MCMC

© Aaron Clauset
June 9, 2008

Generative Models

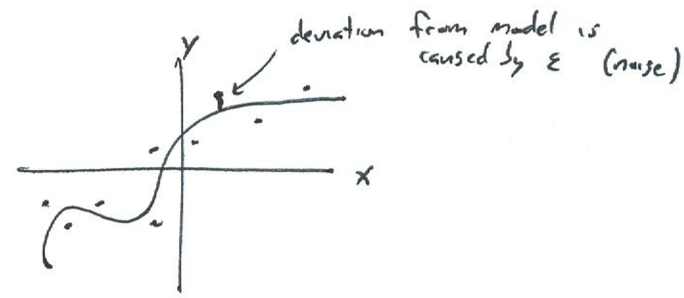
$G(\vec{\Theta})$ a model G with parameters $\vec{\Theta}$

- → generates data eg. edges in a network
- → can be arbitrarily complex
- → most useful models are relatively simple and stochastic (models measurement uncertainty)

$\boxed{\text{Example}}$: regression

$$G: \underbrace{y}_{\text{dependent variable}} = \underbrace{\overbrace{\sum_{k=0}^{n} \alpha_k x^k}^{\text{model is basically a nth-order polynomial.}} + \varepsilon}_{\text{independent variable}}$$

ε ← Gaussian or normal noise $N(0,1)$



$\vec{\Theta}: \alpha_k$

- → if ε is Gaussian, then least-squares is the maximum likelihood estimator (MLE) of $\vec{\Theta}$

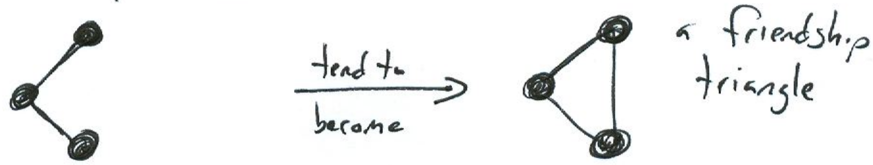
$\boxed{\text{Example}}$: Erdös-Rényi random graphs

$$G : \underset{\substack{\uparrow \\ \text{number of nodes in network}}}{G(n,} \underset{\substack{\nwarrow \\ \text{probability that any edge } (i,j) \text{ exists}}}{p)} \quad \longrightarrow \quad \overbrace{A_{ij}}^{\text{adjacency matrix}} = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{otherwise} \end{cases}$$

→ thus all edges are <u>independent</u> and <u>identically distributed</u>

<u>Comment</u> edges in real-world networks are obviously neither independent nor identically distributed

$\boxed{\text{Example}}$ a friendship triad



so, ER graphs are unlikely to be <u>good</u> models of <u>some</u> real networks

---

<u>Inference</u>

- given some data D <u>and</u> a family of generative models $G(\vec{\Theta})$
- can we estimate the particular Θ that <u>best describes</u> D ?

[redacted]

<u>Maximum Likelihood</u>

let $\mathbf{P}(D|G)$ be probability of observing data D from model G

$$P(D|G) = \prod_{D_i} G(D_i)$$ ← called the <u>likelihood</u> of the data

→ but we want $P(G|D)$, no?

from Bayes rule

$$P(G|D) = \underbrace{P(D|G)}_{\text{likelihood}} \frac{P(G)}{P(D)}$$

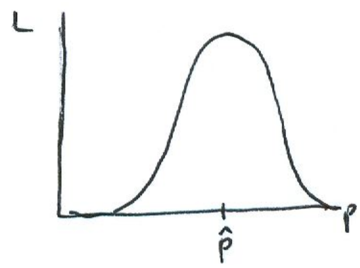
but $P(D)$ = fixed (our specific data)
$P(G) \propto$ constant (a uniform prior)

thus

$$P(G|D) \propto P(D|G)$$

- finding the $\hat{\Theta}$ that maximizes $P(D|G)$ is called the <u>method of maximum likelihood</u>

$\boxed{\text{Example}}$: ER graphs



let $G \equiv$ an ER random graph $G(n,p)$

$$L = P(D|G) = \binom{\binom{n}{2}}{m} p^m (1-p)^{\binom{n}{2}-m}$$

$$\frac{dL}{dp} = \left(\frac{p^m}{p}\right)\left(\frac{(1-p)^{\binom{n}{2}-m}}{(1-p)}\right)\left(m - p\binom{n}{2}\right)$$

set $\frac{dL}{dp} = 0$, solve for p

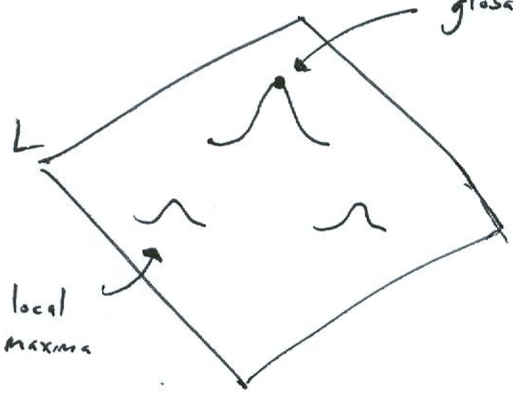
thus $\left.\hat{p} = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)}\right\}$ the MLE for p

- MLEs have many advantages including
  - being asymptotically unbiased
  - having normally distributed errors.
- likelihood can be used for model comparison
- all things considered, a good way to fit models to data.
- but, sometimes L is complicated and/or mathematics fails us.
- how do we find peaks in L despite this?

