

Tutkimustiedonhallinnan peruskurssi

Hannu Toivonen, Marko Salmenkivi, Inkeri Verkamo

`hannu.toivonen, marko.salmenkivi, inkeri.verkamo@cs.helsinki.fi`

Helsingin yliopisto – Tietojenkäsittelytieteen laitos

Johdatus Monte Carloon

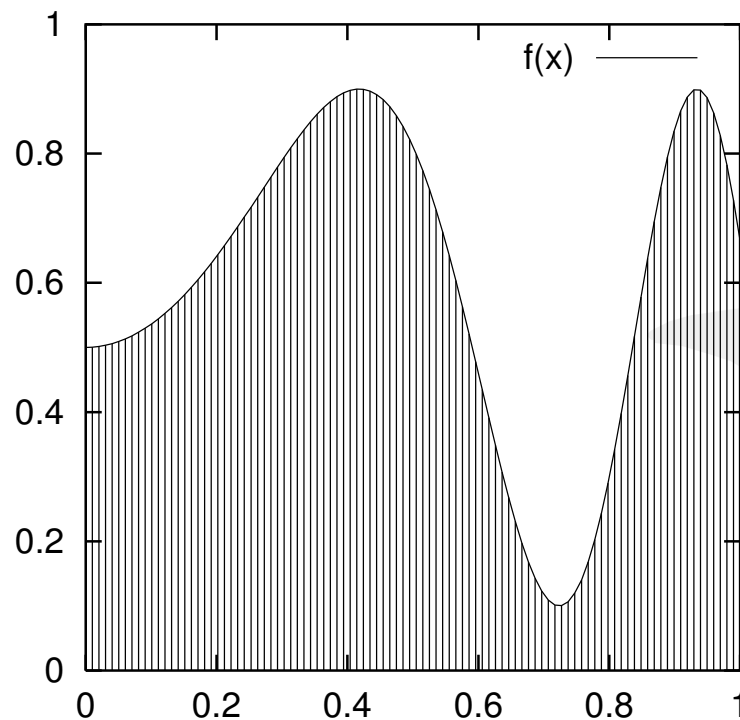


Monte Carlo

- Monte Carlo -menetelmät perustuvat satunnaisuuden hyväksikäyttöön
- joidenkin laskennallisesti vaikeiden ongelmien ratkaisu
- satunnaisuudesta johtuen ratkaisut ovat likimääräisiä
- vrt. satunnaisotos, jonka perusteella arvioidaan muuttujan jakaumaa
 - koska *satunnainen* ja *otos*, tulos on likimääräinen
 - etu: otanta ja jakauman arviointi on tehokasta
- Monte Carlo usein: mallista generoitu joukko satunnaisia pisteitä (ei otos aineistosta)

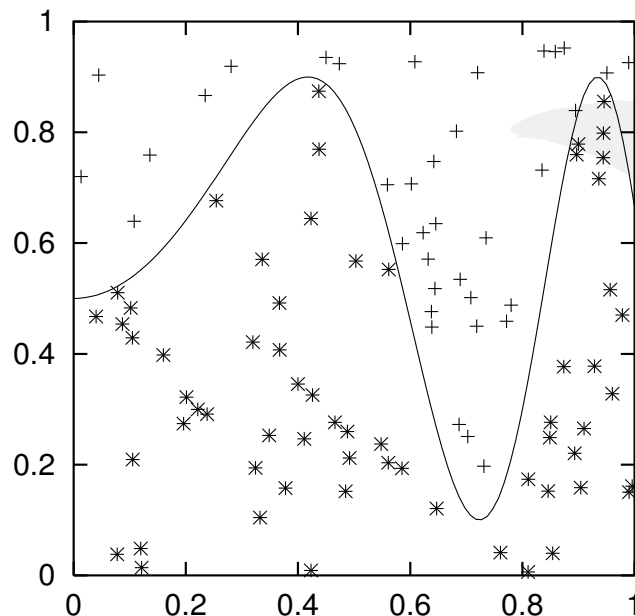
Monte Carlo -integrointi

- tyypillinen Monte Carlon sovellus: integrointi
- halutaan integroida funktio $f : [0, 1] \rightarrow [0, 1]$:
 $\int_0^1 f(x) dx = ?$



Monte Carlo -integrointi (2)

- generoidaan (tasaisesti jakautuneita) satunnaisia pisteitä alueelle $[0, 1] \times [0, 1]$
 - $x \in [0, 1]$ (integroitava väli)
 - $y \in [0, 1]$ ($f(x)$:n mahdolliset arvot kattava alue)
- kuvaajan alle osuneiden pisteiden osuus 0.63



Monte Carlo -integrointi (3)

- vähän yleisempi MC-integrointialgoritmi
 - ol. $f(x) \geq 0$
 - syöte: integroitavan välin alaraja AR ja yläraja ja YR, yläraja MAX $f(x)$:lle, satunnaispisteiden lkm K
- $k = 0$;
- for $i = 1$ to K
 - $x = \text{rand}(\text{AR}, \text{YR})$;
 - $y = \text{rand}(0, \text{MAX})$;
 - if $y \leq f(x)$ then $k++$;
- print $k / K \cdot (\text{YR} - \text{AR}) \cdot \text{MAX}$;

Shakkiesimerkki

- kaksi shakkiohjelmaa, A ja B, pelaavat 15 ottelua
- A voittaa niistä 10 ja B voittaa 5 (oletetaan että peli ei voi päättyä tasan)
- onko A siis parempi kuin B?
- ajatuskoe: jos A ja B ovat yhtä hyviä, niin millä todennäköisyydellä A voittaa sattumalta 10 ottelua?
- Monte Carlo -toteutus ajatuskokeelle
 - simuloi 15 ottelun turnauksia
 - arvo A kunkin ottelun voittajaksi tn:llä 0.5
 - kuinka suuressa osassa turnauksista A voitti vähintään 10 ottelua?

