Statistique et Informatique (3I005)

2016-2017

Nicolas Baskiotis - Pierre-Henri Wuillemin

Université Pierre et Marie Curie (UPMC)
Laboratoire d'Informatique de Paris 6 (LIP6)

Cours 1 :
Probabilités sur des ensembles discrets et dénombrements

Description de l’UE

Objectifs du cours

- Introduction aux domaines :
  - de la théorie des probabilités,
  - de la statistique,
- donner des exemples de leur application en informatique,
- pratiquer les concepts introduits sur des exemples $\longrightarrow$ mini-projets.

Organisation

- Calcul des probabilités (Nicolas Baskiotis – cours 1 à 6) :
  - introduction aux probabilités,
  - exemples d’applications.
- L’inférence statistique (Pierre-Henri Wuillemin – cours 7 à 11) :
  - recueil et analyse des données,
  - estimation, tests et validation.

Description de l'UE (2)

Informations pratiques

- Site Web :
  `http://webia.lip6.fr/~phw/3i005`
  (`http://www-connex.lip6.fr/~baskiotisn` pour le début)
- Organisation en *mini-projets*:
  - TD/TME 1-4: intro + projet 1,
  - TD/TME 5-7 : projet 2,
  - TD/TME 8-11 : projet 3 + révisions.

Evaluation

- Les trois mini-projets sont notés
- les mini-projets comptent dans la note finale *dans tous les cas* !
- un partiel et un examen.

Evaluation

Calcul de la note finale

- Note de Contrôle Continu : le partiel (40%),
- note d'Ecrit : sur 60%, Examen $(70\%$, soit $\sim 40\%$ du total) et note des 3 projets (30%)

Exemple

- un étudiant a eu 12, 13, et 14 aux mini-projets, 12 au partiel et 11 à l'examen. Alors:
- note de CC : $24/40$,
- note Ecrit : $11 * 0.7 + (13 + 14 + 15)/3 * 0.3 = 11.9/20$, donc $35.7/60$,
- note finale : $59.7/100$.

Plan

De quoi parle ce cours . . .

- Qu'est ce que la chance ? le hasard ? le aléas ?
- Comment mesurer le hasard ?
- Comment l'utiliser ?
- Comment modéliser des phénomènes aléatoires ?
- Comment les étudier ?
- Comment les caractériser ?
- Que peut-on prédire ?

Définitions

Probabilités

- La théorie des probabilités : domaine des mathématiques qui étudie les phénomènes *aléatoires*,
- fournit des outils pour étudier les *expériences aléatoires* :
  des expériences qui, répétées dans les mêmes conditions, ne donnent pas nécessairement le même résultat.

Statistique

- La statistique : domaine des mathématiques dans lequel on étudie la collecte, l'analyse, l'interprétation de données,
- en particulier, des données stockées dans les bases de données, sur le Web, ...

Une (très) petite histoire des probabilités et statistiques

- $-XVI^e$ siècle : préhistoire, (Cardan 1501-1576, Galilée 1564-1642)
- XVI^e-$XVII^e$: la découverte du domaine
  - Fermat (161x-1665), Pascal (1623-1662)
  - Huyghens (1629-1695)
- $XVIII^e - XIX^e$: developpement et premières applications scientifiques
  - Montmort (1678-1719), de Moivre (1667-1754)
  - la dynastie Bernouilli : Jacob (1657-1705), Jean (1667-1748), Daniel (1700-1782), Nicolas (1687-1759)
  - Bayes (1700-1761)
  - Buffon (1707-1788), Simpson (1710-1761), D'alembert (1717-1783)
  - Lagrance (1736-1813), Laplace (1749-1827), Poisson (1781-1840)
- $XIX^e - XX^e$: théorie de la mesure, axiomatisation, applications multiples
  - Tchebychev (1821-1894), Emile Borel (1871-1956), Johann Radon (1887-1956), Paul Lévy (1886-1971), Andreï Kolmogorov (1903-1987)
  - Gibbs (1839-1903), Boltzmann (1844-1906), Poincaré (1854-1912) , Pearson (1857-1936), Markov (1856-1922)

Applications : informatique fondamentale

- algorithmique : *tri rapide*
  $\longrightarrow$ meilleure performance “en moyenne” que les autres tris ;
  “en moyenne” $\approx$ les valeurs dans le tableau initial sont aléatoires.
- structure de données : *Table de Hashage*
  $\longrightarrow$ propriété souhaitée de `hashCode` : donner des valeurs différentes aux différentes chaînes de caractères stockées dans la table de hachage.
  $\longrightarrow$ nécessite un modèle (probabiliste) des chaînes de caractères qui seront stockées.
- compression de données.

