Yakov G. Sinai, Abelprisvinner 2014

ENTROPI AV 0,1-SEKVENSER

Vi betrakter et dynamisk system der tilstandsrommet består av alle uendelig lange 0,1-sekvenser og hvor dynamikken er gitt ved en skift-operator, som flytter oss ett hakk lenger ut i sekvensen. Systemet er oppkalt etter en av de store matematikerene på 1600-tallet, Jacob Bernoulli.

Vi betrakter en 50 tegn lang 0,1-sekvens:

11010010001010111011011000101010100011100110100011

Har vi noen grunn til å tro at denne sekvensen er helt tilfeldig generert?

Vi kan gjøre noen enkle analyser for om mulig å bygge opp under et svar.

1. Sekvensen inneholder 25 0-ere og like mange 1-ere. Dette passer med en forventning for en tilfeldig generert sekvens.

2. Hvis vi derimot teller antall ganger vi skifter tegn, finner vi at det skjer 30 ganger, mens vi ikke skifter tegn 19 ganger. I en tilfeldig generert sekvens ville disse tallene normalt være omtrent like store.

3. Sekvensen inneholder flere delsekvenser med 3 like tegn, men ingen med 4 like eller flere. I en tilfeldig valgt tilfeldig generert sekvens med 50 tegn, vil det være 98% sannsynlighet for at det finnes en delsekvens med 4 like tegn. Det at denne sekvensen ikke har noen slik delsekvens er enda en indikasjon på at den ikke er tilfeldig generert.

Vi konkluderer derfor med at vi ikke tror at sekvensen er tilfeldig generert.

Sannheten er at sekvensen er manuelt generert i et forsøk på å produsere noe som skulle kunne framstå som tilfeldig. Men feilen vi gjør, og som analysen avslører oss på, er at i vår streben etter å framstå så lite systematisk som mulig, så skifter vi tegn for ofte.

Entropi av Bernoulli-skjemaer

En 0,1-sekvens kalles også et Bernoulli-skjema. Dersom prosessen er helt tilfeldig vil sannsynligheten p for at neste tegn er 0 være den samme som for at neste tegn er 1, dvs. $p = \frac{1}{2}$. I eksempelet over kan det virke som om 0 og 1 opptrer med like stor sannsynlighet, men at kombinasjonene 01 og 10, som begge innebærer et skift, er mer sannsynlig, si 60/40, enn kombinasjonene 00 og 11. Denne forskjellen i forutsigbarheten om hva som kommer, måles ved systemets **entropi**. Det mest uforutsigbare systemet, dvs. for $p = \frac{1}{2}$, har høyest entropi, i dette tilfellet 0,693. 60/40-systemet har entropi 0,673, altså litt lavere. Generelt vil et Bernoulli-skjema med to utfall med sannsynlighet p og $1 - p$ ha entropi gitt ved

$$E = -p \ln p - (1 - p) \ln (1 - p)$$

Et Bernoulli-skjema kan imidlertid ha flere enn to utfall. Mengden av alle uendelige bokstavsekvenser er et Bernoulli-skjema med 29 utfall. Et enkelt spørsmål om Bernoullli-skjema, opprinnelig stilt av matematikeren John von Neumann, fikk stor betydning for forskningen rundt entropi. Von Neumann lurte på om det er mulig at to strukturelt forskjellige Bernoulli-skjema kan gi samme resultat. Er det for eksempel slik at $BS(\frac{1}{2}, \frac{1}{2})$ og $BS(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ på en eller annen måte kan identifiseres? BS står her for Bernoulli-skjema, og tallene gir oss sannsynligheten for hvert utfall.

$BS(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ angir dermed Bernoulli-skjemaet med tre like sannsynlige utfall.

Svaret på von Neumanns spørsmål kom først i 1970, presentert av den amerikanske matematikeren Donald Ornstein. Og nei, det er ikke mulig at to strukturelt forskjellige Bernoulli-skjema gir oss samme resultat. Grunnlaget for Ornsteins resultat ble lagt av Sinai og Kolmogorov i 1959. Det viste seg nemlig at Kolmogorov-Sinai-entropien var nøyaktig det som skulle til for å skille Bernoulli-skjemaene fra hverandre. Like Bernoulli-skjemaer har samme entropi, og forskjellige Bernoulli-skjemaer har forskjellig entropi.

Et kombinatorisk resultat

Innledningsvis ble det påstått at sannsynligheten for at en tilfeldig 0,1-sekvens med 50 tegn inneholder en delsekvens av minst 4 like tegn er 98%. Vi skal se litt nærmere på en mer generell versjon av denne påstanden.

En 0,1-sekvens av lengde n er tilfeldig generert. Hvor stor er sannsynligheten for at sekvensen inneholder en delsekvens med m etterfølgende like siffer, enten 0 eller 1?

Vi skal gi svaret rekursivt. La X_n være mengden av 0,1-sekvenser av lengde n. Mengden X_n har 2^n elementer. La $q(n, m)$ være antall sekvenser som inneholder m påfølgende like siffer, enten 0 eller 1, og $p(n, m)$ sannsynligheten for å velge ut en slik sekvens fra alle mulige sekvenser, dvs. $p(n, m) = \frac{q(n,m)}{2^n}$.

La $\xi \in X_{n+1}$ være en sekvens som inneholder m påfølgende like tegn. Fjern det siste tegnet fra sekvensen. Da har vi tre muligheter for den reduserte sekvensen:

1. Den inneholder fortsatt en delsekvens med m etterfølgende like tegn, eller
2. Den ender med en delsekvens med presis $m-1$ etterfølgende like tegn, hvilket innebærer at forløperen til de siste $m-1$ tegnene er forskjellig fra sin etterfølger, eller
3. Både 1. og 2. er oppfylt.

Kardinaliteten til mengdene som oppfyller 1.-3. er som følger: Den første har $q(n, m)$ elementer, den andre har 2^{n+1-m} elementer, og den tredje har $2^{n+1-m} \cdot p(n+1-m, m)$ elementer. Det gir oss en rekursjon

$$q(n+1, m) = 2q(n, m) + 2^{n+1-m} - 2^{n+1-m} \cdot p(n+1-m, m)$$

eller for sannsynligheter, dvs. dividert med 2^{n+1},

$$p(n+1, m) = p(n, m) + 2^{-m}(1 - p(n+1-m, m))$$

Det er opplagt at $p(n, 1) = 1$, at $p(n, n) = 2^{1-n}$ og at $p(n, m) = 0$ for $n < m$. Tabellen viser noen flere utregninger for $p(n, m)$:

n/m	2	3	4	5	6
1	0	0	0	0	0
5	0,938	0,5	0,188	0,063	0
10	0,998	0,826	0,465	0,217	0,094
20	1	0,979	0,768	0,458	0,237
35	1	0,999	0,934	0,689	0,410
50	1	1	0,981	0,821	0,544

Vi merker oss at for sekvenser med 50 tegn, så vil et flertall av sekvensene inneholde delsekvenser med 6 etterfølgende like tegn, og hele 98% inneholder delsekvenser med 4 etterfølgende like tegn.