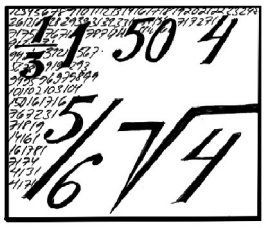

Käänteistä marjanpoimintaa

Jukka Liukkonen
Metropolia Ammattikorkeakoulu

Mistä on kysymys?

Mustikan pinnalla on normaalisti ohut ultraviolettivaloa heijastava vahakerros. Näkyvän valon aallonpituudella kerros aiheuttaa sen, että marja näyttää siniseltä. *Tervamustikaksi* kutsutulta kiiltävänmustalta värimuunnokselta vahakerros puuttuu. Teeret ja punakylkirastaat suosivat sinisiä mustikoita, sillä linnut ovat mieltyneitä ultraviolettivaloon. Tervamustikat ovat melko harvalukuisia, mutta joillain alueilla niitä tavataan runsaastikin.

Sammeli on käynyt keräämässä kipollisen mustikoita. Velipoika Miihkali yrittää arvata, mistä Sammeli on mustikkansa hakenut. Ahkujärven rannoilla kasvavista mustikoista $p_A \cdot 100$ % on tervamustikoita, Bieggajängän laitamien mustikkasadossa tervamustikoiden osuus on $p_B \cdot 100$ %. Muita mustikkapaikkoja lähimaastossa ei ole. Sammelin keräämässä *otoksessa* tervamustikoiden osuus kaikista mustikoista ei välttämättä ole lähelläkään mainittuja prosenttimääriä, mutta keskittyäksemme Miihkalin käyttämään päättelymenetelmään tarkastelemme pelkistettyä tilannetta, jossa Sammelin marjasaaliin tervamustikkapitoisuus on tasan $p_A \cdot 100$ % tai tasan $p_B \cdot 100$ %. Edellisessä tapauksessa marjat tulkitaan Ahkujärven ja jälkimmäisessä Bieggajängän sadon osaksi. Sammeli ei anna Miihkalin tutkia kipponsa sisältöä, vaan näyttää hänelle kiposta summamutikassa ottamansa marjan yksi kerrallaan palauttaen mustikan aina takaisin kippoon. Tällaista kutsutaan *otannaksi takaisinpanolla*. Se jatkuu kunnes laskintaan näpräilevä Miihkali lopulta rohkaistuu ilmoittamaan hyvinkin valistuneen arvauksen mustikoiden alkuperästä. Normaalikäsittelyssä mustikan vahakerros kuluu helposti pois, mutta nyt oletamme vahan pysyvän. Millainen on Miihkalin menetelmä?

Matemaattinen malli

Todennäköisyyslaskennan kielellä ilmaistuna kyseessä on *toistokoe*, jossa perättäiset toistot ovat toisistaan *riippumattomat*. Mustan (M) ja sinisen (S) värin todennäköisyydet ehdoilla, että marja on poimittu Ahkujärveltä (A) tai vastaavasti Bieggajängältä (B), ovat

$$P(M|A) = \frac{P(M \cap A)}{P(A)} = p_A, \quad P(S|A) = 1 - p_A,$$
$$P(M|B) = p_B, \qquad\qquad P(S|B) = 1 - p_B.$$

Sulkeaksemme pois mielenkiinnottomat erikoistapaukset oletamme, että $0 < p_A < 1$, $0 < p_B < 1$ ja $p_A \neq p_B$. Olkoon V_n marjojen värien jono siinä vaiheessa, kun Miihkali on nähnyt n mustikkaa, ja $m(n)$ mustan värin esiintymiskertojen lukumäärä jonossa V_n. Jos esimerkiksi kaksi ensimmäistä mustikkaa ovat sinisiä ja kolmas musta, värijono on $V_3 = (S, S, M)$, ja $m(3) = 1$.

