Lesson 1 : Word prediction via Markov chains

(or Writing like your favorite author)

In this lesson of introductory nature, we start by something easy : word prediction. The idea is to see a text as a sequence of random words. These are not independent & the simplest tool to model weak dependence is Markov Chains. In fact, Andrei Andreevich Markov (1856-1922) designed Markov chains with the example of TEXT in mind - only that states were just "vowel" & "consonant".

I Formal definition :

Let $E = \{1, 2, \dots, n\}$ be a finite set, "the state space"

$P = (P_{ij})_{1 \leq i,j \leq n}$ is a stochastic matrix

ie $\left(\sum_j P_{ij} = 1 \ \forall i \iff P \mathbb{1} = \mathbb{1} \right)$

Then the Markov chain with { state space E, transition matrix P } is the E-valued stochastic process $(X_n)_{n \in \mathbb{N}}$ defined by

- $X_0 \in E$
- $P_{x,y} = \mathbb{P}(X_{n+1} = y \mid X_n = x)$
  $= \mathbb{P}(X_{n+1} = y \mid X_n = x, X_{n-1}, \dots$

This is a generalizat° of independent sequences - here dependence is only through the previous state.

The following important Theorem generalizes the law of large numbers.

Thm [Birkhoff's ergodic Thm for Markov chains]

If X is irreducible

Then $$\frac{1}{T}\sum_{t=0}^{T} f(X_t) \xrightarrow[T\to+\infty]{a.s} \mathbb{E}_{\mathcal{L}(X)=\pi}(f(X)) = \sum_{x\in E} f(x)\,\pi(x)$$

where π is a probability measure called the invariant measure, solution to $\pi P = \pi$

II The historical example Markov Andrei was in disagreement with Pavel Nekrasov that independence is necessary for the (weak) law of large numbers. To that endeavor, in a paper written in 1913, Markov chose a sequence of $T = 20\,000$ letters from Pushkin's "Eugene Onegin", He obtained the Markov chain with state space $E = \{V, C\}$

$$P = \begin{array}{c} \\ V \\ C \end{array}\begin{array}{c} \begin{array}{cc} V & C \end{array} \\ \begin{pmatrix} 0,128 & 0,872 \\ 0,663 & 0,337 \end{pmatrix} \end{array}$$

Computing π from $\pi P = \pi$, one finds $\pi = (0,432; 0,56$

This generalizes easily to the problem of word prediction by setting

E = { words appearing in a text }

$\mathbb{P}$ = Transition matrix for the Markov chain whose realizat° is our text.

↳ How to estimate $\mathbb{P}$? How did Markov do?

Take a long text (say $T = 20\,000$ letters like Markov) & compute $\hat{\mathbb{P}}_T = \begin{matrix} \\ V \\ C \end{matrix} \begin{matrix} \begin{matrix} V & C \end{matrix} \\ \begin{pmatrix} \hat{\alpha}_T & 1-\hat{\alpha}_T \\ 1-\hat{\beta}_T & \hat{\beta}_T \end{pmatrix} \end{matrix}$

where
$$\begin{cases} \hat{\alpha}_T = \dfrac{\#\{1 \le t \le T-1 \mid X_t = X_{t+1} = V\}}{\#\{1 \le t \le T-1 \mid X_t = V\}} \\ \hat{\beta}_T = \dfrac{\#\{ \text{---} \mid X_t = X_{t+1} = C\}}{\#\{ \mid X_t = C\}} \end{cases}$$

↳ Why does it work?

Only the denominators are amenable to Birkhoff's thm

as $\frac{1}{T}\#\{1 \le t \le T-1 \mid X_t = x\} = \frac{1}{T}\sum_{t=1}^{T-1} \mathbb{1}_{\{X_t = x\}} \longrightarrow \pi(x)$

Numerators can be treated with an extended version of Birkhoff. However, we shall prefer a very generic method in statistics & Machine learni

III - Maximum Likelihood Estimation.

Proposition: $\left[\hat{P}_T\right]_{ij} := \dfrac{n_{ij}}{\sum_{j'} n_{ij'}}$ & $n_{ij} := \{1 \leq t \leq T-1 \mid (X_t, X_{t+1}) = (i,j)\}$

defines a matrix $\hat{P}_T$; which is the consistent MLE estimator for P

Proof: • Deriving Likelihood & CV.

Suppose we observe the realization $(x_0, x_1, x_2, \dots, x_T)$, and x_0 is deterministically chosen initial state.

Then the likelihood is the probability of our realizato.

$$\begin{aligned} L &= \mathbb{P}(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) \\ &= \mathbb{P}(X_1 = x_1 \mid X_0 = x_0) \cdots \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}) \, \mathbb{P}(X_1 = x_1 \\ &= \mathbb{P}(X_0 = x_0) \prod_{t=0}^{T-1} P_{x_t, x_t} = \mathbb{P}(X_0 = x_0) \prod_{i,j=1}^{n} P_{ij}^{n_{ij}} \end{aligned}$$

$\rightsquigarrow$ **MLE:**

$$\begin{aligned} \hat{P}_T &:= \operatorname*{Argmax}_{P \text{ stochastic matrix}} \log L(P) \\ &= \operatorname*{Argmax}_{P} \sum_{i,j=1}^{n} n_{ij} \log P_{ij} \end{aligned}$$

Thanks to general theorems of existence of MLE estimators (for example C. Pellegrini's course)

$\hat{P}_T \xrightarrow{T \to +\infty} P$ stochastic matrix.

- Computing $\hat{P}_T$:

/!\ The set of stochastic matrices has n constrains

$\rightsquigarrow$ Method of Lagrange Multipliers

$$\mathcal{L}(P, \lambda) := \sum_{i,j=1}^{n} n_{ij} \log P_{ij} - \sum_{i=1}^{n} \lambda_i \left(\sum_{j=1}^{n} P_{ij} - 1 \right)$$

for $\begin{cases} P \text{ matrix} \\ \lambda \in \mathbb{R}^n \end{cases}$

(P, λ) is critical

$$\iff \begin{cases} \forall i, \frac{\partial \mathcal{L}}{\partial \lambda_i} = 0 \\ \forall i,j, \frac{\partial \mathcal{L}}{\partial P_{ij}} = 0 \end{cases} \iff \begin{cases} \sum_{j=1}^{n} P_{ij} = 1 & \forall i \\ \forall i,j, \ 0 = \frac{n_{ij}}{P_{ij}} - \lambda_i \end{cases}$$

$$\iff \begin{cases} \forall i,j, \ P_{ij} = \frac{n_{ij}}{\lambda_i} \\ 1 = \sum_{j=1}^{n} P_{ij} = \left(\sum_{j=1}^{n} n_{ij} \right) / \lambda_i \end{cases}$$

Hence the result

/!\