Algorithmic Data Science
1st Assignment

Nikos Stamatis
MSc Student
Graduate Programme in Data Science and Machine Learning
SN: 03400115
nikolaosstamatis@mail.ntua.gr

The following three exercises are taken from [LRU20].

Exercise 6.3.1

The supports of the 1-itemsets are $|\text{spt}\{1\}| = 4$, $|\text{spt}\{2\}| = 6$, $|\text{spt}\{3\}| = 8$, $|\text{spt}\{4\}| = 8$, $|\text{spt}\{5\}| = 6$ and $|\text{spt}\{6\}| = 4$, so they are all frequent when we use $s = 4$ as a threshold. For the 2-itemsets we have that $|\text{spt}\{1,2\}| = 2$, $|\text{spt}\{1,3\}| = 3$, $|\text{spt}\{1,4\}| = 2$, $|\text{spt}\{1,5\}| = 1$, $|\text{spt}\{1,6\}| = 0$, $|\text{spt}\{2,3\}| = 3$, $|\text{spt}\{2,4\}| = 4$, $|\text{spt}\{2,5\}| = 2$, $|\text{spt}\{2,6\}| = 1$, $|\text{spt}\{3,4\}| = 4$, $|\text{spt}\{3,5\}| = 4$, $|\text{spt}\{3,6\}| = 2$, $|\text{spt}\{4,5\}| = 3$, $|\text{spt}\{4,6\}| = 3$ and $|\text{spt}\{5,6\}| = 2$.

By computing the expression $h_1(x,y) = xy \mod 11$ for $x \neq y \in \{1,\dots,6\}$ we obtain that bucket 1 contains 5 elements (1 appearance of $\{2,6\}$ and 4 of $\{3,4\}$), bucket 2 contains 5 elements ($2\times\{1,2\}$, $3\times\{4,6\}$), bucket 3 contains 3 elements ($3\times\{1,3\}$), bucket 4 contains 6 elements ($2\times\{1,4\}, 4\times\{3,5\}$), bucket 5 contains 1 element ($1\times\{1,5\}$), bucket 6 contains 3 elements ($3\times\{2,3\}$), bucket 7 contains 2 elements ($2\times\{3,6\}$), bucket 8 contains 6 elements ($4\times\{2,4\}, 2\times\{5,6\}$), bucket 9 contains 3 elements ($3\times\{4,5\}$) and bucket 10 contains 2 elements ($2\times\{2,5\}$).

Out of them, the buckets 1, 2, 4 and 8 are frequent. Therefore, the candidate pairs are the ones that hash into one of these four buckets, namely

$$C_2 = \big\{\{2,6\},\{3,4\},\{1,2\},\{4,6\},\{1,4\},\{3,5\},\{2,4\},\{5,6\}\big\}.$$

Exercise 6.3.2

We only need to apply the second hash function $h_2(x,y) = x + y \mod 9$ to the 2-itemsets that were hashed into frequent buckets, namely the elements of C_2. After applying h_2 we obtain that bucket 1 contains 3 elements ($3\times\{4,6\}$), bucket 2 contains 2 elements ($2\times\{5,6\}$), bucket 3 contains 2 elements ($2\times\{1,2\}$), bucket 4 is empty, bucket 5 contains 2 elements ($2\times\{1,4\}$), bucket 6 contains 4 elements ($4\times\{4,2\}$), bucket 7 contains 4 elements ($4\times\{4,3\}$) and bucket 8 contains 5 elements ($4\times\{3,5\}, 1\times\{2,6\}$).

The buckets 6, 7 and 8 are frequent, therefore the candidate pairs are now

$$\tilde{C}_2 = \big\{\{2,6\},\{3,4\},\{3,5\},\{2,4\}\big\},$$

and have been reduced from 8 to 4.