Tällaisen jonon esiintymistodennäköisyydet ovat

$$P(V_3|A) = P(S|A)P(S|A)P(M|A) = p_A(1-p_A)^2,$$
$$P(V_3|B) = P(S|B)P(S|B)P(M|B) = p_B(1-p_B)^2.$$

Yleisesti

$$P(V_n|A) = p_A^{m(n)}(1-p_A)^{n-m(n)},$$
$$P(V_n|B) = p_B^{m(n)}(1-p_B)^{n-m(n)}.$$

Todennäköisyyden kertolaskusäännöstä

$$P(A|V_n)P(V_n) = P(V_n|A)P(A)$$

saadaan *Bayesin kaava*

$$P(A|V_n) = \frac{P(V_n|A)P(A)}{P(V_n)}.$$

Kokonaistodennäköisyyden kaava

$$P(V_n) = P(V_n|A)P(A) + P(V_n|B)P(B)$$

antaa Bayesin kaavalle muodon

$$P(A|V_n) = \frac{P(V_n|A)P(A)}{P(V_n|A)P(A) + P(V_n|B)P(B)} = \frac{1}{1 + \frac{P(V_n|B)}{P(V_n|A)}\frac{P(B)}{P(A)}}. \quad (1)$$

Jos näiden todennäköisyyksien jono lähestyisi nollaa otannan edistyessä, olisi varmaankin kyse Bieggajängän mustikoista. Jos jono lähestyisi ykköstä, mustikat luultavasti olisivat Ahkujärveltä peräisin.

Mallin suppeneminen

Suppenemistarkastelut saattavat tuntua hankalilta tottumattomasta lukijasta. Edes jonkinlaiseen täsmällisyyteen pyrkivän esityksen yksityiskohtien ymmärtäminen ei ole oleellista itse asian kannalta. Yritämme selvittää, milloin jonolla $P(A|V_n)$ on raja-arvo, ja mikä se on. Ehdollisten todennäköisyyksien $P(V_n|B)$ ja $P(V_n|A)$ suhde on

$$\frac{P(V_n|B)}{P(V_n|A)} = \frac{p_B^{m(n)}(1-p_B)^{n-m(n)}}{p_A^{m(n)}(1-p_A)^{n-m(n)}} = \left[\left(\frac{p_B}{p_A}\right)^{\frac{m(n)}{n}}\left(\frac{1-p_B}{1-p_A}\right)^{1-\frac{m(n)}{n}}\right]^n.$$

Lukumäärien $m(n)$ ja n suhde $m(n)/n$ on mustan värin *suhteellinen frekvenssi* jonossa V_n. Jos marjat ovat Ahkujärven mustikoita, *suurten lukujen lain* ns. *vahvan version* nojalla suhteellinen frekvenssi lähestyy rajaarvonaan todennäköisyyttä p_A siinä mielessä, että

$$P\left(\frac{m(n)}{n} \to p_A\right) = 1.$$

Tällöin sanotaan, että $m(n)/n$ *konvergoi melkein varmasti* (engl. *almost surely*) kohti lukua p_A, ja merkitään

$$\frac{m(n)}{n} \xrightarrow[\text{a.s.}]{} p_A.$$

On periaatteessa mahdollista, että kiposta sattumalta tulee poimituksi esim. koko ajan vain sinisiä marjoja, jolloin $m(n)/n = 0 \to 0 \neq p_A$, mutta tämänkaltainen tilanne on äärimmäisen harvinainen, sen todennäköisyys on 0. Seuraavassa tarkastelemme vain niitä värijonoja, joille $m(n)/n \to p_A$ tavanomaisessa mielessä, kun $n \to \infty$. Sellaisille jonoille johtamamme tulokset pätevät todennäköisyydellä 1 kaikkien värijonojen joukossa. Edellä oletimme marjojen sisältyvän Ahkujärven satoon. Bieggajängän mustikoiden tapauksessa tarkastelemme vastaavasti vain niitä värijonoja, joilla $m(n)/n \to p_B$. Voimme siis kirjoittaa

$$\lim_{n\to\infty}\left[\frac{P(V_n|B)}{P(V_n|A)}\right]^{\frac{1}{n}} = \begin{cases} \left(\frac{p_B}{p_A}\right)^{p_A}\left(\frac{1-p_B}{1-p_A}\right)^{1-p_A} & \text{tapauksessa } A, \\ \left(\frac{p_B}{p_A}\right)^{p_B}\left(\frac{1-p_B}{1-p_A}\right)^{1-p_B} & \text{tapauksessa } B. \end{cases} \quad (2)$$

Suhteen $P(V_n|B)/P(V_n|A)$ raja-arvon määrittämiseksi tarvitsemme pari aputulosta eli lemmaa:

