Algorithmes On-lines

Sélectionner k éléments d'un ensemble S de n éléments distincts.

L'objectif de cet exercice est de concevoir un algorithme $\mathcal{A}$ qui sélectionne k éléments parmi un ensemble $S = \{s_1, s_2, \cdots, s_n\}$ de n éléments distincts de façon uniforme. Nous noterons par $\mathcal{P}_k(S)$ l'ensemble des parties de S ayant exactement k éléments. Nous supposerons connu que $C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$ correspond aux nombres de parties de k éléments d'un ensemble de n éléments distincts. En d'autres termes $|\mathcal{P}_k(S)| = \binom{n}{k}$.

Soit $\mathcal{A}$ un algorithme qui retourne une partie de k éléments parmi S de façon uniforme. Nous noterons X pour la variable aléatoire qui désigne la partie de S retournée par $\mathcal{A}$.

Question 1. Soit C une partie de S de k éléments. Que vaut la probabilité $Pr(X = C)$?

$Pr(X \in \mathcal{P}_k(S)) = 1$. Comme les évènements $X = C'$ et $X = C$ sont disjoints si $C' = C$, on a $\sum_{C\in\mathcal{P}_k(S)} Pr(X = C) = 1$. De plus, comme X est générée de de façon uniforme, on a $Pr(X = C) = Pr(X = C')$ pour toutes parties C' et C de S ayant k éléments. Donc

$$\begin{aligned}\sum_{C\in\mathcal{P}_k(S)} Pr(X = C) &= 1\\ \binom{n}{k} Pr(X = C) &= 1\\ Pr(X = C) &= \frac{k!(n-k)!}{n!}\end{aligned}$$

Question 2. Soit s un élément de S. Notons par $s \in X$ l'événement correspondant au fait que s est un élément de X, partie retournée par l'algorithme $\mathcal{A}$. Calculer les probabilités suivantes : $Pr(s \in X|X = C)$, $Pr(s \in X)$.

$$Pr(s \in X|X = C) = \begin{cases} 1 & si \quad s \in C \\ 0 & sinon \end{cases}$$

$$\begin{aligned}Pr(s \in X) &= \sum_{C\in\mathcal{P}_k(S)} Pr(s \in X|X = C) \cdot Pr(X = C)\\ &= \sum_{C\in\mathcal{P}_k(S) \text{ et } s\in C} Pr(s \in X|X = C) \cdot Pr(X = C)\\ &= \frac{\binom{n-1}{k-1}}{\binom{n}{k}} = \frac{\frac{(n-1)!}{(k-1)!(n-1-k+1)!}}{\frac{n!}{k!(n-k)!}}\\ &= \frac{k}{n}\end{aligned}$$

Nous utiliserons une fonction *Bernoulli*(p) qui retourne 1 avec probabilité p sinon retourne la valeur 0.

Considérons l'algorithme $\mathcal{D}$:

Entrée : Un ensemble $S = \{s_1, s_2, \cdots, s_n\}$ de n éléments distincts.

Sortie : une partie X ayant k éléments de S.

1. $X \leftarrow \emptyset$;
2. pour i allant 1 à n faire
   (a) Si $i \leq k$, alors $X \leftarrow X \cup \{s_i\}$;
   (b) sinon
      i. $Z_i \leftarrow Bernoulli(\frac{k}{i})$;
      ii. Si $Z_i = 1$ alors
         – Choisir un élément ℓ de X de façon uniforme ;
         – $X \leftarrow X \backslash \{\ell\}$;
         – $X \leftarrow X \cup \{s_i\}$;
3. retourner X

Notons par $\mathcal{E}(e, i)$ l'évènement que l'élément e soit dans l'ensemble désigné par la variable X à la fin de l'itération i. L'objectif est de calculer la probabilité que l'élément e soit dans X à la fin de l'itération i, en d'autres termes, de calculer la probabilité $\Pr(\mathcal{E}(e, i))$.

