(a) [15 分]

使用置信度的一个缺点是它忽略了 $Pr(B)$。为什么这是一个缺点？请解释一下提升度和确信度为什么不受这个缺点的影响。

(a). $Coef(A\to B) = Pr(B|A) = \frac{S(AB)}{S(A)}$

由于 $Coef(A\to B)$ 忽略了 $Pr(B)$. 所以当 $Pr(B)$ 取一些极大值，如1时，B在每个篮子中都出现. 此时. $Coef(A\to B)$ 也会很高. 但这是由于B的普遍存在导致的高Coef 而非 AB之间的关联规则导致的高coef. 所以置信度忽略 $Pr(B)$ 是一个缺点的原因是其没有排除一些由于物品B的高频次而导致高置信度的物品组合.

而提升度通过在 $Coef(A\to B)$ 的基础出除以 $S(B)$. 即除以B的频率. 抵消了由于B的高频率所带来的高 $Coef(A\to B)$ 的影响. 比如若 $Coef(A\to B)$ 是由于高频率B而导致的虚高. 会受到高压缩因子 $S(B)$ 的影响而降低. 故提升度不受这个缺点影响.

确信度与提升度相似., $1-S(B)$ 代表A,B独立时B不发生频率. $1-Coef(A\to B)$ 代表实际A发生但B不发生的频率. 同样通过 $1-S(B)$ 这个压缩因子避免了由于高频率B导致确信度虚高的问题. 故确信度不受这个缺点影响.

(b) [15 分]

如果 度量$(A \to B) =$ 度量$(B \to A)$，则称该度量是对称的。在前文介绍的几种度量中，有哪些是对称的？对于每个度量，请证明该度量是对称的，或者提供一个反例表明该度量不是对称的。

(b). 置信度:

取 $S(A)=0.5$ $S(B)=1$ $S(AB)=0.5$

则 $Coef(A\to B)=Pr(B|A)=\frac{S(AB)}{S(A)}=1$.

$Coef(B\to A)=Pr(A|B)=\frac{S(AB)}{S(B)}=0.5$

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

$\Rightarrow$. 置信度不对称.

提升度:

$lift(A\to B)=\frac{Conf(A\to B)}{S(B)}=\frac{S(AB)}{S(A)\cdot S(B)}$

$lift(B\to A)=\frac{Conf(B\to A)}{S(A)}=\frac{S(AB)}{S(A)\cdot S(B)}$

$\Rightarrow$. $lift(A\to B)=lift(B\to A)$

$\Rightarrow$. 提升度是对称的.

确信度:

取 $S(A)=0.4$ $S(B)=0.5$ $S(AB)=0.3$.

则 $Conv(A\to B)=\frac{1-S(B)}{1-Coef(A\to B)}=\frac{\frac{1}{2}}{1-\frac{3}{4}}=2$;

$Conv(B\to A)=\frac{1-S(A)}{1-Coef(B\to A)}=\frac{0.6}{1-0.6}=1.5$

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

$\Rightarrow$. 确信度不对称.

(c) [20 分]

完美蕴含是指相关的条件概率为 1。例如，$P(B|A) = 1$，则称 $A \to B$ 为完美蕴含。如果一个度量对于所有完美蕴含的情况，都能达到其最大可实现值，则称该度量是可取的。这使得我们很容易识别出最佳规则。在上述度量中，哪些是可取的？您可以忽略 0/0 的情况，但不能忽略其他无穷大的情况（即，最大值也可以为无穷大）。同时，您可以通过一个例子来简单解释。

(c) 置信度.

conf(A→B) 的最大可实现值为1

当 Pr(B|A)=1 时

conf(A→B) = Pr(B|A) = 1

⇒ 置信度是可取的.

提升度

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

∴ 当 S(B)→0 时. 其最大可实现值为 +∞.

反例：令. S(B)=0.5.

则 Pr(B|A)=1 时. lift(A→B)=2.

此时 lift(A→B) 未可取到其最大可实现值 +∞.

⇒ 提升度是不可取的.

确信度.

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

∴ 当 1-S(B)→0 而. 1-conf(A→B)→0 时.

Conv(A→B) 的最大可实现值为 +∞

当 Pr(B|A)=1 时. conf(A→B)=1 ⇒ 1-conf(A→B)=

忽略 1-S(B)=0 的情况.

此时 Conv(A→B) → +∞. 取到最大可实现值.

⇒ 确信度是可取的.