Q2) a) $\text{Conf}(A \to B) = \Pr(B|A)$

$$= \frac{\Pr(B \wedge A)}{P(A)}$$

which ignores $P(B)$. In case $P(B)$ is very high or 1, ie, B occurs in every bucket, this rule does not have any relevance even though it has high confidence.

$$\text{Lift}(A \to B) = \frac{\text{Conf}(A \to B)}{S(B)} = \frac{\Pr(A \wedge B)}{\Pr(A)\Pr(B)}$$

Lift does not suffer from this it takes $P(B)$ into consideration and ~~woul~~ would hence be lower for very high $P(B)$.

$$\text{Conv}(A \to B) = \frac{1 - S(B)}{1 - \text{Conf}(A \to B)} = \frac{1 - \Pr(B)}{1 - \text{conf}(A \to B)}$$

This is also safe from this drawback, since it has a $P(B)$ factor in the numerator. What this causes is that the metric goes up as confidence increases but goes down in case $P(B)$ itself is too high.

Q2) b): $\text{Conf}(A \rightarrow B) = \dfrac{\Pr(A \wedge B)}{P(A)}$.

$$\text{conf}(B \rightarrow A) = \frac{\Pr(A \wedge B)}{P(B)}$$

conf $(A \rightarrow B)$ will be equal to conf $(B \rightarrow A)$ only in case $P(A) = P(B)$. Since this may not be always true, confidence is not symmetric.

$$\text{Lift}(A \rightarrow B) = \frac{\Pr(A \wedge B)}{\Pr(A)\Pr(B)}$$

$$\text{Lift}(B \rightarrow A) = \frac{\Pr(B \wedge A)}{\Pr(B)\Pr(A)}$$

As we can see, Lift $(A \rightarrow B)$ will always be equal to Lift $(B \rightarrow A)$ hence it is symmetrical.

$$\text{conv}(A \rightarrow B) = \frac{1 - \Pr(B)}{1 - \text{conf}(A \rightarrow B)} = \frac{1 - \Pr(B)}{1 - \frac{\Pr(A \wedge B)}{\Pr(A)}}$$

$$= \frac{\Pr(A) - \Pr(B)\Pr(A)}{\Pr(A) - \Pr(A \wedge B)}$$

contd.

Similarly,

$$\text{Conv}(B \to A) = \frac{Pr(B) - Pr(B)\,Pr(A)}{Pr(A) - Pr(A \cap B)}$$

Hence, Conv $(A \to B)$ will be equal to Conv $(B \to A)$ only in case $P(A) = P(B)$. Since that may not always be true, Conviction in general is not symmetric.

Q2) c). $conf(A \to B) = \frac{Pr(A \wedge B)}{Pr(A)}$

In case $A \to B$ is a perfect implication, B is always in a basket that contains A. i.e. $P(A \wedge B) = P(A)$.

$\Rightarrow Conf(A \to B) = \frac{Pr(A \wedge B)}{Pr(A)} = \frac{Pr(A)}{Pr(A)} = 1$

which is maximal since a probability can not be > 1.

Hence conf is desirable with max. of 1.

$Lift(A \to B) = \frac{Conf(A \to B)}{Pr(B)} = \frac{Pr(A \wedge B)}{Pr(A)\,Pr(B)}$

Similar to above, in case of perfect implication, $conf(A \to B) = 1$.

$\Rightarrow Lift(A \to B) = \frac{1}{Pr(B)}$.

For a given dataset, $Pr(B)$ can be considered constant. Thus the numerator is maximum when the rule is a perfect implication.

Therefore, lift is also desirable with maximum value of $\frac{1}{Pr(B)}$.

$$\text{conv}(A \to B) = \frac{1 - Pr(B)}{1 - \text{conf}(A \to B)}$$

As we saw, in case of perfect implication
$\text{conf}(A \to B) = 1$

$\Rightarrow$ as $\text{conf}(A \to B) \to 1$
$\text{conv}(A \to B) \to \infty$.

As conf approaches 1, conv. approaches ∞.

Hence, with increase in conf. the conv is higher.
Conviction is also desirable with a max val of ∞.

☆ A special case here is when
$P(B) = 1$. $\Rightarrow \text{conf}(A \to B) = 1$.
in which case it becomes undefined.

