Introduction to Data Science - 1MS041

Benny Avelin

Department of Mathematics

HT 2023

Recall from last time

- We explored how a computer which is fully deterministic, can produce something that looks random, i.e. pseudorandom.

Recall from last time

- We explored how a computer which is fully deterministic, can produce something that looks random, i.e. pseudorandom.
- Pseudorandom sequences

Recall from last time

- We explored how a computer which is fully deterministic, can produce something that looks random, i.e. pseudorandom.
- Pseudorandom sequences
- Period of a dynamical system

Recall from last time

- We explored how a computer which is fully deterministic, can produce something that looks random, i.e. pseudorandom.
- Pseudorandom sequences
- Period of a dynamical system
- We explored ways to go from our rudimentary dynamical sequence to something which is uniform $[0, 1]$.

Recall from last time

- We explored how a computer which is fully deterministic, can produce something that looks random, i.e. pseudorandom.
- Pseudorandom sequences
- Period of a dynamical system
- We explored ways to go from our rudimentary dynamical sequence to something which is uniform $[0, 1]$.
- Linear Congruential Generators

Recall from last time

- We explored how a computer which is fully deterministic, can produce something that looks random, i.e. pseudorandom.
- Pseudorandom sequences
- Period of a dynamical system
- We explored ways to go from our rudimentary dynamical sequence to something which is uniform $[0, 1]$.
- Linear Congruential Generators
- Bernoulli, discrete, shuffling

Recall from last time

- We explored how a computer which is fully deterministic, can produce something that looks random, i.e. pseudorandom.
- Pseudorandom sequences
- Period of a dynamical system
- We explored ways to go from our rudimentary dynamical sequence to something which is uniform $[0, 1]$.
- Linear Congruential Generators
- Bernoulli, discrete, shuffling
- Permutation test

Today

We have so far seen only independent and identically distributed random variables, today we will introduce a simple dependence and explore what that allows us to do.

Today

We have so far seen only independent and identically distributed random variables, today we will introduce a simple dependence and explore what that allows us to do.

Definition (Informal)

A sequence of random variables $X_1, \ldots$ is called a Markov chain if the distribution of X_t only depends on X_{t-1} and not on any X_s before $t-1$.

Markov Chain

Definition

A $\mathbb{R}$-valued **stochastic process** is a parametrized set of RVs. That is, we denote the collection

$$(X_\alpha)_{\alpha \in \mathbb{N}}$$

a $\mathbb{R}$-valued discrete stochastic process.

Markov Chain

Definition

A $\mathbb{R}$-valued **stochastic process** is a parametrized set of RVs. That is, we denote the collection

$$(X_\alpha)_{\alpha \in \mathbb{N}}$$

a $\mathbb{R}$-valued discrete stochastic process.

Example

Our standard i.i.d. sequence $X_1, \ldots, X_n$ is a discrete stochastic process!

Markov chain definition

Definition (Finite Markov Chain)

A stochastic process,

$$\{X_n : n \in \mathbb{N}\}$$

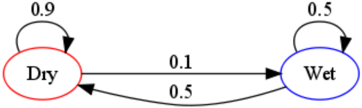
is a **Markov chain** with **state space** $\mathbb{X}$, if for any $t \in \mathbb{N}$ the following holds

$$\mathbb{P}(X_{t+1} = x | X_0, X_1, \ldots, X_t) = \mathbb{P}(X_{t+1} = x | X_t).$$

Simple weather example

Dry Wet Markov chain

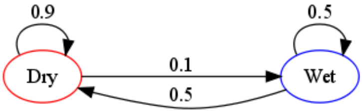
Consider recording whether it is raining or not on a day and calling 'rain = wet' and 'no rain = dry'. Let $\mathbb{X} = \{"dry", "wet"\}$ and let $X_t \in \mathbb{X}$. Consider the following transition probabilities.



Simple weather example

Dry Wet Markov chain

Consider recording whether it is raining or not on a day and calling 'rain = wet' and 'no rain = dry'. Let $\mathbb{X} = \{"dry", "wet"\}$ and let $X_t \in \mathbb{X}$. Consider the following transition probabilities.



