathbb{E}[X^4] - 4\mathbb{E}[X^3Y] + 6\mathbb{E}[X^2Y^2] - 4\mathbb{E}[XY^3] + \mathbb{E}[Y^4] - \frac{1}{36} \\
&= \mathbb{E}[X^4] - 4\mathbb{E}[X^3]\mathbb{E}[Y] + 6\mathbb{E}[X^2]\mathbb{E}[Y^2] - 4\mathbb{E}[X]\mathbb{E}[Y^3] + \mathbb{E}[Y^4] - \frac{1}{36} \\
&= \int_0^1 x^4dx - 4 \times \int_0^1 x^3dx \times \int_0^1 ydy + 6 \times \int_0^1 x^2dx \times \int_0^1 y^2dy - 4 \times \int_0^1 xdx \times \int_0^1 y^3dy + \int_0^1 y^4dy - \frac{1}{36} \\
&= \frac{1}{5} - \frac{1}{2} + \frac{2}{3} - \frac{1}{2} + \frac{1}{5} - \frac{1}{36} = \frac{7}{180}
\end{aligned}$$

I tried to verify the results with the help of a simple program.
I generated 10 datasets of X and Y. Each dataset contains 1 million random variables of X^* and Y^* sampled from uniform distribution. With a function $Z = (X-Y)^2$, the program calculated the means and variances of the results of the function, Z^*, given the datasets of X^* and Y^*. Then, I calculated the mean of the 10 datasets to get the final answer:
$E(Z^*) = 0.16664879, Var(Z^*) = 0.03887961$
which is close to what we calculated.

(b) Now suppose we sample two points independently from a unit cube in d dimensions. Observe that each coordinate is sampled independently from [0, 1], i.e. we can view this as sampling random variables $X_1, ..., X_d, Y_1, ..., Y_d$ independently from [0, 1]. The squared Euclidean distance can be written as $R = Z_1 + ... + Z_d$, where $Z_i = (X_i - Y_i)^2$. Using the properties of expectation and variance, determine $\mathbb{E}[R]$ and $Var[R]$.

Using the properties of expectation,

$$\mathbb{E}[R] = \mathbb{E}[Z_1 + Z_2 + ... + Z_d] = \mathbb{E}[Z_1] + ... + \mathbb{E}[Z_d] = \sum_{i=1}^{d} \mathbb{E}[Z_i] = d \times \mathbb{E}[Z] = \frac{d}{6} \approx 0.167d$$

Using the properties of variance,

$$Var[R] = Var[\sum_{i=1}^{d} Z_i] = \sum_{i=1}^{d} Var[Z_i] + \sum_{i \neq j} Cov[Z_i, Z_j]$$

Since all the random variables $X_1, ..., X_d, Y_1, ..., Y_d$ are independent, $Z_1, ..., Z_d$ are also independent. As a result,

$$Cov[Z_i, Z_j] = 0, for\ i, j \in [1, d]\ and\ i \neq j$$

Consequently,

$$Var[R] = \sum_{i=1}^{d} Var[Z_i] = d \times Var[Z] = \frac{7d}{180} \approx 0.0389d$$

To verify the results with the help of a python program, I generalized a set of R^*, with $d = 10$,
$E(R^*) = 1.66616071, Var(R^*) = 0.38834862$, which is close to what we calculated.

Question 1 continued on next page...

(c) Based on your answer to part (b), compare the mean and standard deviation of R to the maximum possible squared Euclidean distance (i.e. the distance between opposite corners of the cube). Why does this support the claim that in high dimensions, "most points are far away, and approximately the same distance"?

The standard deviation of R is,

$$Std[R] = \sqrt{Var[R]} = \sqrt{\frac{7d}{180}} \approx 0.197\sqrt{d}$$

The $X_1, ..., X_d$ and $Y_1, ..., Y_d$ can be seen as two points with d features. R is the Euclidean distance between the two points. The expectation of the distance is about $0.167d$ and its variance is about $0.197\sqrt{d}$. Because the mean and the variance of the distance are different orders of magnitude, when the number of dimensions increase, the distance between points gets really big.

Question 2 Decision Trees.

(a) See hw1_code.py

(b) Write a function select_model which trains the decision tree classifier using at least 5 different values of max_depth, as well as two different split criteria (information gain and Gini coefficient), evaluates the performance of each one on the validation set, and prints the resulting accuracies of each model.

The output of this function is shown in Table 1.

Depth	Gini	Entropy
5	0.7367	0.7388
10	0.7653	0.7592
12	0.7673	0.7592
14	0.7673	0.7653
15	0.7714	0.7531
16	0.7694	0.7490
18	0.7755	0.7592
20	**0.7776**	0.7531
25	0.7694	0.7714
30	0.7653	0.7612
35	0.7776	0.7653
40	0.7612	0.7755

Table 1: Train the classifier using different values of max_depth, as well as two different split criteria (information gain and Gini coefficient, "Entropy" and "Gini").