Cryptographie et cryptanalyse

Enigma : machine de cryptage allemande pendant la Seconde Guerre mondiale.

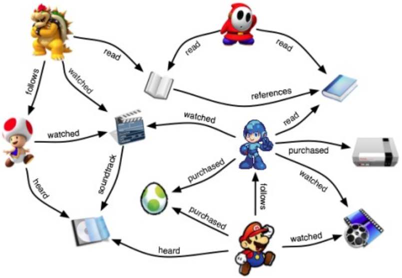
Le décryptage des messages par les alliés a été facilité par un mauvais algorithme de génération de *permutations* aléatoires.

La sécurité des communications sur Internet est gérée par des algorithmes de cryptographie.

Les algorithmes de cryptographie utilisent des générateurs de nombres aléatoires.

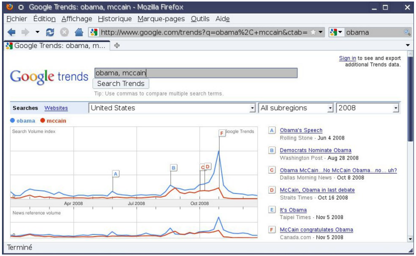
Réciproquement : les cryptanalystes cherchent les *régularités* (déviations par rapport à l'aléatoire) dans les textes cryptés.

Fouille de données



Systèmes de recommandation :
Les clients qui ont acheté ...
ont aussi acheté ...

Analyses statistiques des achats/recherches des différents produits

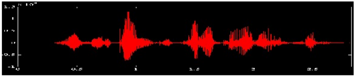


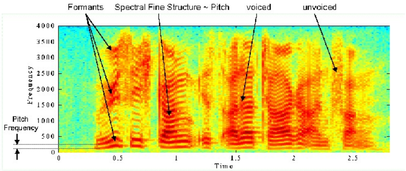
Google Trends : analyse des requêtes effectuées par les utilisateurs de Google.

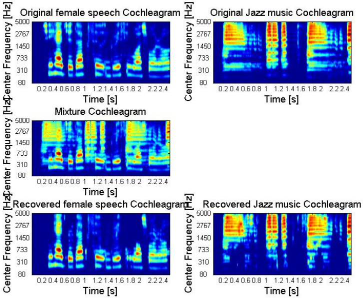
Applications possibles : suivi des intérêts dans une population, détection des épidémies, ...

Reconnaissance de la parole, séparation de sources

http://markus-hauenstein.de

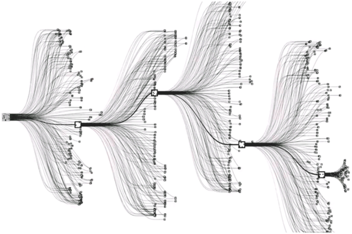






Analyse prédictive/apprentissage automatique

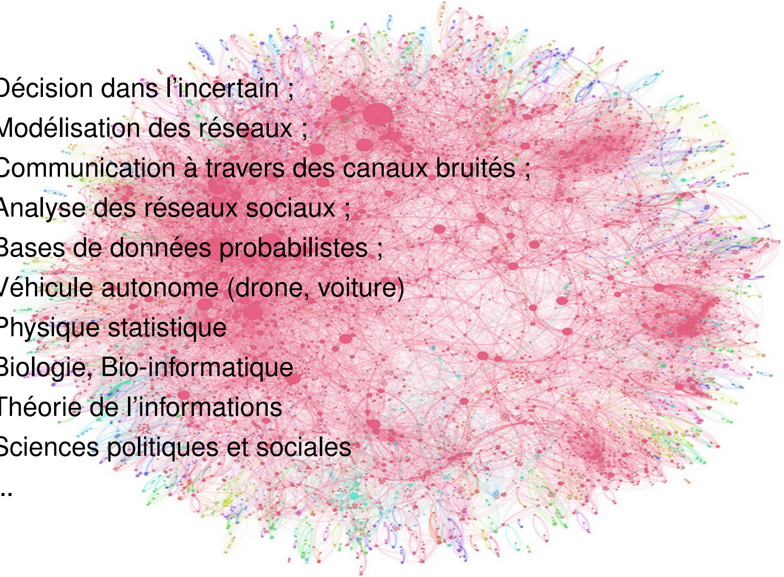
IA de jeu



Autres exemples :

- traduction automatique,
- génération automatique (musique, textes)
- classification d'images, moteur de recherche
- Interface cerveau-machine (BCI)

Et bien d'autres...