We can write this as a matrix

$$P = \begin{bmatrix} \mathbb{P}(X_t = "dry" \mid X_{t-1} = "dry") & \mathbb{P}(X_t = "wet" \mid X_{t-1} = "dry") \\ \mathbb{P}(X_t = "dry" \mid X_{t-1} = "wet") & \mathbb{P}(X_t = "wet" \mid X_{t-1} = "wet") \end{bmatrix}$$

One step

Lets say we know $p_{t-1} = [\mathbb{P}(X_{t-1} = "dry"), \mathbb{P}(X_{t-1} = "wet")]$, then lets compute $p_1 = [\mathbb{P}(X_t = "dry"), \mathbb{P}(X_t = "wet")]$.

One step

Lets say we know $p_{t-1} = [\mathbb{P}(X_{t-1} = "dry"), \mathbb{P}(X_{t-1} = "wet")]$, then lets compute $p_1 = [\mathbb{P}(X_t = "dry"), \mathbb{P}(X_t = "wet")]$.

$$
\begin{aligned}
\mathbb{P}(X_t = "dry") = \mathbb{P}(X_t = "dry" \mid X_{t-1} = "dry")\mathbb{P}(X_{t-1} = "dry") \\
+ \mathbb{P}(X_t = "dry" \mid X_{t-1} = "wet")\mathbb{P}(X_{t-1} = "wet") \\
= (p_{t-1})_0 P(t)_{0,0} + (p_{t-1})_1 P(t)_{1,0}
\end{aligned}
$$

One step

Lets say we know $p_{t-1} = [\mathbb{P}(X_{t-1} = ''dry''), \mathbb{P}(X_{t-1} = ''wet'')]$, then lets compute $p_1 = [\mathbb{P}(X_t = ''dry''), \mathbb{P}(X_t = ''wet'')]$.

$$\begin{aligned}\mathbb{P}(X_t = ''dry'') = \mathbb{P}(X_t = ''dry'' \mid X_{t-1} = ''dry'')\mathbb{P}(X_{t-1} = ''dry'')\\ + \mathbb{P}(X_t = ''dry'' \mid X_{t-1} = ''wet'')\mathbb{P}(X_{t-1} = ''wet'')\\ = (p_{t-1})_0 P(t)_{0,0} + (p_{t-1})_1 P(t)_{1,0}\end{aligned}$$

in the same way

$$\begin{aligned}\mathbb{P}(X_t = ''wet'') = \mathbb{P}(X_t = ''wet'' \mid X_{t-1} = ''dry'')\mathbb{P}(X_{t-1} = ''dry'')\\ + \mathbb{P}(X_t = ''wet'' \mid X_{t-1} = ''wet'')\mathbb{P}(X_{t-1} = ''wet'')\\ = (p_{t-1})_0 P(t)_{0,1} + (p_{t-1})_1 P(t)_{1,1}\end{aligned}$$

Conclusion

In our wet dry chain, we have that $P(t) = P$ for any t, so

$$p_t = p_0 P^t$$

for $t = 0, 1, \ldots$.

See simulation

Homogeneity

Definition

We say that the Markov chain is **homogeneous** if

$$\mathbb{P}(X_{t+1} = y | X_t = x) = \mathbb{P}(X_{s+1} = y | X_s = x) = P_{xy}$$

for all $t, s \in \mathbb{N}$.

Estimation of transition probabilies

- Lets consider the case of the Dry-Wet Markov chain, with the transition matrix P, that we now assume is unknown and we want to estimate it from data.

Estimation of transition probabilies

- Lets consider the case of the Dry-Wet Markov chain, with the transition matrix P, that we now assume is unknown and we want to estimate it from data.
- Each row of P is the conditional distribution, and as such sums to 1.

Estimation of transition probabilies

- Lets consider the case of the Dry-Wet Markov chain, with the transition matrix P, that we now assume is unknown and we want to estimate it from data.
- Each row of P is the conditional distribution, and as such sums to 1.
- Thus there is actually only two parameters to find, we let these be $p_{0,0}$ and $p_{1,1}$.

