idl_07 : Chaînes de Markov pour la génération de texte

On a vu :

- comment classifier des documents (language identification), avec des méthodes statistiques et en entraînant un classifieur.

> Cette semaine : Comment générer automatiquement du texte ?
>
> ❓ sachant un mot donné, prédire le mot suivant dans la phrase.

Rappel sur la notion de probabilité :

- toujours > 0
- compris entre 0 et 1

Statistique vs probabilité :

- Une statistique vise à décrire un évènement dans un vase clos.
- Une probabilité cherche à prédire un évènement avec un degré de certitude.

Une probabilité peut utiliser des statistiques.

> variable aléatoire : Une variable aléatoire est une formalisation mathématique d'une quantité aléatoire, i.e. qui dépend du hasard.
>
> Le terme « variable aléatoire » est trompeur car mathématiquement une variable aléatoire n'est ni une variable ni un objet aléatoire. Formellement, une variable aléatoire est une application (i.e. une fonction) [...] qui associe pour chaque éventualité une valeur.

Le lancer de dé 🎲

Variable aléatoire : valeur prise par le dé.

- Dé bien formé (non pipé)
- Lancement indépendant

- $P(X = 1) = P(X = 2) = [...] = P(X = 6) = 1/6$
- $P(X = 1) + P(X = 2) + [...] + P(X = 6) = 1$

Le lancer de pièce

Variable aléatoire : valeur prise par la pièce (pile ou face).

Probabilité conditionnelle $P(a|b)$ = Probabilité de l'événement "a" sachant l'événement "b".

- $P(X_n = \text{pile} | X_{n-1} = \text{face}) = ?$
- $P(X_n = \text{pile} | X_{n-1} = \text{face}, X_{n-2} = \text{face}) = ?$
- $P(X_n = \text{pile} | X_i = \text{face}, i \in [1, n-1]) = ?$

Voir l'erreur du parieur (https://fr.wikipedia.org/wiki/Erreur_du_parieur)

Comprendre les chaînes de Markov

Toutes les séquence d'événements ne sont pas indépendantes les unes des autres !

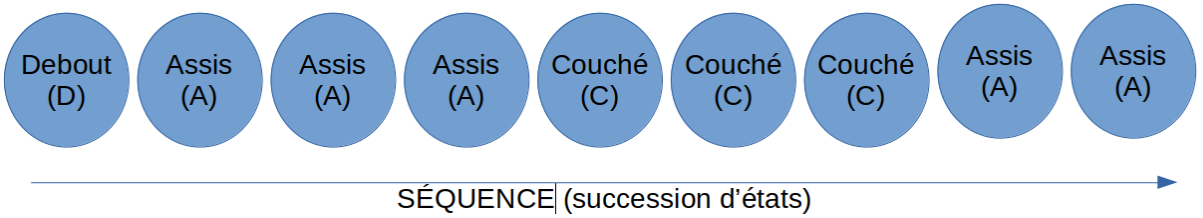
La probabilité d'un événement futur peut dépendre des événements qui précèdent.

▾ Details

- Caractères
- Mots
- Notes de musique
- Position d'un chat sur un canapé
- Météo

> Comment modéliser ces phénomènes ?

Les positions du chat



 On considère ici que la position à t+1 du chat ne dépend que de sa position à t

Peut-on prédire l'état suivant ?

> une chaîne de Markov est un processus de Markov à temps discret, ou à temps continu et à espace d'états discret. Un processus de Markov est un processus stochastique possédant la propriété de Markov.
>
> Un processus stochastique ou processus aléatoire (voir Calcul stochastique) ou fonction aléatoire (voir Probabilité) représente une évolution, discrète ou à temps continu, d'une variable aléatoire.
>
> La Propriété de Markov : l'information utile pour la prédiction du futur est entièrement contenue dans l'état présent du processus et n'est pas dépendante des états antérieurs (le système n'a pas de « mémoire ») :
>
> $P(X_{t+1} = i_{t+1} | X_0 = i_0, X_1 = i_1, \ldots, X_t = i_t) = P(X_{t+1} = i_{t+1} | X_t = i_t)$

Une chaîne de Markov peut avoir un ordre, c'est-à-dire la taille du contexte observé à l'instant t pour prédire l'état suivant.

ordre 1 : $P(X_{t+1} = i_{t+1} | X_t = i_t)$
ordre 2 : $P(X_{t+1} = i_{t+1} | X_t = i_t, X_{t-1} = i_{t-1})$
ordre n : $P(X_{t+1} = i_{t+1} | X_t = i_t, \ldots, X_{(n-1)} = i_{t-(n-1)})$

Application en TAL (NLP)

- Correcteur orthographique
- Reconnaissance de la parole
- Reconnaissance des écrits manuels
- Annotation des catégories des mots (parties de discours)
- Génération de texte (TP de demain)
- Identification d'auteur
- Analyse de sentiment
- Composition musicale

> Un modèle de langue est un modèle mathématique (probabiliste) qui modélise la probabilité qu'un mot apparaisse étant donné un contexte.

Une chaîne de Markov qui prend en contexte n mots est appelée également modèle de langue n-grammes.

Préparation TP demain :

Modélisation : on considère les mots d'une langue comme une variable aléatoire.

- On commence par compter les mots dans un corpus pour créer une distribution de probabilité sur les mots.
  Corpus avec 1 million de mots :
  - P (X = le) = 1000/100000
  - P (X = avec) = 200/1000000
  - P(X = Jean) = 5/1000000
  - etc.

la somme de toutes les probabilités pour tous les mots = 1

> ⚠️ évidemment, cette distribution dépend du corpus de départ !

- Comment trouver la probabilité d'apparition d'un mot sachant le contexte qui précède ?
  Par exemple, on cherche à prédire la suite de "Bob aime les..."

Chercher dans les corpus toutes les occurrences du contexte pour trouver le mot qui peut suivre et sa distribution de probabilités : $P(? \mid \text{Bob aime les})$

- "Bob aime les chiens" : 2
- "Bob aime les glaces" : 1
- "Bob aime les Schtroumpfs" : 1

On obtient :

$$P(X = \text{chiens} \mid \text{Bob aime les}) = 2/4$$
$$P(X = \text{glaces} \mid \text{Bob aime les}) = 1/4$$
$$P(X = \text{Schtroumpfs} \mid \text{Bob aime les}) = 1/4$$

❓ Ici, on a fait une recherche avec un ordre 3 : quel problème cela peut-il poser ?

▾ Details
On peut s'intéresser à d'autres ordres :

- ordre 1 (unigramme) : $P(X \mid \text{les})$
- ordre 2 (bigramme) : $P(X \mid \text{aime les})$
- ordre 3 (trigramme) : $P(X \mid \text{Bob aime les})$

Demain : Travail sur le corpus de votre choix

Voir ce que ça donne en fonction des tailles.

- https://www.gutenberg.org/browse/languages/fr (https://www.gutenberg.org/browse/languages/fr)
- contenu de wikipédia (style "encyclopédique")
- paroles de chanson