Question 3. Quelle est la probabilité ($\Pr(\mathcal{E}(e, k))$) que l'élément e soit dans X à la fin de l'itération k (c'est-à-dire lorsque i vaut k dans l'algorithme) ?

$$\Pr(\mathcal{E}(e, k)) = \begin{cases} 1 & \textit{si} \quad e \text{ est un des éléments } s_1, s_2, s_3, \cdots, s_k \\ 0 & \textit{sinon} \end{cases}.$$

Question 4. Soit s_i le $i$ème élément de S avec $i > k$.

1. Donner les probabilités suivantes : $\Pr(Z_i = 1)$ et $\Pr(Z_i = 0)$.
2. Quelle est la probabilité que s_i soit dans X à la fin de l'itération i ? (Calculer $\Pr(\mathcal{E}(s_i, i))$)

   Pour qu'il soit dans X, il faut que $Z_i = 1$, la probabilité qu'il soit dans X après l'étape i correspond à la probabilité que $X_i = 1$.

$$\Pr(\mathcal{E}(s_i, i)) = \frac{k}{i}.$$

Question 5. Soit e un élément dans X à la fin de l'itération k. Quelle est la probabilité ($\Pr(\mathcal{E}(e, k+1) | \mathcal{E}(e, k))$) qu'il reste dans X à la fin de l'itération $k+1$? Pour cela, calculer cette probabilité dans les deux cas suivants :

1. sachant que $(Z_{k+1} = 0)$
2. sachant que $(Z_{k+1} = 1)$

Si l'évènement $(Z_{k+1} = 0)$ se réalise, alors aucun élément sera supprimé pendant l'itération $k+1$ et l'élément e sera toujours dans X. Comme $\Pr(Z_{k+1} = 0) = \frac{1}{k+1}$, on a

$$\Pr(\mathcal{E}(e, k+1)|\mathcal{E}(e,k) \text{ et } Z_{k+1} = 0) = \frac{1}{k+1}$$

Si l'évènement $(Z_{k+1} = 1)$ se réalise, un seul élément sera supprimé : il est choisit de façon uniforme. Donc, un élément de X ne sera pas supprimé avec probabilité $\frac{k-1}{k}$.

$$\Pr(\mathcal{E}(e, k+1)|\mathcal{E}(e,k) \text{ et } Z_{k+1} = 1) = \frac{k-1}{k}$$

Donc :

$$\begin{aligned}\Pr(\mathcal{E}(e,k+1)|\mathcal{E}(e,k)) &= \Pr(Z_{k+1}=0)\Pr(\mathcal{E}(e,k+1)|\mathcal{E}(e,k) \text{ et } Z_{k+1}=0)) + \Pr(Z_i=1)\Pr(\mathcal{E}(e,k+1)|\mathcal{E}(e,k) \text{ et } Z_{k+1}=1) \\ &= 1 \cdot \frac{1}{k+1} + \frac{k-1}{k} \cdot \frac{k}{k+1} \\ &= \frac{k}{k+1}\end{aligned}$$

La première ligne du calcul correspond à la définition de la probabilité conditionnelle. La deuxième est du au fait que $\Pr(Z_{k+1} = 1) = \frac{k}{k+1}$.

Question 6. Montrer que la probabilité que l'élément e soit dans X à la fin de l'itération $k+1$ vaut $\frac{k}{k+1}$?

Soit e un élément dans $\{s_1, s_2, s_3, \cdots, s_k\}$.

$$\begin{aligned}\Pr(\mathcal{E}(e,k+1)) &= \Pr(\mathcal{E}(e,k))\Pr(\mathcal{E}(e,k+1)|\mathcal{E}(e,k)) + \Pr(non\mathcal{E}(e,k))\Pr(\mathcal{E}(e,k+1)|non\mathcal{E}(e,k)) \\ &= 1\Pr(\mathcal{E}(e,k+1)|\mathcal{E}(e,k)) + \Pr(\mathcal{E}(e,k+1)|non\mathcal{E}(e,k)) \\ &= \frac{k}{k+1}\end{aligned}$$