Consider the log-loss and consider the corresponding risk of the full joint probability density up to n, i.e.

$$\mathbb{E}[\ln(p_n(X_1, \ldots, X_n))]$$

Consider the log-loss and consider the corresponding risk of the full joint probability density up to n, i.e.

$$\mathbb{E}[\ln(p_n(X_1, \ldots, X_n))]$$

Lets see what happens with $n = 2$, then using the tower property

$$\begin{aligned} \mathbb{E}[\ln(p_2(X_2, X_1))] &= \mathbb{E}[\ln(p_2(X_2 \mid X_1)p_1(X_1))] \\ &= \mathbb{E}[\ln(p_2(X_2 \mid X_1))] + \mathbb{E}[\ln(p_1(X_1))] \end{aligned}$$

Consider the log-loss and consider the corresponding risk of the full joint probability density up to n, i.e.

$$\mathbb{E}[\ln(p_n(X_1, \dots, X_n))]$$

Lets see what happens with $n = 2$, then using the tower property

$$\begin{aligned}\mathbb{E}[\ln(p_2(X_2, X_1))] &= \mathbb{E}[\ln(p_2(X_2 \mid X_1)p_1(X_1))] \\ &= \mathbb{E}[\ln(p_2(X_2 \mid X_1))] + \mathbb{E}[\ln(p_1(X_1))]\end{aligned}$$

doing this for n we get

$$\begin{aligned}\mathbb{E}[\ln(p_n(X_n, \dots, X_1))] &= \mathbb{E}[\ln(p_n(X_n \mid X_{n-1}, \dots, X_1)p_{n-1}(X_{n-1}, \dots, X_1))] \\ &= \mathbb{E}[\ln(p_n(X_n \mid X_{n-1})] + \mathbb{E}[\ln(p_{n-1}(X_{n-1}, \dots, X_1))] \\ &= \sum_{i=2}^{n} \mathbb{E}[\ln(p_i(X_i \mid X_{i-1}))] + \mathbb{E}[\ln(p_1(X_1))]\end{aligned}$$

Consider the log-loss and consider the corresponding risk of the full joint probability density up to n, i.e.

$$\mathbb{E}[\ln(p_n(X_1, \ldots, X_n))]$$

Lets see what happens with $n = 2$, then using the tower property

$$\begin{aligned}\mathbb{E}[\ln(p_2(X_2, X_1))] &= \mathbb{E}[\ln(p_2(X_2 \mid X_1)p_1(X_1))] \\ &= \mathbb{E}[\ln(p_2(X_2 \mid X_1))] + \mathbb{E}[\ln(p_1(X_1))]\end{aligned}$$

doing this for n we get

$$\begin{aligned}\mathbb{E}[\ln(p_n(X_n, \ldots, X_1))] &= \mathbb{E}[\ln(p_n(X_n \mid X_{n-1}, \ldots, X_1)p_{n-1}(X_{n-1}, \ldots, X_1))] \\ &= \mathbb{E}[\ln(p_n(X_n \mid X_{n-1})] + \mathbb{E}[\ln(p_{n-1}(X_{n-1}, \ldots, X_1))] \\ &= \sum_{i=2}^{n} \mathbb{E}[\ln(p_i(X_i \mid X_{i-1}))] + \mathbb{E}[\ln(p_1(X_1))]\end{aligned}$$

However, let us interpret X_1 as the initial state of the process, so we skip the last term.

Empirical risk

Risk for Markov chain

The risk of a Markov chain starting in X_1

$$R(p) = \sum_{i=2}^{n} \mathbb{E}[\ln(p_i(X_i \mid X_{i-1})]$$

Empirical risk

Risk for Markov chain

The risk of a Markov chain starting in X_1

$$R(p) = \sum_{i=2}^{n} \mathbb{E}[\ln(p_i(X_i \mid X_{i-1})]$$

For each of these terms we only have one observation, so the empirical risk is just plugging the values in for each of these terms, i.e.

