emulating data generation process - collect statistics from it - do that simulation

Economics 144
Economic Forecasting

Lecture 17
Monte Carlo Approximations and Markov Chains

Both monte carlo approximations and markov chains

Dr. Randall R. Rojas

Today's Class

- Monte Carlo Approximations
- Markov Chains: Historical Background
- Introduction to Markov Chains
  - Two-State Example
  - Three-State Example
- Properties of Markov Chains
- Types of Markov Chains
- Theorems
- MCMC Markov Chain Monte Carlo
- Gibbs Sampling Algorithm Understand it conceptually

Stochastic Processes

Monte Carlo Approximations

- Q: If X is a r.v where $X \sim f(X)$ then $g(X)$ is also a r.v., but what is the probability distribution of $g(X)$?

Example: Let $X \sim f(X) = \text{Unif}[1,4]$ and Y=g(X) =(X-3)^2, find $E[Y]$.

what does this mean?

final exam coverage - additional OH - recommendation letter/grad school

Step 1: $p(X) = 1/4 \; \forall \; X$ values in {1,2,3,4}

Step 2: $g(X)$

X = 1 → g(1) = 4

X = 2 → g(2) = 1

X = 3 → g(3) = 0

X = 4 → g(4) = 1

Step3: $P(g(X)) = P(Y)$

$$Y = 0, 1, 4$$

$$P(Y = 0) = 1/4, P(Y = 1) = 1/2, P(Y = 4) = 1/4$$

Step 4: $E[g(X)] = \sum Y \, P(Y)$

$$= 0(1/4) + 1(1/2) + 4(1/4)$$

$$= 3/2$$

Monte Carlo Approximations

- Q: What if we don't' know $P(Y)$? Do we need it?
- Example: Let $X \sim f(X) = \text{Unif}[1,4]$ and $Y = g(X) = (X-3)2$, find $E[Y]$.

using probabilities of X

$$E[Y] = E[g(X)] = \sum Y\,P(Y) = \sum g(X)\,P_X(X)$$

$$= (1\text{-}3)^2\,P(X=1) + (2\text{-}3)^2\,P(X=2) + (3\text{-}3)^2\,P(X=3) + (4\text{-}3)^2\,P(X=4)$$

$$= 4\,(1/4) + 1(1/4) + 0(1/4) + 1(1/4)$$

$= 3/2$ → same as before without knowing $P(Y)$ ☺

Conclusion: We do not need to know $P(Y)$ in order to compute $E[Y]$ provided we know $P(X)$.

= ***Foundation of MC Approximations!***

Monte Carlo Approximations

Application: Approximating Integrals

- How can we estimate: $F = \int_a^b f(x)dx$?

→

$$\langle F^N \rangle = (b-a)\frac{1}{N}\sum_{i=0}^{N-1} f(X_i)$$

average - bra-cket(different operations from left to right) - quantum-mechanics

Choose the points x at random between a and b, evaluate $f(x)$→ compute the area $A = l \times w$

repeated sampling help me get very good approximation

→ $P(\lim_{x\to\infty} \langle F^N \rangle = F) = 1$ → $E\left[\langle F^N \rangle\right] = F$

(Converges in probability) (by the LLN)

Algorithmic Implementation of Sequential Bayesian Learning

Application: Approximating Integrals Generalization (arbitrary PDF)

Since $F = \int_D f(\vec{x})d\vec{x} = \int_D \frac{f(\vec{x})}{p(\vec{x})} p(\vec{x})d\vec{x}$ for any PDF *p*

but $\int_D \frac{f(\vec{x})}{p(\vec{x})} p(\vec{x})d\vec{x} = E\left[\frac{f(\vec{x})}{p(\vec{x})}\right], x \sim p$

then $F_N = \frac{1}{N}\sum_{i=1}^{N} \frac{f(\vec{x}_i)}{p(\vec{x}_i)}$.

p is arbitary - use arbitrary function, so long as it's not 0, to compute it

→ (by the LLN) as $\boxed{N \to \infty, F_N \xrightarrow{a.s} E\left[\frac{f(\vec{x})}{p(\vec{x})}\right] = F}$

Monte Carlo Approximations

- Example: Approximate the integral: $F = \int_0^1 4\sqrt{1-x^2}dx = \pi$

Let $X_1, X_2, \ldots, X_n$ = i.i.d r.v.s. ~ unif[0,1].→$F^1 = f(X_1), \ldots, F^N = f(X_N)$ are i.i.d. r.v's with mean $\langle F^N \rangle$ and $E[\langle F^N \rangle] = F = \pi$

expected value - take a point, plug in my function - always take the average

average - simulate that

For $N = 1000$
→ $E[\langle F^N \rangle] = 3.15$

π

For $N = 10^5$
→ $E[\langle F^N \rangle] = 3.14$

Monte Carlo Approximations

- Example: Value at Risk Stock prices follow a random walk

  Given a current stock price (of your choice), what is the lowest it can go (worst-case scenario) 20 days from now with 99% probability of occurrence?