Finding Good Models

likelihood function $\approx -(\text{energy of system})$
(inference) (physics)

global maximum of L ≈ global minimum of E



- each point in L landscape = particular model $G(\theta)$
- finding good models = searching over L for high scores
- can "average" properties of interest over all the good models

  i.e. $\langle x \rangle = \int_\theta x\, p(D|G)\, d\theta$

How to search?

→ any optimization technique could work

e.g.
1. hill climbing (gradient ascent on $dL/d\theta$) → good if unimodal
2. simplex method (Nelder-Mead) → good if not too glassy
3. etc.

$\boxed{\text{Today: Markov chain Monte Carlo}}$ (MCMC) → good in general

- MCMC is a sampling algorithm, not an optimization technique
- approximates $P(D|G)$ by sampling each state G with prob. $\propto$ likelihood
- uses a correlated random walk to sample states
- define MCMC so that equilibrium distribution of walker positions equals $P(D|G)$

Comment

MCMC has many applications : volume of a convex body in d dimensions
: particle diffusion
: estimating $P(D|G)$

and has strong connections with Bayesian inference, statistics, physics, and computer science

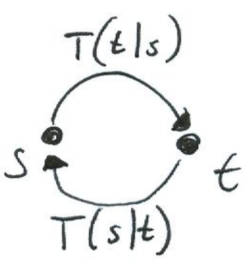
Requirements of MCMC

① ERGODICITY : every state s is reachable in Markov chain from every other state t with prob. > 0

② DETAILED BALANCE : there are no periodic cycles in walk and equilibrium distribution is the target distribution

Creating Detailed Balance

(s, t denote particular models.)



T denotes the transition matrix

detailed balance requires (flux in) = (flux out) for every state

$$\rightarrow p(s)T(t|s) = p(t)T(s|t)$$

for convenience, we split $T(t|s) = \underbrace{g(t|s)}_{\text{probability of proposing } t \text{ as next step}} \underbrace{a(t|s)}_{\text{probability of accepting proposal}}$

thus, substituting and rearranging

$$\frac{p(s)}{p(t)} = \frac{g(s|t)\,a(s|t)}{g(t|s)\,a(t|s)}$$

} detailed balance

→ $p(s)$ is the target distribution we want to sample

e.g. Boltzmann distribution

$$p(s) = \frac{1}{Z} e^{-E_s/kT}$$

How to choose g and a? (How do we generate proposals and how do we decide if we should accept?)

① without loss of generality, let $g(s|t) = g(t|s)$ → equal proposal rates

→ e.g. choose t uniformly given s

thus
$$\frac{p(s)}{p(t)} = \underbrace{\frac{a(s|t)}{a(t|s)}}_{\text{acceptance ratio}}$$

② for efficiency, we want to maximize acceptance ratio

$\boxed{A}$ MINIMIZATION (physics)

suppose $E(t) < E(s)$, then let $a(t|s) = 1$

→ always accept if $E(t)$ is smaller

and
$$a(s|t) = \frac{p(s)}{p(t)} = e^{-\beta(E(s)-E(t))}$$

$\beta = \frac{1}{kT}$ Boltzmann factor

→ otherwise accept sometimes

MCMC thus samples states according to the Boltzmann distribution of energies

$\boxed{B}$ MAXIMIZATION (inference)

suppose $L(t) > L(s)$, then let $a(t|s) = 1$

$\rightarrow$ always accept if likelihood improves

and $a(s|t) = \dfrac{L(s)}{L(t)} = e^{\log L(s) - \log L(t)} = e^{\Delta \log L}$

MCMC samples states (models) with probability proportional to their likelihood

$$a(s|t) = \begin{cases} e^{\Delta \log L} & \text{if } \Delta \log L < 0 \\ 1 & \text{otherwise} \end{cases}$$

this procedure called Metropolis - Hastings.

general MCMC recipe:

1. choose initial state / parameterization uniformly at random
2. PROPOSAL choose an adjacent state uniformly at random
3. accept this move with probability $a(s|t)$
4. goto #2

note: simulated annealing is MCMC but where $a(s|t)$ has a "temperature" dependence T and T decreases over time (often exponentially slowly). this produces an optimization technique.