Lemma 1. Olkoon $0 < x < 1$, $0 < y < 1$ ja $x \neq y$. Tällöin

(a) $\left(\frac{x}{y}\right)^y\left(\frac{1-x}{1-y}\right)^{1-y} < 1,$

(b) $\left(\frac{x}{y}\right)^x\left(\frac{1-x}{1-y}\right)^{1-x} > 1.$

Todistus: Jälkimmäinen väite (b) seuraa edellisestä käänteislukuihin siirtymällä, joten riittää todistaa (a). Se on yhtäpitävä epäyhtälön

$$f(x) = x^y(1-x)^{1-y} < y^y(1-y)^{1-y} \quad (3)$$

kanssa. Tässä $f(y) = y^y(1-y)^{1-y}$. Funktion f derivaatta muuttujan x suhteen on

$$\begin{aligned} f'(x) &= yx^{y-1}(1-x)^{1-y} + x^y(1-y)(1-x)^{-y}\cdot(-1) \\ &= \left(\frac{1-x}{x}\right)\left(\frac{1-x}{x}\right)^{-y}y + \left(\frac{1-x}{x}\right)^{-y}(y-1) \\ &= \left[\left(\frac{1-x}{x}+1\right)y - 1\right]\left(\frac{1-x}{x}\right)^{-y} \\ &= \left[\frac{y}{x} - 1\right]\left(\frac{1-x}{x}\right)^{-y}. \end{aligned}$$

Tässä tulossa jälkimmäinen tekijä on positiivinen, joten ensimmäinen tekijä määrää tulon etumerkin. Täten $f'(x) > 0$ eli f on aidosti kasvava, kun $0 < x < y$.

Vastaavasti $f'(x) < 0$ eli f on aidosti vähenevä, kun $y < x < 1$. □

Lemman perusteella siis

$$\left(\frac{p_B}{p_A}\right)^{p_A}\left(\frac{1-p_B}{1-p_A}\right)^{1-p_A} < 1,$$

ja

$$\left(\frac{p_B}{p_A}\right)^{p_B}\left(\frac{1-p_B}{1-p_A}\right)^{1-p_B} > 1.$$

Lemma 2. Olkoon $\lim_{n\to\infty} a_n = a$, missä $0 < a \neq 1$. Tällöin

$$\lim_{n\to\infty} a_n^n = \begin{cases} 0, & 0 < a < 1, \\ \infty, & a > 1. \end{cases}$$

Todistus: Perustamme todistuksen lukujonon rajaarvon määritelmään. Olkoon $\lambda = (a+1)/2$ lukujen a ja 1 keskiarvo. Tapauksessa $0 < a < 1$ on olemassa sellainen positiivinen kokonaisluku n_1, että $0 < a_n < \lambda < 1$ aina, kun $n > n_1$. Indeksin arvoilla $n > n_1$ pätee $0 < a_n^n < \lambda^n \to 0$.

Tapauksessa $a > 1$ on olemassa sellainen positiivinen kokonaisluku n_2, että $a_n > \lambda > 1$ aina, kun $n > n_2$. Indeksin arvoilla $n > n_2$ pätee $a_n^n > \lambda^n \to \infty$. □

Soveltamalla lemmaa raja-arvoyhtälöön (2) saamme yhtälön

$$\lim_{n\to\infty} \frac{P(V_n|B)}{P(V_n|A)} = \begin{cases} 0 & \text{tapauksessa } A, \\ \infty & \text{tapauksessa } B. \end{cases}$$

Lopuksi otamme taas kaikki värijonot mukaan, jolloin tavallisen suppenemisen tilalle astuu melkein varma suppeneminen. Kaavan (1) perusteella

$$P(A|V_n) = \frac{1}{1 + \dfrac{P(V_n|B)}{P(V_n|A)}\dfrac{P(B)}{P(A)}} \xrightarrow[\text{a.s.}]{} \begin{cases} 1 & \text{tapauksessa } A, \\ 0 & \text{tapauksessa } B. \end{cases} \quad (4)$$

Laskentakaava

Tarkoituksenamme on johtaa käytännöllinen *iteraatioon* eli toistoon perustuva kaava todennäköisyyksien $P(A|V_n)$ laskemiseksi. Olkoon F_n jonon V_n viimeinen alkio. Kun F_n poistetaan jonosta V_n, jäljelle jää jono V_{n-1}. Kaavasta (1) saadaan Bayesin kaavan avulla

$$\begin{aligned} &P(A|V_n) \\ &= \frac{P(V_n|A)P(A)}{P(V_n|A)P(A) + P(V_n|B)P(B)} \\ &= \frac{P(F_n|A)P(V_{n-1}|A)P(A)}{P(F_n|A)P(V_{n-1}|A)P(A) + P(F_n|B)P(V_{n-1}|B)P(B)} \\ &= \frac{P(F_n|A)\dfrac{P(V_{n-1}|A)P(A)}{P(V_{n-1})}}{P(F_n|A)\dfrac{P(V_{n-1}|A)P(A)}{P(V_{n-1})} + P(F_n|B)\dfrac{P(V_{n-1}|B)P(B)}{P(V_{n-1})}} \\ &= \frac{P(F_n|A)P(A|V_{n-1})}{P(F_n|A)P(A|V_{n-1}) + P(F_n|B)P(B|V_{n-1})}. \end{aligned}$$