  Different values the price can take - 1% below - worst case scenario

- Model: Geometric Brownian Motion → $S_{t+1} = S_t \times (1 + \mu\Delta t + \sigma\varepsilon\sqrt{t})$
  - Standard model for stock price movements

emulate ending price
- looking at worst case
- having some kind of simple models

job interviews:
End up using Monte Carlo Simulations(Nuclear physics)

Imperative that u have certainty about how much energy can be released
William: randomly through 52 cards on floor - get straight flush - extremely difficult - do it many times and see how many times I got outcome/times I threw cards
- Neutron-diffusion problem
- Von Neuemann - Metropolis - Monte-Carlo Casino
- Use idea of simulating to get pretty accurate estimates
- promoted in mainstream science - RAM corporation/airforce
- beyond scope of physics

- Quantum monte-carlo - mean field theory
- simulating results - see what's the most likely outcome

Assumptions:

$S_0 = \$100$ Price - not acutally returns

$\mu = 0.1$ (10% drift)

$\sigma = 0.2$ (20% volatility)

ε ~unif[0,1]

sample from a uniform 0 to 1

Price drops below some threshold 10 times

← VaR = $66.28

Markov Chains: Historical Background

- **Andrey Markov** (Chebyshev was his advisor)

  - 1906 –First paper on the subject of Markov Chains

Looking at mathematical framework to describe how only the most recent event influences the final outcome

MA - short term memory - developing how would u describe mathematically events that are related in time - Paul Eherenfest - Diffusion model

- Paul Eherenfest: Diffusion model, Early 1900s

  - Statistical interpretation of the second law of thermodynamics: The entropy of a closed system can only increase. Proposed the "Urn Model" to explain diffusion.

Measure of disorder - ice cube very hard to rearrange itself back to atomic position of original particles - reversability of entropy - entropy: measure of disorder - shannon entropy

- Albert Einstein: Brownian Motion, 1905

  - Realized Brownian (after Robert Brown 1827) motion would provide a magnifying glass into the world of the atom. Brownian motion has been extensively modeled by Markov Chains.

Describe pollen particles in water - causing jagged motion
Mathematicaly he would describe it - Short term moemory that seem kinda random, but finally we have equations that characterizes these trajectories
- Beginning of markov chains

Introduction

- A **Markov Chain** is a mathematical model that describes a set of states and transitions between them.
  - Each transition probability is the probability of moving from one state to another in one step.
  - The transition probabilities are independent of the past, and depend only on the two states involved. The matrix of transition probabilities is called the transition matrix.

state + time component
Cold right now - hot/warm tomorrow
Historically evaluate what u seen - build a proper markov chains
state - transition probabilities that will take u from one state to the next

moving back and forth between these

Introduction

position, velocity, accelaration right now

- A **Markov Chain** consists of three elements:
  - A State Space $\boldsymbol{x}$ = set of values the chain can take
  - Note: Your textbook uses **S** = State Space

    all possible states in which a system can be

  - A Transition Probability $p(x^{t+1}|x^t)$ =probability of moving from state x^t to x^{t+1}

    Have to be neighboring probabilities
    Whole time - looking at the next step

  - In general, p_{ij} = conditional probability of being in state x_i at step $t+1$ given that the process was in state x_j at step t.

