HW Solution

CS 512: Data Mining, Fall 2022 Version: **1.0**

Submitted by:

- **Yifei Liu**: **yifeil6**

1

Solution:

Holer

2

Solution:

2.A. Numerator is the number of samples o' "near" the sample o, denominator is the number of samples in D, thus LHS is the ratio of samples from D that is close to o.

Therefore this ratio is less than $\pi \Leftrightarrow$ few samples from D is close to $o \Leftrightarrow o$ is a distance-based outlier.

2.B.

$$\begin{aligned}
&\frac{\|o'|dist(o,o') \leq r\|}{\|D\|} \leq \pi \\
&\Leftrightarrow \|o'|dist(o,o') \leq r\| \leq \|D\|\pi \\
&\Leftrightarrow \|o'|dist(o,o') \leq r\| < \lceil \pi\|D\| \rceil \\
&\Leftrightarrow \|o'|dist(o,o') > r\| > \lceil \pi\|D\| \rceil \\
&\Leftrightarrow \|o'|dist(o,o') > r\| > k
\end{aligned}$$

Therefore if $dist(o,o_k) > r \forall o_k$, then $\|o'|dist(o,o') > r\| > k$, thus o is an outlier. ■