(Q3) a) To prove:

$$d(x,y) + d(y,z) \geq d(x,z) \quad \forall (x,y,z)$$

where,

$$d(x,y) = 1 - sim(x,y)$$
$$= 1 - Pr[h(x) = h(y)]$$
$$= Pr[h(x) \neq h(y)] \quad \text{—— (A)}$$

We can make the following observations -

(1) The event $h(x) \neq h(y)$ $\forall (x,y)$ is a binary event and can take values true or false.

(2) $h(x) \neq h(y)$ implies that ~~both~~ ~~either~~ ~~atleast~~ one of the following must hold

- $h(x) \neq h(z)$
- $h(y) \neq h(z)$

Because ~~if~~ both do not hold, by trasitivity property of equality, $h(x) = h(z) = h(y)$ which is against our assumption above.

Hence,

We can say that for event $h(x) \neq h(y)$ to occur, one of $h(x) \neq h(z)$ or $h(y) \neq h(z)$ must occur.

Therefore,

$$Pr[h(x) \neq h(y)] \leq Pr[h(x) \neq h(z)] + Pr[h(y) \neq h(z)]$$

from equation (A)

$$\boxed{d(x,y) \leq d(x,z) + d(y,z)}$$

Q3) b) $Sim_{over}(A,B) = \dfrac{|A \cap B|}{\min(|A|, |B|)}$

Assume $A = \{1, 2, 3\}$
$B = \{1\}$
$C = \{2\}$

$Sim_o(A,B) = 1/1 = 1$

$Sim_o(A,C) = 1/1 = 1$

$Sim_o(B,C) = 0/1 = 0$

$\Rightarrow$ $d(A,B) = 1 - 1 = 0$
$d(A,C) = 1 - 1 = 0$
$d(B,C) = 1 - 0 = 1$

These distances do not obey the triangle inequality as
$d(A,B) + d(A,C) \not\geq d(B,C)$

Hence, there is no LSH scheme for overlap similarity.

Q3.) c) $Sim_{Dice}(A,B) = \dfrac{|A \wedge B|}{\frac{1}{2}(|A| + |B|)}$

$$= \frac{2\,|A \wedge B|}{|A| + |B|}$$

Assume $A = \{1, 2\}$

$B = \{1\}$

$C = \{2\}$

$Sim_D(A,B) = \dfrac{2 \times 1}{3} = 2/3$

$Sim_D(A,C) = \dfrac{2 \times 1}{3} = 2/3$

$Sim_D(B,C) = 0/3 = 0$

$\Rightarrow$ $d(A,B) = 1 - 2/3 = 1/3$

$d(A,C) = 1 - 2/3 = 1/3$

$d(B,C) = 1 - 0 = 1$

These distances do not obey the triangle inequality as

$d(A,B) + d(A,C) \not\geq d(B,C)$

Hence this similarity does not have a LSH scheme.

Q4.) a) $W_j = \{x \in A : g_j(x) = g_j(z)\}$ $(1 \leq i \leq L)$

That is W_j is the set of elements in the same bucket as z hashed by g_j

Also, $T = \{x \in A : d(x, z) > c\lambda\}$.

$\Rightarrow Pr[g(x) = g(z)]$
$= [Pr[h(x) = h(z)]]^k$

$\because$ H is $(\lambda, c\lambda, P_1, P_2)$-sensitive

$$[Pr[h(x) = h(z)]]^k \leq p_2^k$$
$$\leq p_2^{\lceil \log_{1/p_2} n \rceil}$$
$$\leq n^{\log_{1/p_2} p_2}$$
$$\leq n^{-\log_{1/p_2} 1/p_2}$$
$$\leq n^{-1}$$

$\Rightarrow Pr(g(x) = g(z)) \leq 1/n$. — (A)

To prove

$$Pr\left[\sum_{1}^{L} |T \cap W_j| \geq 3L\right] \leq 1/3$$

Using markov inequality on LHS.

$$\Pr\left[\sum_1^L |T \cap w_j| \geq 3L\right] \leq \frac{E\left[\sum_1^L |T \cap w_j|\right]}{3L} \quad \text{— (B)}$$

LHS represents the probability that all the 3L points that we gather from L buckets are greater than $c\lambda$ from the query point, ie, an error condition.