Otannan peruskäsitteitä

- mihin edellinen esimerkki perustuu?
- hyödyllisiä tilastollisia käsitteitä
 - satunnaismuuttuja
 - otos vs. populaatio
 - tunnusluku vs. parametri
 - otosjakauma
 - riippumaton toistokoe

Satunnaismuuttuja

- *satunnaismuuttuja* X on suure, joka saa eri arvoja eri todennäköisyyksillä
 - esim. $X = 1$ jos A voitti pelin ja $X = 0$ jos A hävisi
- satunnaismuuttujalla voi olla tietty hyvin määritelty arvo $X = x$
 - X on muuttuja, x on sen arvo
 - kun peli on pelattu, X :llä on arvo 0 tai 1

Otos vs. populaatio

- käytettävissä oleva aineisto on *otos* ilmiöstä
- n havaintoa: $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$
- koko ilmiön perusjoukkoa, josta otos on peräisin, nimitetään usein *populaatioksi*
- X_i :n jakaumasta käytetään usein nimitystä *populaatiojakauma*
- *tilastollinen päättely*: populaation ominaisuuksien arviointi otoksen perusteella

Tunnusluku vs. parametri

- populaatiojakauman tunnuslukuja kutsutaan usein *parametreiksi*, tällöin
 - tunnusluvun arvo on otoksen funktio
 - esim. otoskeskiarvo $\bar{x}$, otoskeskihajonta σ_x
 - parametrin arvo on koko populaation funktio (“todellinen” tunnusluku)
 - esim. satunnaismuuttujan odotusarvo $EX = \mu_X$ ja keskihajonta $VX = \sigma_X$
- käsitteellisesti siis kaksi eri asiaa, vaikkakin lähellä toisiaan

Otosjakauma

- mielivaltainen tunnusluku t
- tietty otoksen koko n
- t :n *otosjakauma* = tunnusluvun t (teoreettinen) jakauma kaikissa n :n kokoisissa otoksissa (iid)
- otosjakauma **ei** viittaa otoksen arvojen jakaumaan (engl. *sampling distribution* = otosjakauma, *sample distribution* = (yksittäisen) otoksen arvojen jakauma)
- sanat vaarallisen lähellä toisiaan, kysymys aivan eri asioista

Otosjakauma (2)

- tunnusluku t on myös satunnaismuuttuja
 - määritellään perusjoukoksi eli populaatioksi kaikki mahdolliset $n:n$ kokoiset otokset
 - tarkastellaan satunnaismuuttujaa $Y = t$
 - $Y:n$ populaatiojakauma = $t:n$ otosjakauma
- edelleen voidaan tarkastella $t:n$ (otos)jakauman tunnuslukuja (p-arvo tulee olemaan esimerkki tästä)

Jakaumien parametreja

- muuttujan *odotusarvo* EX on sen keskimääräinen arvo:

$$EX = \sum_x xP(X = x) \text{ tai } \int_x x f(x) dx$$

- *varianssi* σ_X^2 kuvaa jakauman leveyttä:

$$\sigma_X^2 = E(X - EX)^2$$

- *hajonta* eli standardipoikkeama:

$$\sigma = \sqrt{\sigma_X^2}$$

Iid-oletus

- *iid-oletus* (independent and identically distributed)
 - *riippumaton toistokoe*
 - kaikki X_i :t noudattavat samaa jakaumaa ja ovat toisistaan riippumattomia
- iid-oletus on yleinen
 - tämän kurssin MC-menetelmät tuottavat iid otoksia
- jos iid-oletus on voimassa, niin erityisesti
 - $E(\bar{x}) = EX$
 - $E(\sigma_x) = \sigma_X$
 - eli otoskeskiarvo ja -hajonta ovat harhattomia estimaatteja vastaaville parametreille

Shakkiesimerkki (2)

- mihin aiempi shakkiesimerkki siis perustui?
- satunnaismuuttuja $X_i = 1$, jos A voitti pelin i , muuten 0
- ajatuskokeessa X_i :n populaatiojakauma tunnettu: voiton ja tappion todennäköisyys $P(X = 1) = P(X = 0) = 0.5$
- tunnusluku $t = \sum_{i=1}^{15} X_i$, eli A:n voittamien otteluiden määrä 15 ottelusta
- t :n otosjakauman approksimointi
 - satunnaismuuttuja t :n populaatio: 15 ottelun turnaukset
 - t :n (otos)jakaumaa approksimoidaan tuottamalla satunnaisotoksia iid 15 ottelun turnauksista

Shakkiesimerkki (3)

- Monte Carlo -approksimaatio: otoksista laskettujen tunnusluvun arvojen jakauma
- voimakas abstraktio
 - 15 ottelun otos kaikkien mahdollisten pelien äärettömästä populaatiosta helppo toteuttaa
 - ei mitään oletuksia pelin tai pelaajien luonteesta, pelien jakaumista, tms.
- Monte Carlo on *parametriton* menetelmä (ei käytä oletuksia jakaumien muodoista)
- lisäämällä otosten määrää päästään mielivaltaisen tarkkaan approksimaatioon (suurten lukujen laki)