Symmetriasyistä vastaavat yhtälöt pätevät myös vaihtamalla A ja B keskenään:

$$P(A|V_n) = \frac{P(F_n|A)P(A|V_{n-1})}{P(F_n|A)P(A|V_{n-1}) + P(F_n|B)P(B|V_{n-1})},$$

$$P(B|V_n) = \frac{P(F_n|B)P(B|V_{n-1})}{P(F_n|A)P(A|V_{n-1}) + P(F_n|B)P(B|V_{n-1})}.$$

Merkitsemme $P_0(A) = P(A)$, $P_0(B) = P(B)$, $P_n(A) = P(A|V_n)$ ja $P_n(B) = P(B|V_n)$, kun $n > 0$. Näillä merkinnöillä todennäköisyyksille saadaan helppokäyttöinen iteratiivinen laskentakaava

$$\begin{cases} P_0(A) = P(A) \\ P_0(B) = P(B) \end{cases}$$

$$\begin{cases} P_n(A) = \dfrac{P(F_n|A)P_{n-1}(A)}{P(F_n|A)P_{n-1}(A) + P(F_n|B)P_{n-1}(B)} \\ P_n(B) = \dfrac{P(F_n|B)P_{n-1}(B)}{P(F_n|A)P_{n-1}(A) + P(F_n|B)P_{n-1}(B)} \end{cases} \quad (5)$$

Alin yhtälö on mukana vain symmetrian havainnollistamisen takia. Laskennassa se kannattaa korvata yhtälöllä $P_n(B) = 1 - P_n(A)$. Jos jompikumpi todennäköisyyksistä $P_n(A)$ ja $P_n(B)$ on tasan 1, jolloin toinen on tasan 0, iteroinnin jatkaminen ei enää muuta todennäköisyyksiä. Verrattuna laskentakaavaan (1) iteraatio (5) on kätevä siinä mielessä, että koko havaintojonoa V_n ei tarvitse säilyttää muistissa, vaan ainoastaan jompikumpi edellisistä todennäköisyyksistä $P_{n-1}(A)$ ja $P_{n-1}(B)$. Toinen riittää, sillä $P_{n-1}(A) + P_{n-1}(B) = 1$. Tietysti $P(M|A)$, $P(M|B)$, $P(S|A)$ ja $P(S|B)$ tarvitaan myös, mutta ne säilyvät samoina koko laskennan ajan. Kun vaiheen $n-1$ jälkeen tulee uusi havainto (so. katsotaan seuraavan mustikan väri F_n), vaiheen $n-1$ *posterioritodennäköisyydet* $P_{n-1}(A)$ ja $P_{n-1}(B)$ otetaan vaiheen n *prioritodennäköisyyksiksi*, ja lasketaan uudet posterioritodennäköisyydet $P_n(A)$ ja $P_n(B)$. Kertoimia $P(F_n|A)$ ja $P(F_n|B)$ kutsutaan nimellä *uskottavuus* (engl. *likelihood*), ja nimittäjä $P(F_n|A)P_{n-1}(A) + P(F_n|B)P_{n-1}(B)$ on *normeerausvakio*. Viimeksi mainitun ainoana tehtävänä on varmistaa, että $P_n(A) + P_n(B) = 1$.

Esimerkki 1. Tervamustikoiden osuudet kaikista mustikoista ovat tavallisesti sen verran pieniä, että havainnollista esimerkkiä ajatellen musta mustikka toistuu aivan liian harvoin. Sen takia tässä simulaatiossa liioitellaan: Ahkujärven marjoista 60 % ja Bieggajängän marjoista 30 % oletetaan tervamustikoiksi. Sammeli keräsi marjat Ahkujärveltä, mutta Miihkali ei sitä etukäteen tiedä.

