Combinatoire, Probabilités, Statistiques

Adrian TANASĂ

université de Bordeaux – Licence Informatique (2ème année, S3)

2022-2023

Administratif

- **Équipe pédagogique :**
  - CM/responsable UE : Adrian TANASĂ
    LaBRI (bât. A30, bureau 303)
    ntanasa@u-bordeaux.fr
  - Chargés de TD : P. Bonnet, Ph. Duchon, A. Larbi, C. Legrand, F. Mazoit, Z. Varin
- 12 séances cours (mardi matin (8h - 9h20)), TD, TD Machine
- **Page Moodle du cours :**
  https ://moodle1.u-bordeaux.fr/course/view.php ?id=10789
  transparents cours, énoncés TD, énoncés TD Machine, annales DS, annales examen terminal *etc.*
- **Évaluation :**
  - un devoir surveillé, coefficient 0.5 (25 nov.)
  - examen terminal, coefficient 0.5 (janvier)
  - si la note de l'examen est meilleure que celle du DS, on ne garde que l'examen
  - même chose en session 2

Équipe pédagogique

- **Cours :** Ph. Duchon
- **TD et TDM :**
  - **A1** : F. Mazoit
  - **A2** : A. Larbi
  - **A3** : Ph. Duchon
  - **A4** : P. Bonnet
  - **A5** : C. Legrand
  - **A6** : Z. Varin

Contenu de l'UE

Probabilités et **statistiques** : disciplines étudiant les situations d'*incertitude* ("le hasard", c'est quoi ?)

- **Combinatoire :** étude des objets discrets (combinaisons, mots, chemins, arbres. . .)
  - on *compte* les possibilités ;
    variante : on les *énumère* (passe en revue une par une)
  - parfois un préalable indispensable à certains calculs de probabilités
  - utilisé en particulier pour
    analyser la complexité des algorithmes - analyse d'algorithmes
- **Probabilités :** on définit un modèle pour une expérience, et on *prédit* par le calcul ce qu'il est plausible d'observer
- **Statistiques :** à partir de données concrètes, on essaie de *diagnostiquer*
  - décrire de manière concise les données observées
  - proposer des valeurs plausibles pour les paramètres d'un modèle
  - "est-ce que les données sont raisonnablement compatibles avec l'hypothèse que. . ."

Notions-clef du cours

- ensembles : union, intersection, cardinal d'un ensemble *etc.*
- combinatoire : classe combinatoire, suite de comptage, permutations, coefficients binomiaux, arbres binaires, mots de Dyck *etc.*
- probabilités discrètes : variables aléatoires, indépendance, espérance, variance *etc.*

Plan du cours

1. Quelques notions de la théorie des ensembles : union, intersection, cardinal d'un ensemble *etc.*
2. Combinatoire : classe combinatoire, suite de comptage, codage des objets par des mots, comptage (chemins, compositions d'un entier, pavages), arbres binaires complets, formule de récurrence pour le comptage (nombres de Catalan), codage bijectif des arbres par les mots (ou chemins) de Dyck, obtention d'une formule pour le comptage des mots de Dyck (par le principe de réflexion), autres preuves du nombre de mots de Dyck par la bijection de Rémy et pas le lemme cyclique
3. Probabilités discrètes : variables aléatoires, loi d'une variable aléatoire, indépendance, espérance, variance, lois de probabilités usuelles (Bernoulli, binomiale, géométrique, Poisson) et certaines de leurs propriétés
4. Probabilités continues (si le temps le permettra)
5. Statistiques (si le temps le permettra)

Pourquoi un cours de probas-stats en informatique ?

Probabilités et statistiques sont présentes dans de nombreux sous-domaines de l'informatique, aussi bien au niveau de la théorie que des applications... difficile de les éviter !

- **Conception d'algorithmes :** énormément d'algorithmes "probabilistes" existent qui ont des performances bien meilleures que ce qu'atteignent des algorithmes "déterministes" d'un degré comparable de simplicité.
- **Analyse d'algorithmes :** les techniques probabilistes permettent des analyses précises de la complexité d'algorithmes
- **Analyse statistique des données :** classification, apprentissage... font beaucoup appel aux probabilités et aux statistiques ; idem dans beaucoup de branches de l'intelligence artificielle (IA)

Pourquoi des probas - algo PageRank de Google

L'algorithme **PageRank (PR)** (algo de Google) : l'algorithme d'analyse des liens concourant au système de classement des pages Web

PR mesure quantitativement la popularité d'une page web.

Le PageRank n'est qu'un indicateur parmi d'autres dans l'algorithme qui classe les pages du Web dans les résultats de recherche de Google

le concept mathématique qui a rendu possible le calcul du PageRank : le thm. de point fixe (thm. en anlyse mathématique)

PR est basé sur l'application de la théorie des chaînes de Markov (*i.e.* des probabilités) !