  - Initial condition distribution $\boldsymbol{\pi(0)}$= probability of being in any one of the possible states at $t=0$.

    bdfore I go in the system, what are the probabilities of each state at time 0 - Prior

Introduction

- Weather Forecast Example:
  - Assume on any given day it is either *sunny* (S)or *rainy* (R). Therefore, our state space $x = \{S, R\}$
  - Historical meteorological data reveal that:

$$\left.\begin{aligned} p(x^{Tomorrow} = S \mid x^{Today} = S) &= 0.7 \\ p(x^{Tomorrow} = S \mid x^{Today} = R) &= 0.2 \\ p(x^{Tomorrow} = R \mid x^{Today} = S) &= 0.3 \\ p(x^{Tomorrow} = R \mid x^{Today} = R) &= 0.8 \end{aligned}\right\} \quad P = \begin{array}{c c} & \begin{array}{cc} S^{t+1} & R^{t+1} \end{array} \\ \begin{array}{c} S^t \\ R^t \end{array} & \underbrace{\begin{pmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{pmatrix}}_{\text{Transition Probability}} \end{array}$$

rows add up to one

What we seen historically and the data - not making them up - have to get them from historical data

Exhaustive in the sense that it can only be one of the two states

  - Initial State: Today is a sunny day→ $x^0 = [1\ 0]$

what is the system in today?

current state of the system

  - Note: 1 = 100% sunny and 0% rainy

Introduction

- Q1: What is the probability that it will be sunny tomorrow given that today is sunny?

Multiply it

$$\rightarrow x^1 = x^0 P = [1\ 0] \begin{pmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{pmatrix} = [0.7\ 0.3]$$

Conclusion: There is a 70% that it will be sunny tomorrow given that it sunny today, and a 30% that it will be rainy.

Introduction

- Q2: What is the probability that it will be sunny two days from now given that today is sunny?

$$\rightarrow x^2 = x^1 P = (x^0 P)P = x^0 P^2$$

Higher orders - eventually made it stuck - steady-state solution

$$= [1\ 0] \begin{pmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{pmatrix} \begin{pmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{pmatrix}$$

$$= [0.55\ 0.45]$$

Conclusion: There is a 55% that it will be sunny two days from now given that it sunny today, and a 45% that it will be rainy.

Eventually plateaus

Introduction

- Q3: What is the probability that it will be sunny three weeks from now given that today is sunny?

$$\rightarrow x^{21} = x^{0} P^{21} = [1\ 0] \begin{pmatrix} 0.4 & 0.6 \\ 0.4 & 0.6 \end{pmatrix}$$

$$= [0.40\ 0.60]$$

Conclusion: There is a 40% that it will be sunny ^{21}two days from now given that it sunny today, and a 60% that it will be rainy.

Markov Model Representation

The **burn in period** for the Markov chain is the number of transitions it takes the chain to move from the initial conditions to the stationary distribution.

Introduction

- Financial Market Example:
  - Assume on any given day the market is either bullish, bearish, or stagnant. Therefore, our state space

$$x = \{Bull, Bear, Stagnant\}$$

Care more about equilibrium - we knew today we're in a bear market - what is the steady state solution - what we observed going forward
- If we do - which one is most likely - how long do I have to wait

P (To) (From)	**Bull**	**Bear**	**Stagnant**
Bull	0.900	0.075	0.025
Bear	0.150	0.800	0.050
Stagnant	0.250	0.250	0.500

Initial State: Today the market is Bearish → $x^0 = [0\ 1\ 0]$
→Find the steady state solution (Burn in Time)

Markov Model Representation

standard tool
Markov Chain
Power of the matrix

Introduction

- A sequence of trials of an experiment is a Markov Chain if:
  1. The outcome of each experiment is one of a set of discrete states;
  2. The outcome of an experiment depends only on the present state, and not on any past states.

- If the time parameter is discrete $\{t_1, t_2, t_3, \ldots\ldots\}$, it is called Discrete Time Markov Chain (DTMC).

- If time parameter is continues, ($t \geq 0$) it is called Continuous Time Markov Chain (CTMC).

Properties of Markov Chains

matrix have to be square
- if not, fill in by padding
- line up data -

Transition Matrix

$$P = pij = \text{State } t+1 \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \overbrace{\begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} \end{array}}^{\text{State } t}$$

- It is square matrix , since all possible states must be used both as rows and as columns.
- All entries are between 0 and 1, because all entries represent probabilities.
- The sum of the entries in any row must be 1, since the numbers in the row give you the probability of changing from the state at the left to one of the states indicated across the top.

Types of Markov Chains

- Def: Regular Markov Chains
  - A Markov chain is a **Regular Markov** chain if some power of the transition matrix has only positive entries. That is, if we define the (i, j) entry of P^n to be $p^n{}_{ij}$, then the Markov chain is regular if there is some n such that $p^n{}_{ij} > 0$ for all (i,j).