From (A), we know than for pts (x, z) such that $d(x,z) \geq c\lambda$

$$\Pr(g(x) = g(z)) \leq 1/n$$

suppose t elements from T fall into w_j for any j

$\Rightarrow \Pr[\text{having } t \text{ elements from } T \text{ in } w_j] \leq 1/n^t$ — (C)

From equation (B),

$$\frac{E\left[\sum_1^L |T \cap w_j|\right]}{3L}$$

$$= \frac{E[|T \cap w_1|] + E[|T \cap w_2|] + \cdots + E[|T \cap w_L|]}{3L}$$

From (C)

$$\leq \frac{1 + 1 \cdots 1}{3L} \leq \frac{1k}{3k} \leq 1/3$$

Q4) b) $x^* \in A$: $d(x^*, z) \leqslant \lambda$

To prove -

$$\Pr[\forall 1 \leq j \leq L, g(x^*) \neq g(z)] \leq 1/e.$$

$$\Pr[\forall 1 \leq j \leq L, g(x^*) \neq g(z)]$$

$$= [\Pr[g(x^*) \neq g(z)]]^L$$

(x^* & z do not hash to any of the buckets)

$$= [1 - \Pr[g(x^*) = g(z)]]^L$$

$\because$ $g \in G$ is an and construct for $h \in H^k$, we get

$$= [1 - [\Pr[h(x^*) = h(z)]]^k]^L \quad \text{—— (A)}$$

Also, H is a family with $(\lambda, c\lambda, p_1, p_2)$ sensitive

$$\Pr[h(x^*) = h(z)] \geqslant p_1$$

$$\Rightarrow 1 - \Pr[h(x^*) = h(z)] \leq 1 - p_1$$

(A) becomes -

$$[1 - [\Pr[h(x^*) = h(z)]]^k]^L \leq$$

$$[1 - p_1^k]^L \quad \text{—— (B)}$$

We know.

$$k = \log_{1/p_2} n$$

Hence,

$$p_1^k = p_1^{[\log_{1/p_2} n]}$$

$$= n^{[\log_{1/p_2} p_1]} \quad \text{(base shift)}$$

$$= n^{\left[-\frac{\log_e 1/p_1}{\log_e 1/p_2}\right]} \quad \text{(base change)}$$

$$= n^{-e} \quad \text{where } e = \frac{\log 1/p_1}{\log 1/p_2}.$$

Equation (B) becomes

$$\Rightarrow [1 - n^{-e}]^L$$

Since, $\forall x \in R$, $(1 - x/n)^n \leq e^{-x}$

$$\Rightarrow (1-x)^1 \leq e^{-x}$$

$$\Rightarrow [1 - n^{-e}] \leq e^{-n^{-e}}$$

$$\Rightarrow [1 - n^{-e}]^L \leq [e^{-n^{-e}}]^L$$

$$\leq e^{-L/n^e}$$

$$\because L = n^e$$

$$\Rightarrow [1 - n^{-e}]^L \leq e^{-1} \leq 1/e$$

$$\Rightarrow \boxed{P_r[\forall 1 \leq j \leq L, g(x^*) \neq g(z)] \leq 1/e}$$

Q4) c) ~~fo~~ To prove:

Point chosen is $c\lambda$-ANN

That is, the pt chosen (x) is such that

$d(x,z) \leq c\lambda$, where z is the query.

We know that x is a point chosen uniformly from L buckets and is among the ~~to~~tal of $3L$. In case the total of L buckets $> 3L$, then from part 4@ we know the probability

$\boxed{\text{Pr[Choosing } 3L \text{ from } L \text{ buckets where all points are greater than } c\lambda \text{ dist]} \leq 1/3}$ — (A).

Also, from 4(b) we know that for a pt.

$x^* \in A: d(x^*, z) \leq \lambda$

$pr[g_j(x^*) \neq g_j(z), 1 \leq j \leq L] \leq 1/e$.

Suppose there are q pts. that are within λ distance from z.

$\boxed{\text{Pr[none of } q \text{ pts map to same bucket as } z] \leq 1/e^q}$ — (B)

~~From (A), (B)~~

~~Pr[point cho~~

from (A) & (B) equations

$\Pr[\text{point chosen has dist} \geq c\lambda] \leq \frac{1}{3} + \frac{1}{e^{q}}$

$\Pr[\text{point chosen is } (c,\lambda)\text{-ANN}] \geq 1 - \frac{1}{3} - \frac{1}{e^{q}}$

$\geq \frac{2}{3} - \frac{1}{e^{q}}$