- Décision dans l'incertain ;
- Modélisation des réseaux ;
- Communication à travers des canaux bruités ;
- Analyse des réseaux sociaux ;
- Bases de données probabilistes ;
- Véhicule autonome (drone, voiture)
- Physique statistique
- Biologie, Bio-informatique
- Théorie de l'informations
- Sciences politiques et sociales
- ...

Plan

Une probabilité ?

Trois sachets, un croissant ...

- Pourquoi a-t-on une chance sur 3 de trouver le croissant ?
- Est-ce toujours le cas ?

Une probabilité ?

Trois sachets, un croissant . . .

- Pourquoi a-t-on une chance sur 3 de trouver le croissant ?
- Est-ce toujours le cas ?

La probabilité d'un événement

- c'est la fréquence d'apparition de l'événement, le nombre de fois où il apparaît rapportée au nombre d'expériences.
- Notion d'événement, d'expérience et de répétition.

Un peu plus compliqué

30 croissants, 30 pains au chocolat, 20 pains aux raisins, 10 pains au lait, 10 chaussons

Qu'est ce qui est équiprobable ?
Quelle est la probabilité :

- d'un pain au chocolat ?
- si 5 croissants ont été tirés avant ?
- qu'un pain soit tiré ?

Plusieurs tirages

Qu'elle est la probabilité :

- de tirer aucun croissants au bout de deux tirages ? au bout de trois ?
- de tirer au moins un croissant ?

Probabilités sur les ensembles discrets

Pour modéliser une expérience aléatoire :

Nous avons besoin des notions de :

- *événement élémentaire* : un résultat simple non composé de l'expérience
- *événement* : un résultat simple ou composé de plusieurs événements élémentaires de l'expérience
- *univers* : l'ensemble de tous les résultats possibles

Formalisation

- Soit Ω, un ensemble dénombrable, appelé univers,
  - Ω représente l'ensemble des résultats possibles d'une expérience aléatoire
- un élément $\omega \in \Omega$ est *un événement élémentaire*,
- un sous-ensemble E de Ω est un *événement*.

Probabilités sur les ensembles discrets

Exemple : lancer simultané de trois dés

- événement élémentaire : $(i,j,k) \in \{1,\dots 6\}^3$
- L'univers (l'ensemble des événements possibles) est $\Omega = \{(i,j,k) \mid i \in \{1,..,6\}, j \in \{1,..,6\}, k \in \{1,..,6\}\} = \{(1,1,1),(1,1,2),...,(1,2,1),(1,2,2)...,(6,6,6)\}$, $card(\Omega) = 6^3 = 216$
- Si on considère chaque dé équilibré, alors chaque événement élémentaire est équiprobable : $\forall (i,j,k) \in \Omega, P((i,j,k)) = \dfrac{1}{216}$.
- $E = \{(1,1,1),(2,2,2),...(6,6,6)\}$ représente l'événement : *les 3 dés sont égaux*.

Bien définir un problème

Problème du prince de Toscanne

Pourquoi en lançant trois dés, obtient-on plus souvent un total de 10 points qu'un total de 9 points, alors qu'il y a 6 façons d'obtenir ces deux totaux?

9 pts	10 pts
$6+2+1$	$6+3+1$
$5+2+2$	$6+2+2$
$5+3+1$	$5+4+1$
$4+3+2$	$5+3+2$
$4+4+1$	$4+4+2$
$3+3+3$	$4+3+3$

Premier réflexe : définir les événements et l'univers !

Bien définir un problème

Problème du prince de Toscanne

Pourquoi en lançant trois dés, obtient-on plus souvent un total de 10 points qu'un total de 9 points, alors qu'il y a 6 façons d'obtenir ces deux totaux?