Exercise 6.4.1

The supports of the 1-itemsets in the sample are $|\text{spt}\{1\}| = 1$, $|\text{spt}\{2\}| = 2$, $|\text{spt}\{3\}| = 3$, $|\text{spt}\{4\}| = 3$, $|\text{spt}\{5\}| = 2$ and $|\text{spt}\{6\}| = 1$, so they are all frequent when we use $s = 1$ as a threshold. For the 2-itemsets in the sample we have that $|\text{spt}\{1,2\}| = 1$, $|\text{spt}\{1,3\}| = 1$, $|\text{spt}\{1,4\}| = 0$, $|\text{spt}\{1,5\}| = 0$, $|\text{spt}\{1,6\}| = 0$, $|\text{spt}\{2,3\}| = 2$, $|\text{spt}\{2,4\}| = 1$, $|\text{spt}\{2,5\}| = 0$, $|\text{spt}\{2,6\}| = 0$, $|\text{spt}\{3,4\}| = 2$, $|\text{spt}\{3,5\}| = 1$, $|\text{spt}\{3,6\}| = 0$, $|\text{spt}\{4,5\}| = 2$, $|\text{spt}\{4,6\}| = 1$ and $|\text{spt}\{5,6\}| = 1$.

So the frequent itemsets in the sampled dataset are:

$$\begin{aligned}\text{SF} = \big\{&1,2,3,4,5,6,\\ &12,13,23,24,34,35,45,46,56,\\ &123,234,345,456\big\}.\end{aligned}$$

Since all the singletons are frequent in the sample, every 2-itemset not in SF must be in the negative border NB. So $\{14,15,16,25,26,36\} \subseteq \text{NB}$.

We can show that no 3-itemset can belong to the negative border: Suppose that abc is a triple in the NB. If 1 is contained in it, then only the triple 123 is available, which is frequent. If 6 is contained in it, then only the triple 456 is available which is also frequent. So neither 1 nor 6 may be contained in abc, leaving only the digits 2, 3, 4 and 5. The triples which can be formed by them are 234 and 345 which are frequent and 235 and 245 which contain the infrequent 25. Therefore, any 3-itemset is either frequent, or has an infrequent 2-itemset as a strict subset, implying that no 3-itemset may belong to the negative border:

$$\text{NB} = \big\{14,15,16,25,26,36\big\}.$$

We then check the support of each element of the negative border in the whole dataset. We obtain that $|\text{spt}\{14\}| = 2$, $|\text{spt}\{15\}| = 1$, $|\text{spt}\{16\}| = 0$, $|\text{spt}\{25\}| = 2$, $|\text{spt}\{26\}| = 1$ and $|\text{spt}\{36\}| = 2$, so none of them is frequent in the whole dataset. This means that the set of frequent itemsets in the whole dataset contains exactly the elements of SF which are also frequent in the whole data set.

Among the elements of SF, the singletons are clearly frequent. Also $\{24\}, \{34\}$ and $\{35\}$ are frequent, so the set of frequent itemsets is

$$F = \{1, 2, 3, 4, 5, 6, 24, 34, 35\}.$$

Exercise 3

a) We pick $x = m$ and $y = 2m$. Then for any a, b, $h_{a,b}(x) = am + b \mod m = b \mod m$ and $h_{a,b}(y) = 2am + b \mod m = b \mod m = h_{a,b}(x)$. We showed that the probability that a function $h_{a,b}$ will have the property that $h_{a,b}(x) = h_{a,b}(y)$ is equal to one, therefore the family cannot be universal.

b) We mention a few results which will simplify our proof:

Lemma 1. *Let $m \in \mathbb{N}$ and $p > m$ be a prime number. If $a \mod p \equiv_m b \mod p$, then $a \equiv_p b + im$ for some $i = 0, \pm 1, \ldots, \pm\lfloor \frac{p}{m} \rfloor$.*

Proof. Suppose that

$$\begin{aligned} a &= k_a p + r_a, \\ r_a &= \tilde{k}_a m + \tilde{r}_a, \\ b &= k_b p + r_b, \\ r_b &= \tilde{k}_b m + \tilde{r}_b. \end{aligned}$$

By our assumption, $\tilde{r}_a = \tilde{r}_b = \tilde{r}$, so

$$\begin{aligned} a &= k_a p + \tilde{k}_a m + \tilde{r}, \\ b &= k_b p + \tilde{k}_b m + \tilde{r}, \end{aligned}$$

which yields that $a - b = (k_a - k_b)p + (\tilde{k}_a - \tilde{k}_b)m$, namely $a \equiv_p b + im$ for $i = \tilde{k}_a - \tilde{k}_b$. Since $\tilde{k}_a, \tilde{k}_b \in \{0, 1, \ldots, \lfloor \frac{p}{m} \rfloor\}$, their difference $i = \tilde{k}_a - \tilde{k}_b$ will belong to $\{-\lfloor \frac{p}{m} \rfloor, \ldots, -1, 0, 1, \ldots, \lfloor \frac{p}{m} \rfloor\}$. □

