AI3602 Data Mining: Homework 3

Xiangyuan Xue (521030910387)

1. Suppose that there are exactly k common elements between S and T. By definition, the Jaccard similarity between S and T is given by

$$J(S,T)=\frac{|S\cap T|}{|S\cup T|}=\frac{k}{2m-k}$$

At the same time, the probability of this event is given by

$$\mathbb{P}\left[|S\cap T|=k\right]=\mathbb{P}\left[J(S,T)=\frac{k}{2m-k}\right]=\frac{C_m^kC_{n-k}^{m-k}}{C_n^m}$$

If $n\geq 2m$ holds, the expected value of $J(S,T)$ is given by

$$\mathbb{E}\left[J(S,T)\right]=\sum_{k=0}^{m}\frac{k}{2m-k}\frac{C_m^kC_{n-k}^{m-k}}{C_n^m}$$

Otherwise, there are at least $2m-n$ common elements between S and T. Then the expected value of $J(S,T)$ should be rewritten as

$$\mathbb{E}\left[J(S,T)\right]=\sum_{k=2m-n}^{m}\frac{k}{2m-k}\frac{C_m^kC_{n-k}^{m-k}}{C_n^m}$$

To conclude, the expected value of Jaccard similarity between S and T is given by

$$\mathbb{E}\left[J(S,T)\right]=\sum_{k=\max\{0,2m-n\}}^{m}\frac{k}{2m-k}\frac{C_m^kC_{n-k}^{m-k}}{C_n^m}$$

where the summation cannot be further simplified.