9 pts	10 pts
$6+2+1$	$6+3+1$
$5+2+2$	$6+2+2$
$5+3+1$	$5+4+1$
$4+3+2$	$5+3+2$
$4+4+1$	$4+4+2$
$3+3+3$	$4+3+3$

Soit $\Omega_{i,j,k}$ l'évènement : les chiffres $\{$i,j,k$\}$ sont affichés sur les dés, sans tenir compte de l'ordre :

$$\text{pour } i \geq j \geq k, \Omega_{i,j,k} = \{(i,j,k),(j,i,k),(j,k,i),(k,j,i),(k,i,j),(i,k,j)\}$$

Il y a 6 événements $\Omega_{i,j,k}$ qui donnent une somme à 10, et 6 qui donnent une somme à 9.

Bien définir un problème

Problème du prince de Toscanne

Pourquoi en lançant trois dés, obtient-on plus souvent un total de 10 points qu'un total de 9 points, alors qu'il y a 6 façons d'obtenir ces deux totaux?

9 pts	10 pts
$6+2+1$	$6+3+1$
$5+2+2$	$6+2+2$
$5+3+1$	$5+4+1$
$4+3+2$	$5+3+2$
$4+4+1$	$4+4+2$
$3+3+3$	$4+3+3$

Les événements $\Omega_{i,j,k}$ ne sont pas "équiprobables" :

- si $i \neq j \neq k \neq i$, alors 6 manières d'obtenir le résultat, $P(\Omega_{i,j,k}) = \frac{6}{216}$,
- si $i = j \neq k$, alors 3 manières, $P(\Omega_{i,j,k}) = \frac{3}{216}$
- si $i = j = k$, alors une seule possibilité, $P(\Omega_{i,j,k}) = \frac{1}{216}$

$P(\{i+j+k=9\}) = \frac{6+3+6+6+3+1=25}{216}$ et $P(\{i+j+k=10\}) = \frac{6+3+6+6+3+3=27}{216}$

Probabilités sur des ensembles discrets (2)

Mesure de probabilité : caractérise l'aléa

- elle définie une probabilité pour chaque événement, qui correspond à la fréquence d'apparition de l'événement par définition, entre 0 et 1
- la probabilité de l'univers est de 1 : au moins un des événements de l'univers se réalise lors d'une expérience
- la probabilité qu'aucun événement arrive est donc de 0.

Elle est entièrement définie par les probas des événements élémentaires.

Formalisation

- Soit $\mathcal{P}(\Omega)$ l'ensemble des sous-ensembles de Ω. Une mesure de probabilité sur Ω est une fonction $P : \mathcal{P}(\Omega) \rightarrow [0, 1]$ vérifiant:
  - $P(\Omega) = 1$ (Ω est l'événement certain),
  - pour tout événement E, $P(E) \geq 0$,
  - Pour toute suite $(E_i)_{i \in \mathbb{N}}$ d'événements deux à deux disjoints (*incompatibles*, $E_i \cap E_j = \emptyset, i \neq j$) : $P(\bigcup_i E_i) = \sum_i P(E_i)$.
- Fonction de masse p associée à P : $\forall \omega \in \Omega, p(\omega) = P(\{\omega\})$
- Pour tout événement E: $P(E) = \sum_{\omega \in E} p(\omega)$

Probabilités sur des ensembles discrets (3)

Probabilité uniforme

- Considérons un ensemble *fini* Ω. La probabilité uniforme sur Ω est définie par la fonction de masse : $p(\omega) = \frac{1}{card(\Omega)}$.
- En particulier pour tout événement E, $P(E) = \dfrac{card(E)}{card(\Omega)}$.

Interprétation: si on répète (indéfiniment) l'expérience aléatoire

- le résultat de l'expérience sera ω avec une fréquence de $P(\{\omega\})$,
- un événement E se produit avec une fréquence $P(E)$
  $\rightarrow$ le résultat appartient à l'ensemble E avec une fréquence $P(E)$.

Exemple : lancer simultané de trois dés

Soit E l'événement : *La somme des trois chiffres est inférieure stricte à* 5, alors

$$P(E) = \frac{4}{256}.$$

En effet : $E = \{(1,1,1),(1,1,2),(1,2,1),(2,1,1)\}$.