Lemma 2. *For any prime number p and $m \in \mathbb{N}$ such that $1 < m < p$, we have that $\lfloor \frac{p}{m} \rfloor = \lfloor \frac{p-1}{m} \rfloor$.*

Proof. It suffices to show that for any integer $n \in \mathbb{N}$ with $n < \frac{p}{m}$, we also have that $n < \frac{p-1}{m}$. Suppose not. Then there exists some n such that

$$\frac{p-1}{m} \le n < \frac{p}{m}. \tag{1}$$

The last relation implies that $0 < p - mn \le 1$, so it has to be that $p = mn + 1$. But then,

$$\begin{aligned} \left\lfloor \frac{p}{m} \right\rfloor &= \left\lfloor n + \frac{1}{m} \right\rfloor = n \quad \text{and} \\ \left\lfloor \frac{p-1}{m} \right\rfloor &= \left\lfloor \frac{nm}{m} \right\rfloor = \lfloor n \rfloor = n, \end{aligned}$$

which contradicts (1). □

The following theorem [Fra03, Thm 4.11] asserts the existence and uniqueness of the solution of a linear system.

Theorem 3. *Let $m \in \mathbb{N}$ and $a \in [m]$ with $(a, m) = 1$. For every $b \in [m]$ the equation $ax \equiv b \pmod m$ has a unique solution modulo m.*

Fermat's Little Theorem [Fra03, Thm 4.8] specifies the inverse of an element when working modulo a prime number p.

Theorem 4 (Fermat)**.** *Let $a \in \mathbb{Z}$ and p be a prime such that p does not divide a. Then $a^{p-1} \equiv 1 \pmod p$.*

In particular, under the previous assumptions we also have that $aa^{p-2} \equiv 1 \pmod p$, so $a^{-1} = a^{p-2}$ when working modulo p.

We now return to the exercise. Let $x \neq y \in U$ and pick a function $h_{a,b}$ with $h_{a,b}(x) = h_{a,b}(y)$. Then $(ax + b) \mod p \equiv_m (ay + b) \mod p$. By Lemma 1, there exists some $i = 0, \pm 1, \ldots, \pm\lfloor \frac{p}{m} \rfloor$ such that $ax + b \equiv_p ay + b + im$, namely

$$a(x - y) \equiv_p im. \tag{2}$$

Notice that the case where $i = 0$ can be excluded, as otherwise we would obtain that p divides $a(x-y)$ which clearly cannot hold.[1] Also the negative values of i can be omitted as they correspond to the positive ones modulo p.

Since $(x - y, p) = 1$, the inverse of $x - y$ exists modulo p (Theorem 3) and by Fermat's Little Theorem it is equal to $(x - y)^{-1} = (x - y)^{p-2}$. So equation (2) implies that

$$a \equiv_p im(x - y)^{p-2} \tag{3}$$

for some $i \in \{1, \ldots, \lfloor \frac{p}{m} \rfloor\}$. From the last relation it follows that there are at most $\lfloor \frac{p}{m} \rfloor$ choices for a. On the other hand, for b we have a total of p choices. Therefore, the set $\{h_{a,b} : h_{a,b}(x) = h_{a,b}(y)\}$ has a cardinality of at most

[1] Then p would have to divide either a or $x - y$, but both of these numbers belong to $\{1, \ldots, p - 1\}$.