$$\hat{R}(p) = \sum_{i=2}^{n} \ln(p_i(x_i \mid x_{i-1}))$$

Minimizing the empirical risk

Recall that we said that for our Dry-Wet chain we only need two values $p_{0,0}, p_{1,1}$. Thus we can express everything in terms of these now, as

$$p_i(x_i \mid x_{i-1}) = p_{x_i, x_{i-1}}$$

Minimizing the empirical risk

Recall that we said that for our Dry-Wet chain we only need two values $p_{0,0}, p_{1,1}$. Thus we can express everything in terms of these now, as

$$p_i(x_i \mid x_{i-1}) = p_{x_i, x_{i-1}}$$

lets introduce the numbers

$$n_{i,j} = \text{number of transitions from state } i \text{ to } j \text{ , } i = 0, 1$$

Minimizing the empirical risk

Recall that we said that for our Dry-Wet chain we only need two values $p_{0,0}, p_{1,1}$. Thus we can express everything in terms of these now, as

$$p_i(x_i \mid x_{i-1}) = p_{x_i, x_{i-1}}$$

lets introduce the numbers

$$n_{i,j} = \text{number of transitions from state } i \text{ to } j \ , i = 0, 1$$

then

$$\begin{aligned}\hat{R}(p) &= \sum_{i,j=0}^{1} \ln(p_{i,j}) n_{i,j} \\ &= \sum_{i=0}^{1} \ln(p_{i,i}) n_{i,i} + \ln(1 - p_{0,0}) n_{0,1} + \ln(1 - p_{1,1}) n_{1,0}\end{aligned}$$

$$\hat{R}(p) = \sum_{i=0}^{1} \ln(p_{i,i})n_{i,i} + \ln(1 - p_{0,0})n_{0,1} + \ln(1 - p_{1,1})n_{1,0}$$

$$\hat{R}(p) = \sum_{i=0}^{1} \ln(p_{i,i}) n_{i,i} + \ln(1 - p_{0,0}) n_{0,1} + \ln(1 - p_{1,1}) n_{1,0}$$

$$\partial_{p_{0,0}} \hat{R}(p) = \frac{1}{p_{0,0}} n_{0,0} - \frac{1}{1 - p_{0,0}} n_{0,1} = 0$$
$$\partial_{p_{1,1}} \hat{R}(p) = \frac{1}{p_{1,1}} n_{1,1} - \frac{1}{1 - p_{1,1}} n_{1,0} = 0$$

$$\hat{R}(p) = \sum_{i=0}^{1} \ln(p_{i,i}) n_{i,i} + \ln(1 - p_{0,0}) n_{0,1} + \ln(1 - p_{1,1}) n_{1,0}$$

$$\partial_{p_{0,0}} \hat{R}(p) = \frac{1}{p_{0,0}} n_{0,0} - \frac{1}{1 - p_{0,0}} n_{0,1} = 0$$
$$\partial_{p_{1,1}} \hat{R}(p) = \frac{1}{p_{1,1}} n_{1,1} - \frac{1}{1 - p_{1,1}} n_{1,0} = 0$$

Minima

$$p_{0,0} = \frac{n_{0,0}}{n_{0,1} + n_{0,0}}$$
$$p_{1,1} = \frac{n_{1,1}}{n_{1,0} + n_{1,1}}$$

How to simulate a Markov Chain

Definition (Random mapping representation (RMR))

A **random mapping representation** (RMR) of a transition matrix $P := (P(x,y))_{(x,y)\in\mathbb{X}^2}$ is a function

$$\rho(x,w) : \mathbb{X} \times \mathbb{W} \to \mathbb{X} \ , \tag{1}$$

along with a $\mathbb{W}$-valued random variable W, satisfying

$$\mathbb{P}\left(\{\rho\left(x,W\right) = y\}\right) = P(x,y), \quad \text{for each } (x,y) \in \mathbb{X}^2 \ . \tag{2}$$

Theorem

Let $W_1, \ldots, \overset{\mathrm{IID}}{\sim} F$ *such that* (ρ_t, W_t) *is a RMR for a transition matrix* P_t*, for all* $t \in \mathbb{N}$*. Then if* $X_0 \sim \mu_0$*,*

$$X_t := \rho_t(X_{t-1}, W_t), t \in \mathbb{N},$$

is a Markov chain with initial distribution μ_0 *and transition matrix* P_t *at time* t*.*

Apply it on data

Lets apply our newly found formula on some data