2 (a) In the extreme case when $Pr(B) = 1$

A. B are statistically independent but $conf(A \to B) = 1$

On the other hand, $Pr(B)$ is related to $S(B)$, so

$lift(A \to B)$ and $conv(A \to B)$ are able to capture the information of $Pr(B)$.

(b) i. $conf(A \to B)$

When $Pr(B) = 1$

$$conf(A \to B) = conf(B \to A)$$

iff $$1 = Pr(A|B)$$

iff $$Pr(A) = 1$$

so $conf(A \to B)$ is not always symmetric.

ii $lift(A \to B)$

$$lift(A \to B) = \frac{conf(A \to B)}{S(B)} = \frac{S(A \cup B)}{S(A) \cdot S(B)} = \frac{conf(B \to A)}{S(A)} = lift(B \to A)$$

iii $conv(A \to B)$

$$conv(A \to B) = \frac{1 - S(B)}{1 - \frac{S(A \cup B)}{S(A)}} = \frac{S(A) - S(A) \cdot S(B)}{S(A) - S(A \cup B)} = 1 + \frac{S(A \cup B) - S(A) \cdot S(B)}{S(A) - S(A \cup B)}$$

When $S(A) \neq S(B)$ and $S(A \cup B) \neq S(A) \cdot S(B)$

$$conv(A \to B) \neq conv(B \to A)$$

(c) i $\operatorname{conf}(A \to B)$

For perfect implication $A \to B$,

$\operatorname{conf}(A \to B) = 1$ = maximumm achievable value

ii $\operatorname{lift}(A \to B)$

For perfect implication $A \to B$,

$\operatorname{lift}(A \to B) = \dfrac{1}{s(B)} \neq N$ (maximum achievable value) in general,

so lift is not desirable.

iii $\operatorname{conv}(A \to B)$

For perfect implication $A \to B$,

$\operatorname{conv}(A \to B) = \dfrac{1 - s(B)}{0} = +\infty$ (ignore $\frac{0}{0}$ cases)

so conv is desirable.

(d) (e) See the codes

Q3 (a) Let E_i be the event that the i-th non-zero column is not chosen.

Then $P(E_1 \cdots E_m) = P(E_1) \cdot P(E_2|E_1) \cdots P(E_1 \cdots E_m | E_1 \cdots E_{m-1})$

$$\leq P(E_1) \cdot P(E_2) \cdots P(E_m)$$

$$= (P(E_1))^m = \left(\frac{n-k}{n}\right)^m$$

Note that $P(E_2|E_1) \leq P(E_2)$ (and similar inequalities) holds naturally in this case.

(b) $\left(\frac{n-k}{n}\right)^m = \left(1 - \frac{k}{n}\right)^m = \left(1 - \frac{1}{\frac{n}{k}}\right)^m$

We can make $m \approx 10 \frac{n}{k}$

i.e $k \approx 10 \frac{n}{m}$

then $\left(1 - \frac{1}{\frac{n}{k}}\right)^m \approx \left(1 - \frac{1}{\frac{n}{k}}\right)^{10\frac{n}{k}} \approx e^{-10}$

(c)

S_1	S_2
1	0
0	0
1	1

Jaccard similarity $= \frac{1}{2}$

random cyclic permutation gives $\frac{2}{3}$

Q4. (a) By Markov's inequality

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

$$= \frac{E[|T \cap W_1|]}{3}$$

since $\mathcal{H}$ is $(\lambda, c\lambda, P_1, P_2)$-sensitive

$$\leq \frac{n P_2^{\,k}}{3}$$

$$= \frac{n \cdot \frac{1}{n}}{3} = \frac{1}{3}$$

(b) $$\Pr[\forall\, 1 \leq j \leq L,\ g_j(x^*) \neq g_j(z)] \leq (1 - P_1^{\,k})^L$$

$$= \left(1 - P_1^{\log_{1/P_2}(n)}\right)^{n^{\frac{\log(\frac{1}{P_1})}{\log(\frac{1}{P_2})}}}$$

$$= \left(1 - \frac{1}{P_1^{\log_n(\frac{1}{P_2})}}\right)^{P_1^{\log_n(\frac{1}{P_2})}} < \frac{1}{e}$$

(c) By (a) (b), with probability $\geq \left(1 - \frac{1}{3} - \frac{1}{e}\right)$

we have $< 3L$ points in T chosen in step 3

and at least 1 point not in T chosen in step 3.

(d) See the codes