$\lfloor \frac{p}{m} \rfloor \cdot p$. In total, we have $p(p-1)$ distinct functions $h_{a,b}$, so the probability of obtaining a collision is equal to:

$$\begin{aligned}\frac{|\{h_{a,b} : h_{a,b}(x) = h_{a,b}(y)\}|}{p(p-1)} &\leq \frac{\lfloor \frac{p}{m} \rfloor \cdot p}{p(p-1)} = \frac{\lfloor \frac{p}{m} \rfloor}{p-1} \\ &= \frac{\left\lfloor \frac{p-1}{m} \right\rfloor}{p-1} \\ &\leq \frac{\frac{p-1}{m}}{p-1} \\ &= \frac{1}{m},\end{aligned}$$

by Lemma 2.

c) The family can be shown to be universal with the same argument as in b). The only difference is that for $x \neq y \in [p]$, the element $x-y$ now belongs to $\{-(p-1), \ldots, -1, 1, \ldots, p-1\}$. However, each of the elements of this set is still prime with respect to p, so $(x-y)^{-1}$ exists modulo p. This implies that the argument between equations (2) and (3) still holds. The rest of the proof remains unchanged as it does not take into account the range of $x-y$.

Exercise 4

a) Suppose that the elements $i = 1, 2, \ldots, n$ have been successively inserted into the list. When searching for an element $i \in \{1, \ldots, n\}$, we follow the exact same steps as when we inserted the same element. Let S_i denote the number of steps required for either of these two procedures. Then the mean number of steps required to insert all the n elements into the list is equal to $\frac{S_1 + \ldots + S_n}{n}$.

Similarly, after all the n elements have been inserted and we pick one at random using the discrete uniform distribution on $\{1, \ldots, n\}$, the mean number of steps required to find it in the list is equal to $\frac{1}{n}(S_1 + \ldots + S_n)$.

b) As before, we denote by S_j the number of steps required in order to successfully insert the j-th element into the list. By the uniformity assumption, the probability that S_j is at least equal to k is

$$P\left[S_j \geq k\right] = \frac{j-1}{m} \cdot \frac{j-2}{m-1} \cdot \ldots \cdot \frac{j-k}{m-k+1} \leq a_{j-1}^k, \quad (4)$$

where $a_{j-1} = \frac{j-1}{m}$ denotes the weight factor of the list after the first $j-1$ elements have been inserted in it. Recall that for a discrete random variable X taking values on $\mathbb{N}$, its expectation $\mathbb{E}[X]$ is given by the formula $\mathbb{E}[X] = \sum_{n=1}^{\infty} P[X \geq n]$ [Gut13, Thm 12.1]. Using it, we can bound the mean value of S_j by

$$\begin{aligned}\mathbb{E}[S_j] &= \sum_{k=1}^{\infty} P\left[S_j \geq k\right] \\ &\leq \sum_{k=1}^{\infty} a_{j-1}^k \\ &\leq \frac{1}{1-a_{j-1}} \\ &= \frac{m}{m-(j-1)}. \end{aligned} \quad (5)$$

The mean number of steps M when considering all the n elements is then

$$\begin{aligned}M &= \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}[S_j] \\ &\leq \frac{m}{n} \sum_{j=1}^{n} \frac{1}{m-(j-1)} \\ &= \frac{m}{n} \left(\frac{1}{m-n+1} + \ldots + \frac{1}{m} \right) \\ &\leq \frac{m}{n} \int_{m-n}^{m} \frac{1}{x} dx \quad (6) \\ &= \frac{m}{n} \ln \frac{m}{m-n} \\ &= \frac{1}{a} \ln \frac{1}{1-a}, \quad (7)\end{aligned}$$

where in relation (6) we used the elementary inequality $\frac{1}{k} + \ldots + \frac{1}{m} \leq \int_{k-1}^{m} \frac{1}{x} dx$ which holds for any two integers $1 < k < m$ by a simple inspection of the graph of $f(x) = \frac{1}{x}$ and the definition of the Riemann integral.

References

[Fra03] J. Fraleigh, *Μια εισαγωγή στην άλγεβρα*, Πανεπιστημιακές Εκδόσεις Κρήτης, 4η Έκδοση, 2003. ISBN: 9789607309716 Cited on p. 2

[Gut13] A. Gut, *Probability: A Graduate Course*, Springer Texts in Statistics 75, Springer-Verlag New York, 2013. DOI: 10.1007/978-1-4614-4708-5 Cited on p. 3

[LRU20] J. Leskovec, A. Rajaraman, J. Ullman, *Mining of Massive Datasets*, Cambridge University Press, 3rd Edition 2020. ISBN: 9781108476348 Cited on p. 1