Probabilités sur des ensembles discrets (4)

Propriétés

- $P(\varnothing) = 0$, ($\varnothing$ est l'événement impossible)
- $P(\bar{E}) = 1 - P(E)$ ($\bar{E}$: complémentaire de E dans Ω),
- $P(E \cup F) = P(E) + P(F) - P(E \cap F)$
- $E \subset F \Rightarrow P(F) = P(F \backslash E) + P(E) \Rightarrow P(E) \leq P(F)$
  ($F \backslash E$: ensemble des éléments de F qui ne sont pas dans E),
- $P(\bigcup_i E_i) \leq \sum_i P(E_i)$

Exemple : lancer simultané de trois dés

Quelle est la probabilité :

- d'obtenir aucun 6 ?
- Que la somme fasse 9 ou que $i = j = k$?

Plan

Le problème des partis
Problème du Chevalier de Méré (Pascal, Fermat)

Deux joueurs jouent à pile ou face

- 128 euros sont en jeu, 64 de la poche de chacun;
- le premier joueur marque 1 point si pile sort, sinon c'est le second;
- le premier qui arrive à 7 points gagne tout.
- La partie s'arrête sur un score donné (x, y) subitement;
- comment répartir équitablement l'argent ?

En particulier :

- si le score est de $(6, 6)$?
- si le score est de $(6, 5)$?
- si le score est $(1, 0)$?
- si le score est $(6, 0)$?

Le problème des partis
Problème du Chevalier de Méré (Pascal, Fermat)

Deux joueurs jouent à pile ou face

- 128 euros sont en jeu, 64 de la poche de chacun;
- le premier joueur marque 1 point si pile sort, sinon c'est le second;
- le premier qui arrive à 7 points gagne tout.
- La partie s'arrête sur un score donné (x, y) subitement;
- comment répartir équitablement l'argent ?

Solution de Fermat

- On note p pour pile, f pour face;
- soit p le nombre de points manquant au premier joueur, q au second pour gagner;
- une succession de partie : une séquence $pffppf..$
- dénombrer les parties favorables à chaque joueur.

Dénombrements

Dénombrement de n-uplets

Soit E un ensemble fini de taille n, et k un entier.

- nombre de k-uplets d'éléments de E: n^k (tirage avec remise, ordonné)
- nombre de k-uplets d'éléments distincts (tirage sans remise, ordonné, nombre d'arrangements de k parmi n) : $A_n^k = n \times (n-1) \times ... \times (n-k+1)$
- Nombre de permutations (cas $n = k$): $n! = n \times (n-1) \times ... \times 1$

Exemple du jeu de 52 cartes.

- On peut :
  - tirer avec ou sans remise
  - considérer la séquence des cartes (tirage ordonné) ou l'ensemble obtenu (non ordonné).
- Quelle est la probabilité d'obtenir :
  - que des dames sur 4 tirages ?
  - 1,2,3,4 de la même couleur et dans l'ordre ?
  - d'obtenir une configuration donnée après un mélange aléatoire ?

Nombre de sous-ensembles

Nombre de sous-ensembles

- Le nombre de sous-ensembles distincts de cardinal k contenus dans E: $C_n^k = \frac{n!}{k!(n-k)!}$ (tirage sans remise non ordonné, C_n^k s'appelle aussi le *nombre de combinaisons de* k *parmi* n *éléments*
- Formule du binôme de Newton

$$(x+y)^n = \sum_{k=0}^{n} C_n^k x^{n-k} y^k \Rightarrow card\left(\mathcal{P}(\Omega)\right) = 2^n$$

- $C_n^k = \frac{A_n^k}{k!}$.

Tirage de 2 cartes sans remise

Quelle probabilité de n'obtenir aucune dame ?

$\frac{C_{48}^2}{1326} = 0.149$

Dénombrements : exemples (2)

Exemple : PMU

Un joueur parie toujours sur le même résultat :

- pour le quarté : les chevaux 1, 2, 3 et 4 vont terminer la course en premier (dans cet ordre).
- pour le 2 sur 4 : les chevaux 1 et 2 seront dans les 4 premiers arrivés.

On suppose qu'il y a toujours 15 chevaux dans une course, et que l'ordre d'arrivée des chevaux suit une probabilité uniforme.

Quelle est la probabilité que le joueur gagne au quarté et au 2 sur 4 ?