En utilisant les deux questions précédentes,

$$\Pr(\mathcal{E}(e,k+1)) = \begin{cases} \frac{k}{k+1} & si \quad e \text{ est un des éléments } s_1, s_2, s_3, \cdots, s_k, s_{k+1} \\ 0 & \end{cases}$$

Nous allons prouver ***par récurrence*** grâce aux prochaines questions que la probabilité qu'un élément e dans $\{s_1, \cdots, s_i\}$ soit dans X à la fin de l'itération i vaut $\frac{k}{i}$: autrement dit, $\Pr(\mathcal{E}(e,i)) = \frac{k}{i}$. La question 6 montre que l'hypothèse de récurrence est vrai pour $i = k+1$. Supposons que l'hypothèse de récurrence est vraie pour pour i. Par la suite nous allons prouver qu'elle est vraie pour $i+1$.

Question 7. Soit e un élément dans X à la fin de l'itération i. Quelle est la probabilité qu'il reste dans X à la fin de l'itération $i+1$? Pour cela, calculer cette probabilité dans les deux cas suivants :

1. sachant que $(Z_{i+1} = 0)$

2. sachant que $(Z_{i+1} = 1)$

   Si l'évènement $(Z_{i+1} = 0)$ se réalise, alors aucun élément sera supprimé pendant l'itération i et l'élément e sera toujours dans X. Comme $\Pr(Z_{i+1} = 0) = \frac{i+1-k}{i+1}$, on a

   $$\Pr(\mathcal{E}(e, i+1) | \mathcal{E}(e, i) \text{ et } Z_{i+1} = 0) = \frac{i+1-k}{i+1}$$

   Si l'évènement $(Z_{i+1} = 1)$ se réalise, un seul élément sera supprimé : il est choisit de façon uniforme. Donc, un élément e de X ne sera pas supprimé avec probabilité $\frac{i}{i+1}$.

   $$\Pr(\mathcal{E}(e, i+1) | \mathcal{E}(e, i) \text{ et } Z_{i+1} = 1) = \frac{i}{i+1}$$

   Donc :

   $$\begin{aligned}\Pr(\mathcal{E}(e, i+1) | \mathcal{E}(e, i)) &= \Pr(Z_{i+1} = 0) + \Pr(Z_{i+1} = 1) \Pr(\mathcal{E}(e, i+1) | \mathcal{E}(e, i+1) \cap Z_{i+1} = 1) \\ &= \frac{i+1-k}{i+1} + \frac{k}{i+1} \frac{k-1}{k} \\ &= \frac{i}{i+1}\end{aligned}$$

Question 8. Quelle est la probabilité que l'élément e soit dans X à la fin de l'itération $i+1$?

- Si $e \notin \{s_1, \cdots, s_{i+1}\}$, alors $\Pr(\mathcal{E}(e, i)) = 0$
- Si $e \in \{s_1, \cdots, s_i\}$,

  $$\begin{aligned}\Pr(\mathcal{E}(e, i+1)) &= \Pr(\mathcal{E}(e, i+1) | \mathcal{E}(e, i)) \cdot \Pr(\mathcal{E}(e, i)) \\ &= \frac{i}{i+1} \cdot \frac{k}{i} \\ &= \frac{k}{i+1}\end{aligned}$$

- Si $e = s_{i+1}$, $Pr(\mathcal{E}(s_{i+1}, i+1)) = \frac{k}{i+1}$

Question 9. Quelle est la probabilité que l'élément e soit dans X à la fin de l'itération n ?

Il faut réutiliser la question précédente en focalisant sur $i+1 = n$.

- Si $e \in \{s_1, \cdots, s_{n-1}\}$, alors $\Pr(\mathcal{E}(e, n)) = \frac{k}{n}$
- Si $e = s_n$, $Pr(\mathcal{E}(s_n, n)) = \frac{k}{n}$

Question 10. Peut-on adapter l'algorithme $\mathcal{D}$ en un algorithme on-line ? Justifier votre réponse (si c'est oui, donner l'algorithme, si c'est non, donnez vos raisons) ?

oui : car l'algorithme n'utilise pas le nombre d'éléments de S.