# Knowledge Discovery in Database - Assignment 1

**Alex Romelt, Alex Vonig, Yu Xiang**

Date: October 27, 2017

## Problem 1-1

(a) Classification; unsupervised

(b) Association analysis; supervised

(c) Outliers detection; unsupervised

(d) Association Analysis; unsupervised

(e) cluster analysis; unsupervised

## Problem 1-2

(a) A and B are not independent.
If A and B are independent, then

$$P(A \cap B) = P(A) * P(B)$$

However, $A \cap B = \emptyset$, $P(A \cap B) = 0$, while $P(A) \neq 0, P(B) \neq 0$, thus $P(A \cap B) \neq P(A) * P(B)$, which means A and B can not be independent.

(b) Proof of Bayes' Rule using only Conditional Probability.
Based on conditional probability, we have:

$$P(B|A) = \frac{P(AB)}{P(A)}, therefore, P(AB) = P(B|A)P(A)$$

So, we will have

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

## Problem 1-3

(a) Define $H1, T1; H2, T2$ as the possible results of the first and second coins; $\{i, i = 1, 2, ..., 12\}$ as the possible outcomes of the fair dice.
The sample space of the possible outcome is as follows:

$$\{H1H2i, H1T2i, T1H2i, T1T2i, i = 1, 2, ...12\}$$

The size of the sample space is 48.

(b) Define $I_{i>5}$ as the case that the dice shows a number greater than 5. Notice, the result of the dice is independent with the result of the coins, also the results of coins are independent with each other. Thus,

$$P(A_1) = P(H1H2I_{i>5}) = P(H1)P(H2)P(I_{i>5}) = \frac{1}{2} * \frac{1}{2} * \frac{7}{12} = \frac{7}{48}$$

(c) Only considering the result of the coins, the probablity of event $A$ (at least one head shown) is equal to one minus the probablity of event $B$ (no head shown). Thus $P(A) = 1 - P(B) = 1 - P(T1T2) = 1 - \frac{1}{2} * \frac{1}{2} = \frac{3}{4}$
The dice result is indepenent with the coins. Thus the probablity of at least one coins turns up (head shown) and the dice shows a number greater than 7 is:

$$P(A_2) = P(AI_{i>7}) = P(A)P(I_{i>7}) = \frac{3}{4} * \frac{5}{12} = \frac{5}{16}$$

(d) Event $A$ (no head shown), $B$(one head shown), $C$(two heads shown) are mutually exclusive, and sum together to a probablity of one. Thus

$$\begin{aligned}
P(A_3) &= P(A_3, A) + P(A_3, B) + P(A_3, C) \\
&= P(A_3|A)P(A) + P(A_3|B)P(B) + P(A_3|C)P(C) \\
&= 0 * P(A) + P(I_{i<=2})P(B) + P(I_{i<=4})P(C) \\
&= 0 * \frac{1}{4} + \frac{1}{6} * \frac{1}{2} + \frac{4}{12} * \frac{1}{4} \\
&= \frac{1}{6}
\end{aligned}$$

## Problem 1-4

(a) Define $S$ as studying computer science, $O$ as studying other subjects, we know

$$P(A_{42}|S) = 0.01, P(A_{42}|O) = 2 * 10^{-5}, P(S) = 0.25$$

Studying computer science and other subjects are mutually exclusive. Thus the probablity that $A_{42}$ is found:

$$P(A_{42}) = P(A_{42}, S)P(S) + P(A_{42}, O)P(O) = 0.01 * 0.25 + 2 * 10^{-5} * (1 - 0.25) = 2.515 * 10^{-3}$$

(b) (1) The probability that Computer Science is studied given $A_{42}$ is found:

$$P(S|A_{42}) = \frac{P(SA_{42})}{P(A_{42})} = \frac{P(A_{42}|S)P(S)}{P(A_{42})} = \frac{0.01 * 0.25}{2.515 * 10^{-3}} = 0.994$$

(2) Define $nA$ as event that $A_{42}$ is not found, the probability that Computer Science is studied given $A_{42}$ is not found:

$$P(S|nA) = \frac{P(S * nA)}{P(nA)} = \frac{P(nA|S)P(S)}{P(nA)} = \frac{(1 - P(A_{42}|S)P(S)}{P(nA)} = \frac{(1 - 0.01) * 0.25}{1 - 2.515 * 10^{-3}} = 0.248$$