CHAPTER 1. Data Mining

Raymond

November 20, 2022

1 Exercise 1.2.1

Original problem setting: 1. There are one billion people who might be evil-doers; 2. Everyone goes to a hotel one day in 100, i.e. has a probability of 0.01 to go to an arbitrary hotel in one arbitrary day; 3. A hotel holds 100 people; 4. We shall examine hotel records for 1000 days; 5. Def of evil-doers: a pair of people who on two different days were both at the same hotel.

Problem: $\mathbb{E}(\#Evil-doers)$

Step: 1) $\mathbf{P}$(Two people go to hotel on same day) $=0.01*0.01=0.0001$.

2) $\mathbf{P}$(Two people go to the same hotel on same day) $=0.0001*10^{-5}=10^{-9}$.

3) $\mathbf{P}$(Two people go to the same hotel on two same day) $=(10^{-9})^2=10^{-18}$.

4) #Pair of People$=C^2_{1,000,000,000}\approx(10^9)^2/2=5*10^{17}$, and #Pair of Day$=C^2_{1,000}\approx(10^3)^2/2=5*10^5$.

5) $\mathbb{E}(\#Evil-doers)=10^{-18}*10^{17}*10^5*25=2.5*10^5$

1.1 a. Number of Observation was Raised to 2000

#Pair of Day$=C^2_{2,000}\approx(2*10^3)^2/2=2*10^6$ changed, so $\mathbb{E}(\#Evil-doers)=10^{-18}*10^{17}*10^6*10=10^6$

1.2 b. Number of People was Raised to 2 billion with 200,000 hotels

$\mathbf{P}$(Two people go to the same hotel on same day) $=0.0001*0.5*10^{-5}=0.5*10^{-9}$.

$\mathbf{P}$(Two people go to the same hotel on two same day) $=(0.5*10^{-9})^2=2.5^{-19}$.

#Pair of People$=C^2_{2,000,000,000}\approx(2*10^9)^2/2=2*10^{18}$

$\mathbb{E}(\#Evil-doers)=2.5^{-19}*10^{18}*10^5*10=2.5*10^5$

1.3 b. Evil-doers: At the same hotel at three different day

Steps: 1) $\mathbf{P}$(Two people go to hotel on same day) $=0.01*0.01=0.0001$.

2) $\mathbf{P}$(Two people go to the same hotel on same day) $=0.0001*10^{-5}=10^{-9}$.

3) $\mathbf{P}$(Two people go to the same hotel on Three same day) $=(10^{-9})^3=10^{-27}$.

4) #Pair of People$=C^2_{1,000,000,000}\approx(10^9)^2/2=5*10^{17}$, and #Combination of Three Days$=C^3_{1,000}\approx(10^3)^3/2/2=2.5*10^8$.

5) $\mathbb{E}(\#Evil-doers)=10^{-27}*5*10^{17}*2.5*10^8=0.125$

2 Exercise 1.2.2

Step: 1) $\mathbf{P}$(Two people purchase the same set items) $=1/C^{10}_{1000}/C^{10}_{1000}\approx(1000^{-10}*10!)^2=(1000^{-10}*4*10^6)^2=1.6*10^{-7}$.

2) $\mathbf{P}$(Two people go to the supermarket on same day) $=365^{-2}$.

3) #Pair of People$=C^2_{100,000,000}\approx(10^8)^2/2=5*10^{15}$, and #Pair of Days$=C^2_{365}\approx 365^2/2$.

4) $\mathbb{E}(\#Terrorists)=1.6*10^{-7}*365^{-2}*365^2/2*5*10^{15}=4*10^8$

3 Exercise 1.3.1

$IDF = \log_2(N/n_i)$, so a) if it appears in 40 documents, $IDF = \log_2(10,000,000/40)$. b) if it appears in 10,000 documents, $IDF = \log_2(10,000,000/10,000)$

4 Exercise 1.3.2

$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$. If word w appears a) once, $TF.IDF = \frac{1}{15} * \log_2(10,000,000/320)$, b) five times, $TF.IDF = \frac{15}{15} * \log_2(10,000,000/320)$.

5 Exercise 1.3.3

This c cannot be divided by 3 or 5.

6 Exercise 1.3.4

Consider $(1+a)^b$, we can rewrite it as $(1+a)^{1/a}(ab)$, then substitute a $= \frac{1}{x}$ and $\frac{1}{a}$ = x, we will have $(1+\frac{1}{x})^{x(ab)}$. If a is small, we will have e^{ab} as the approximate to $(1+a)^b$.

a) $(1.01)^{500} = (1+\frac{1}{100})^{500} = e^{0.01*500} = e^5$
b) $(1.05)^{1000} = e^{0.05*1000} = e^{50}$
c) $(0.9)^{40} = e^{-0.1*40} = e^{-4}$