Regular markov chain
- all positive and not get just one
- common form of markov chian

Example (1)

$$P = \begin{pmatrix} 0.4 & 0.6 \\ 0.4 & 0.6 \end{pmatrix}$$

All entries are positive

Example (2)

$$P = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}$$

If you compute P^2, all entries are positive

Types of Markov Chains

- Def: Absorbing Markov Chains
  - A state S_k of a Markov chain is called an **Absorbing State** if, once the Markov chains enters the state, it remains there forever.
  - A Markov chain is called an absorbing chain if It has at least one absorbing state.
  - In an absorbing Markov chain, a state which is not absorbing is called **Transient**.
  - For every state in the chain, the probability of reaching an absorbing state in a finite number of steps is nonzero.

$$P = \begin{pmatrix} 0.3 & 0.6 & 0.1 \\ 0 & \mathbf{1} & 0 \\ 0.6 & 0.2 & 0.2 \end{pmatrix}$$

$P_{22} = 1$, $P^2_{22}=1$, ...,$P^n_{22}=1$

Types of Markov Chains

- Def: Irreducible (or Ergodic) Markov Chains
  - A Markov chain is called an **irreducible** or **ergodic** chain if it is possible to go from every state to every state (not necessarily in one move).
  - i and j communicate if they are accessible from each other. This is written $i \leftrightarrow j$

Ergotic/irreducable markov chians

Example (1)

$$P = \begin{array}{c c} & \begin{array}{cc} S_1 & S_2 \end{array} \\ \begin{array}{c} S_1 \\ S_2 \end{array} & \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \end{array}$$

This chain is ergodic since it is possible to move from any state to any other state

Example (2) Ehrenfest's urn

$$\mathbf{P} = \begin{array}{c c} & \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/4 & 0 & 3/4 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 3/4 & 0 & 1/4 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

In this example, if we start in state 0 we will, after any even number of steps, be in either state 0, 2 or 4, and after any odd number of steps, be in states 1 or 3. Thus this chain is ergodic but not regular.

Classification of States (Summary)

shows u which is the more "common" state

Let p_{jj} be the probability of returning to state j after leaving j.

- A state j is said transient if $p_{jj} < 1$
- A state j is said recurrent if $p_{jj} = 1$
- A state j is said absorbing if $p_{jj} = 1$
- A state j is said periodic if it is only visited in a number of steps which is multiple of an integer $m > 1$, called the period. A state j is said aperiodic otherwise
- A Markov chain is aperiodic if the period of each state is equal to 1
- A state j is said reachable from a state i if there is a path from i to j in the state transition diagram.
- A subset S of states is said closed if there is no transition leaving S.

Let T_{jj} be the average recurrent time, i.e., the time it takes to return to j

- A recurrent state j is positive recurrent if $E[T_{jj}]$ is finite.
- A recurrent state j is null recurrent if $E[T_{jj}] = \infty$.

Theory
behind it

Theorems

1. If a Markov chain has a finite state space, then at least one state is recurrent.
2. If i is a recurrent state and j is reachable from i, then state j is recurrent.
3. If S is a finite closed irreducible set of states, then every state in S is recurrent.
4. If i is a positive recurrent state and j is reachable from i, then state j is positive recurrent.
5. If S is a closed irreducible set of states, then every state in S is positive recurrent or every state in S is null recurrent or every state in S is transient.
6. If S is a finite closed irreducible set of states, then every state in S is positive recurrent.

Markov Chain Monte Carlo (MCMC)

- Design a Markov Chain on finite state space

$$\text{state space}: x^{(i)} \in \{x_1, x_2, \ldots, x_s\}$$

$$\text{Markov property}: \; p(x^{(i)} \mid x^{(i-1)}, \ldots, x^{(1)}) = T(x^{(i)} \mid x^{(i-1)})$$

…such that when simulating a trajectory of states from it, it will explore the state space spending more time in the most important regions (i.e. where p(x) is large)

Markov Chain Monte Carlo (MCMC)

- Supposing you browse this for infinitely long time, what is the probability to be at page x_i.
- No matter where you started off.