n	F_n	$P_n(A)$	$P_n(B)$	Kommentti
0		0.45	0.55	Arvattu priorijakauma
1	S	0.3185841	0.6814159	Sininen tukee Bieggajänkää
2	S	0.2108346	0.7891654	
3	M	0.3482467	0.6517533	Musta tukee Ahkujärveä
4	M	0.5165919	0.4834081	
5	M	0.6812537	0.3187463	
6	M	0.8104115	0.1895885	
7	M	0.8952788	0.1047212	
8	M	0.9447463	0.0552537	
9	S	0.9071536	0.0928464	Sininen tukee Bieggajänkää
10	M	0.9513168	0.0486832	Musta tukee Ahkujärveä
11	S	0.9178054	0.0821946	Sininen tukee Bieggajänkää
12	S	0.8645118	0.1354882	
13	S	0.7847669	0.2152331	
14	S	0.6756932	0.3243068	
15	M	0.8064641	0.1935359	Musta tukee Ahkujärveä
16	M	0.8928648	0.1071352	
17	S	0.8264577	0.1735423	Sininen tukee Bieggajänkää
18	S	0.7312771	0.2687229	
19	M	0.8447834	0.1552166	Musta tukee Ahkujärveä
20	M	0.9158619	0.0841381	

Aika hyvin asettuivat todennäköisyydet oikean vaihtoehdon eli Ahkujärven kannalle jo 20 ensimmäisellä iteraatiokierroksella. Ise asiassa paras tulos saatiin jo kierroksella 8, mutta sen jälkeen tuli sattumalta paljon sinisiä mustikoita, jotka houkuttelivat todennäköisyyttä Bieggajängän suuntaan. Suurten lukujen laki alkaa kuitenkin "vääjäämättä" (so. todennäköisyydellä 1) toimia jossain vaiheessa. Kierroksen 20 jälkeen Miihkali tietää yli 90 %:n varmuudella, että kyse on Ahkujärven mustikoista – vai tietääkö sittenkään? Viimeiseltä riviltä nähdään todennäköisyys sille, että marjat on poimittu Ahkujärveltä ehdolla, että Sammelin suorittamassa "jälkipoiminnassa" on saatu värisarakkeen mukainen mustikkajono. Todennäköisyydet ovat päteviä vain, mikäli ensimmäisen rivin priorijakauma on todellisen tilanteen mukainen. Priorijakauma tarkoittaa, että 45 % Sammelin poimintaretkistä suuntautuu Ahkujärvelle ja loput 55 % Bieggajängälle. Vaikka priorijakauma olisi arvattu täysin pieleen, iteraation jossain vaiheessa alkaa selvästi näkyä, kummasta paikasta marjat ovat peräisin. Väärästä priorijakaumasta lähteneen iteraation antamat ehdolliset välitodennäköisyydet ovat enemmän tai vähemmän roskaa, mutta tavoitteenahan onkin lopputulos, siis kumman vaihtoehdon kohdalle todennäköisyys lopulta kasaantuu.

Yleistykset

Edellä esitetty oli kenties yksinkertaisin mahdollinen esimerkki *tilastollisesta inversiosta*. Se yleistyy ilmeisellä tavalla tilanteisiin, joissa mustikkapaikkoja on K kpl, siis esim. $A_1, \ldots, A_K$. Nyt kahden todennäköisyyden $P_n(A)$ ja $P_n(B)$ muodostaman jakauman paikalla on todennäköisyyksistä $P_n(A_1), \ldots, P_n(A_K)$ muodostuva jakauma. Malliin on myös helppoa ottaa useampi väri. Iteraatio (5) ei muutu paljon:

$$\begin{cases} P_0(A_1) = P(A_1) \\ \quad\vdots \\ P_0(A_K) = P(A_K) \end{cases}$$

$$\begin{cases} P_n(A_1) = \dfrac{P(F_n|A_1)P_{n-1}(A_1)}{\sum_{k=1}^{K} P(F_n|A_k)P_{n-1}(A_k)} \\ \quad\vdots \\ P_n(A_K) = \dfrac{P(F_n|A_K)P_{n-1}(A_K)}{\sum_{k=1}^{K} P(F_n|A_k)P_{n-1}(A_k)} \end{cases}$$

Värien lisääntyminen merkitsee vain muuttujan F_n mahdollisten arvojen lisääntymistä. Tällöin pitää tietää myös näiden arvojen todennäköisyydet jokaisen marjapaikan A_k kohdalla. Kun marjastusesimerkistä irrottaudutaan ja etäännytään, saadaan yleistykset kaikenlotteisille diskreeteille ja jatkuville jakaumille. Nämä yleistykset toimivat yötäpäivää lukemattomien arkikäytössä olevien laitteiden prosessoreissa, mutta näistä asioista kertomisen jätän viisaammille.