(c) Now let's stick with the hyperparameters which achieved the highest validation accuracy. Extract and visualize the first two layers of the tree. Your visualization may look something like what is shown below, but it does not have to be an image: it is perfectly fine to display text. It may be hand-drawn.

With the parameters max_depth as 20 and criteria as "Gini", the visualization of the first two layers of the decision tree is shown in Figure 1.

Question 3 Information Theory.

Question 3 continued on next page...

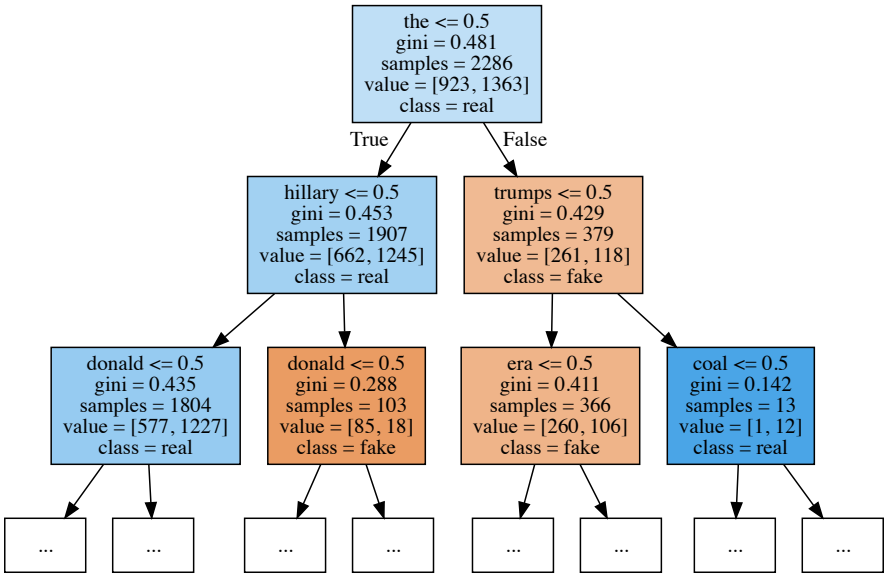

Figure 1: Visualization of the first two layers of the decision tree.

(a) Prove that the entropy $H(X)$ is non-negative.

The $p(x)$ is the probability mass function; consequently, $0 \leq p(x) \leq 1$.
As a result, $\frac{1}{p(x)} \geq 1$; then, $log_2(\frac{1}{p(x)}) \geq 0$.
For each term in the summation H(X), $p(x)log_2(\frac{1}{p(x)}) \geq 0$.
So we have

$$H(X) = \sum_x p(x)log_2(\frac{1}{p(x)}) \geq 0$$

(b) Prove that $KL(p||q)$ is non-negative. $KL(p||q) = \sum_x p(x)log_2\frac{p(x)}{q(x)}$

Apply Jensen's inequality and the concavity of log,

$$\begin{aligned} -KL(p||q) &= -\sum_x p(x)log_2\frac{p(x)}{q(x)} = \sum_x p(x)log_2\frac{q(x)}{p(x)} \\ &= \mathbb{E}[log_2\frac{q}{p}] \leq log_2(\mathbb{E}[\frac{q}{p}]) \\ &= log_2(\sum_x p(x)\frac{q(x)}{p(x)}) = log_2 1 = 0 \end{aligned}$$

As a result, $KL(p||q) \geq 0$.

(c) The Information Gain or Mutual Information between X and Y is $I(Y;X) = H(Y) - H(Y|X)$. Show that $I(Y;X) = KL(p(x,y)||p(x)p(y))$, where $p(x) = \sum_y p(x,y)$ is the marginal distribution of X.

Question 3 [(c)] continued on next page...

The proof is as follows.

$$\begin{aligned} KL(p(x,y)||p(x)p(y)) &= \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y) log p(x,y) - \sum_{x,y} p(x,y) log p(x) - \sum_{x,y} p(x,y) log p(y) \\ &= -\sum_{x,y} p(x,y) log \frac{1}{p(x,y)} + \sum_{x,y} p(x,y) log \frac{1}{p(x)} + \sum_{x,y} p(x,y) log \frac{1}{p(y)} \\ &= -H(X,Y) + H(X) + H(Y) \end{aligned}$$

Using the chain rule of entropy,

$$\begin{aligned} KL(p(x,y)||p(x)p(y)) &= -H(X,Y) + H(X) + H(Y) \\ &= H(Y) - H(Y|X) \\ &= I(Y;X) \end{aligned}$$