Pourquoi des probas - algo PageRank de Google (suite)

principe de base : attribuer à chaque page une valeur (ou score) proportionnelle au nombre de fois que passerait par cette page un utilisateur parcourant le graphe du Web en cliquant aléatoirement, sur un des liens apparaissant sur chaque page.

formellement, le déplacement de l'utilisateur est une marche aléatoire sur le graphe du Web (le graphe orienté dont les sommets représentent les pages du Web et les arcs les hyperliens)

En supposant que l'utilisateur choisisse chaque lien indépendamment des pages précédemment visitées il s'agit d'un processus de Markov

Le PageRank est alors simplement la probabilité stationnaire d'une chaîne de Markov

Pourquoi des probas (suite)

- De manière générale, on peut estimer que les compétences suivantes font parties du bagage naturel d'un informaticien :
  - **programmer** (simuler) un modèle probabiliste simple à des fins d'évaluation et de test
  - **modéliser** une situation d'incertitude en décrivant un modèle probabiliste
  - **analyser** (calculer, prédire) les caractéristiques prévisibles d'un modèle probabiliste.
- **En TD :** des exercices de probas classiques
- **En TDM :** on *programme* des simulations ; aide pour développer l'intuition probabiliste.
  langage : PYTHON.

Partie I : Quelques notions de la théorie des ensembles

Vocabulaire et notations : ensembles

- **Description d'ensemble :**
  1. soit en listant ses éléments (*exemple* : $E = \{1, 2, 3, 4, 5, 6\}$),
  2. soit par une description des éléments
     (*exemple* : E est l'ensemble des entiers compris entre 1 et 6)

  l'ordre "ne compte pas" !
- *Notation :* $x \in E$ ("x appartient à E" ; "E contient x") pour dire que x est un élément de l'ensemble E
  exemple : $1 \in E$.
- *Notation :* $A \subset B$ ("A est inclus dans B", "B inclut A", ou "A est un sous-ensemble de B") pour dire que tout élément de A est aussi un élément de B
  exemple : $\{1, \ldots, 5\} \subset E$.

Vocabulaire et notations : ensembles (2)

- **Égalité de deux ensembles :** deux ensembles sont égaux s'ils ont exactement les mêmes éléments
- **Description d'un sous-ensemble :** "ensemble des éléments de A qui satisfont la propriété P", $\{x \in A : P(x)\}$
- **Ensembles particuliers :** $\emptyset$, $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$...
- *Convention :* $[[a, b]]$ désigne l'ensemble de tous les **entiers** compris entre a et b (inclus).
  ex. : $[[1, 3]] = \{1, 2, 3\}$

Vocabulaire et notations : opérations ensemblistes

- **Opérations ensemblistes :** union, intersection, produit cartésien : $A \cup B$, $A \cap B$, $A \times B$; A^n.
  exemples : $A = \{1, 2, 3\}, B = \{3, 4\}$
  $A \cup B = \{1, \dots, 4\}$,
  $A \cap B = \{3\}$,
  $A \times B = \{(1,3), (1,4), (2,3), (2,4), (3,3), (3,4)\}$,
  $A^2 = A \times A =$
  $\{(1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3)\}$,
  $A^3 = A^2 \times A = \{(1,1,1), \dots\}$
- ***Différence :*** $A - B = \{x \in A : x \notin B\}$ *(attention, ça ne suppose pas que B soit inclus dans A) ; peut aussi être noté* $A \backslash B$,
  exemples : $A - B = \{1, 2\}$.

Vocabulaire et notations : opérations ensemblistes (2)

- $\mathcal{P}(A) = \{B : B \subset A\}$: "powerset" de A, l'ensemble de tous les sous-ensembles de A (c'est bien un *ensemble d'ensembles* : un ensemble dont les éléments sont eux-mêmes des ensembles).
- **Exemple :** $\mathcal{P}(\{1,2,3\}) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$
- **Notations d'unions ou d'intersections itérées :** si pour chaque $i \in I$, A_i est un ensemble,
  - $\bigcup_{i \in I} A_i$ est l'ensemble de tous les éléments qui sont dans **au moins un** des ensembles A_i ;
  - $\bigcap_{i \in I} A_i$ est l'ensemble de tous les éléments qui sont dans **tous** les ensembles A_i.

Ensembles finis, cardinal

- Notion d'ensemble **fini** ou **infini** ; pour un ensemble fini A, son **cardinal** est son nombre d'éléments, noté $\#A$.
- **Attention** au vocabulaire : un ensemble fini a un nombre fini d'éléments ; l'intervalle $[0, 1]$ (dans $\mathbb{R}$) est borné, il n'est pas fini.
- Deux ensembles sont **disjoints** si leur intersection est l'ensemble vide ; quand on parle de plus de deux ensembles, il faut distinguer deux notions :
  - des ensembles $(A_i)_{i \in I}$ sont **deux à deux disjoints** si, quelques soient les deux ensembles A_i et A_j avec $i \neq j$, ces deux ensembles sont disjoints (aucun élément n'appartient à plus d'un des ensembles) ;
  - des ensembles $(A_i)_{i \in I}$ sont **globalement disjoints** si leur intersection à tous est vide (aucun élément n'appartient à chacun des ensembles).

Épisode suivant

Vocabulaire sur les fonctions :

- images,
- préimages,
- cardinaux
- *etc.*

Combinatoire :

- classe combinatoire,
- suite de comptage
- *etc.*