Larry page: user data

Pagerank Algorithm

Markov Chain Monte Carlo (MCMC)

Google vs. MCMC

- Google is given **T** and finds $p(x)$
- MCMC is given $p(x)$ and finds **T**
- But it also needs a transition probability distribution to be specified.
- Q: Do all MCs have a stationary distribution? →No

Markov Chain Monte Carlo (MCMC)

Conditions for existence of a unique stationary distribution

- MCMC samplers are irreducible and aperiodic MCs that converge to the target distribution
- These 2 conditions are not easy to impose directly
- Reversibility (also called 'detailed balance') is a sufficient (but not necessary) condition for $p(x)$ to be the stationary distribution.

system that has to be in some state of equilibirum

$$p(x^t)T(x^{t-1}|x^t) = p(x^{t-1})T(x^t|x^{t-1})$$

$$\rightarrow p(x^t) = \sum_{x^{t-i}} p(x^{t-1})T(x^t|x^{t-1})$$

  - It is easier to work with this condition

Markov Chain Monte Carlo (MCMC)

MCMC Algorithms

- Metropolis-Hastings
- Metropolis
  - Mixtures and Blocks
- Gibbs Sampling statistical sampling
- Sequential Monte Carlo & Particle Filters
- Others
- Auxiliary variable samplers
  - Hybrid Monte Carlo
  - Tries to avoid 'random walk' behavior, i.e. to speed up convergence
- Reversible jump MCMC
  - For comparing models of different dimensionalities (in 'model selection' problems)
- Adaptive MCMC

Particle Filter Algorithm

Don't want to waste resources with all other particles - resamples - reseamples from within these - re-weight, re-sample

Positions with updated weights - when these converges - that's your solution

Converges and intercept

Gibbs Sampling Algorithm

- Suppose $p(x,y)$ is a p.d.f. or p.m.f. that is difficult to sample from directly.
- Suppose, though, that we can easily sample from the conditional distributions $p(x/y)$ and $p(y/x)$.
- The Gibbs sampler proceeds as follows:
  1. Set x and y to some initial starting values
  2. Sample x/y, then sample y/x, then x/y, and so on.

Gibbs Sampling Algorithm

1. Set (x_0, y_0) to some starting value.
2. Sample $x_1 \sim p(x|y_0)$, i.e., from the conditional distribution $X|Y=y_0$

   Current state: (x_1, y_0)

   Sample $y_1 \sim p(y|x_1)$, i.e., from the conditional distribution $Y|X=x_1$

   Current state: (x_1, y_1)
3. Sample $x_2 \sim p(x|y_1)$, i.e., from the conditional distribution $X|Y=y_1$

   Current state: (x_2, y_1)

   Sample $y_2 \sim p(y|x_2)$, i.e., from the conditional distribution $Y|X = x_2$

   Current state: (x_2, y_2)

   ⋮

Repeat iterations 2 and 3, N times

Gibbs Sampling Algorithm

- Once completed, This procedure defines a sequence of pairs of random variables $(X_0, Y_0), (X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots$ that satisfy the property of being a Markov chain.
- The conditional distribution of (X_i, Y_i) given all of the previous pairs depends only on (X_{i-1}, Y_{i-1})
- (x_0,y_0) chosen to be in a region of high probability under $p(x,y)$, but often this is not so easy.
- We run the chain for N iterations and discard the first B samples $(X_1,Y_1),\ldots,(X_B,Y_B)$. This is called burn-in.
  - Typically: if you run the chain long enough, the choice of B does not matter.
- The performance of an MCMC algorithm, i.e., how quickly the sample averages converge, is referred to as the mixing rate.

Gibbs Sampling Algorithm

- Regression Example
  - R Code: https://github.com/stablemarkets/BayesianTutorials
  - Article: https://www.r-bloggers.com/bayesian-simple-linear-regression-with-gibbs-sampling-in-r/

- Model: $y_i \sim N(\beta_0 + \beta_1 x_i, \phi)$
- Estimate β_0 and β_1
- Assume Normal priors on β_0 and β_1
- Assume Inverse Gamma on the variance ϕ

$$
\begin{aligned}
y_i|\beta_0, \beta_1, \phi &\sim N(\beta_0 + \beta_1 x_i, \phi) \\
\beta_0|\mu_0, \tau_0 &\sim N(\mu_0, \tau_0) \\
\beta_1|\mu_1, \tau_1 &\sim N(\mu_1, \tau_1) \\
\phi|\alpha, \gamma &\sim IG(\alpha, \gamma)
\end{aligned}
$$

Bayesian Model

We will use Gibbs sampling to produces samples from the posterior distribution of each parameter of interest

as u increase number of partciples, things slow down a lot

Note: LS estimates are

$b_0 = 4.1$ and $b_1 = 5.2$

Powerful, but very powerful computing

Behind the scenes - have probability regression