Insinöörimatematiikka: Todennäköisyyslaskenta

Demonstraatio 4, 8.3.2024

Jos ei toisin mainita, tehtävissä entropia määritetään käyttäen bittiä informaation yksikkönä.

1. Satunnaismuuttuja X saa arvot 1, 2, 3, 4 ja 5 todennäköisyyksillä $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{5}$, $\frac{1}{6}$ ja $\frac{1}{20}$, Laske entropia $H(X)$.

   Mallivastaus:

$$\begin{aligned} H(X) &= -\frac{1}{3}\log_2\frac{1}{3} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{5}\log_2\frac{1}{5} - \frac{1}{6}\log_2\frac{1}{6} - \frac{1}{20}\log_2\frac{1}{20} \\ &= 2,13963\ldots \end{aligned}$$

2. Satunnaismuuttujien X ja Y mahdolliset arvot ovat joukossa $\{1,2,3,4\}$ ja näillä on seuraavan taulukon mukainen yhteisjakauma:

Y \ X	1	2	3	4
1	0.06	0.02	0.06	0.04
2	0.02	0.08	0.06	0.10
3	0.08	0.04	0.08	0.08
4	0.10	0.08	0.08	0.02

   Taulukko tulkitaan ilmeisellä tavalla: esim. $\mathbb{P}(X=1,Y=2)=0.02$ ja $\mathbb{P}(X=3,Y=1)=0.06$. Voidaan merkitä myös $\mathbb{P}(X=i,Y=j)=\mathbb{P}(i,j)$.

   Määritä todennäköisyydet $\mathbb{P}(X=i)$ ja $\mathbb{P}(Y=i)$, kun $i\in\{1,2,3,4\}$. Ovatko satunnaismuuttujat X ja Y riippumattomat? Ohje: Esim. Todennäköisyys sille, että Y saa arvon 1 lasketaan 1. rivin todennäköisyyksien summana (miksi?)

   Mallivastaus: Todennäköisyys $\mathbb{P}(X=1)$ saadaan summaamalla 1. sarakkeen todennäköisyydet yhteen (kokonaistodennäköisyys): $\mathbb{P}(X=1)=0.06+0.02+0.08+0.10=0.26$. Samoin $\mathbb{P}(X=2)=0.22$, $\mathbb{P}(X=3)=0.28$ ja $\mathbb{P}(X=4)=0.24$ Rivien todennäköisyydet summaamalla saadaan $\mathbb{P}(Y=1)=0.18$, $\mathbb{P}(Y=2)=0.26$, $\mathbb{P}(Y=3)=0.28$ ja $\mathbb{P}(Y=4)=0.28$

   Satunnaismuuttujat ovat riippuvat, koska esimerkiksi $\mathbb{P}(X=1)\cdot\mathbb{P}(Y=1)=0.26\cdot 0.18=0.0468$, mikä on erisuuri kuin $\mathbb{P}(X=1,Y=1)$

3. Laske edellisen tehtävän satunnaismuuttujille entropiat $H(X)$, $H(Y)$, $H(X\mid Y)$ sekä $I(X\mid Y)$. Ohje: Ehdollinen entropia voidaan laskea kaavalla $H(X\mid Y)=H(X,Y)-H(Y)$

   Mallivastaus:

   $H(X)=-(0.26\log_2 0.26+0.22\log_2 0.22+0.28\log_2 0.28+0.24\log_2 0.24)=1.99422\ldots,$

   $H(Y)=-(0.18\log_2 0.18+0.26\log_2 0.26+0.28\log_2 0.28+0.28\log_2 0.28)=1.97904\ldots$

   Mieti: Miksi nämä entropiat ovat lähellä lukua 2?

   Ohjetta noudattaen laskentaan yhteisentropia:

   $H(X,Y)=-(0.06\log_2 0.06+0.02\log_2 0.02+0.06+\ldots+0.02\log_2 0.02)=3.85418\ldots$

   Tästä saadaan ehdollinen entropia $H(X\mid Y)=H(X,Y)-H(Y)=1.87514\ldots$ ja lopulta Informaatio $I(X\mid Y)=H(X)-H(X\mid Y)=0.11908\ldots$.

4. Jääkiekkojoukkue voittaa kotona pelatessaan todennäköisyydellä $\frac{4}{5}$, mutta vierasjoukkueena pelatessaan vain todennäköisyydellä $\frac{1}{4}$. Pelipaikka on todennäköisyydellä $\frac{1}{3}$ kotihalli ja $\frac{2}{3}$ vieras halli.

   Millä todennäköisyydellä joukkue voittaa pelin? Olkoon X satunnaismuuttuja, joka saa arvot $\{\text{voitto}, \text{häviö}\}$ ja Y satunnaismuuttuja, joka saa arvot $\{\text{kotihalli}, \text{vierashalli}\}$. Laske $I(X \mid Y)$.

   Mallivastaus: Kokonaistodennäköisyyden perusteella laskettu voittotodennäköisyys on

$$\frac{4}{5} \cdot \frac{1}{3} + \frac{1}{4} \cdot \frac{2}{3} = \frac{13}{30} = 0.4333\ldots.$$

   Tällöin $H(X) = -(0.4333 \log_2 0.4333 + 0.5667 \log_2 0.5667) = 0.987123\ldots$ Kysyttyä informaatio voidaan laskea, jos tunnetaan ehdollinen entropia. Lasketaan tätä varten

$$H(X \mid \text{kotihalli}) = -(\frac{4}{5} \log_2 \frac{4}{5} + \frac{1}{5} \log_2 \frac{1}{5}) = 0.464386\ldots.$$

   ja

$$H(X \mid \text{vierashalli}) = -(\frac{1}{4} \log_2 \frac{1}{4} + \frac{3}{4} \log_2 \frac{3}{4}) = 0.811278\ldots.$$

   Ehdollinen entropia saadaan tällöin odotusarvona

$$H(X \mid Y) = \frac{1}{3} H(X \mid \text{kotihalli}) + \frac{2}{3} H(X \mid \text{vierashalli}) = 0.695647\ldots$$

   ja kysytty informaatio

$$I(X \mid Y) = H(X) - H(X \mid Y) = 0.291476\ldots.$$

5. Satunnaismuuttuja X saa arvokseen kirjaimia A, B, …, G taulukon

A	B	C	D	E	F	G
$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$

   mukaisilla todennäköisyyksillä. Satunnaismuuttujan arvoista muodostuva jono koodataan bittijonoksi taulukon

A	B	C	D	E	F	G
0	110	111	1000	1001	1010	1011

   mukaisesti. Määritä entropia $H(X)$ ja selvitä kuinka pitkä bittijono keskimäärin koodaa yhden kirjaimen.

   Voidaanko koodaus bittijonoksi tehdä paremmin, siis siten, että koodaavan bittijonon pituus olisi keskimäärin pienempi? Selvitä lopuksi mitkä satunnaismuuttujan arvot ovat koodattuna bittijonoon 11110010110100110101011111100110110. Mistä johtuu että koodattu kirjainjono löytyy yksikäsitteisesti?

   Mallivastaus:

$$H(X) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{8} \log_2 \frac{1}{8} \ldots - \frac{1}{16} \log_2 \frac{1}{16} = \frac{9}{4} = 2.25.$$

   Keskimääräisen bittijonon pituus yhden kirjaimen koodaukselle on

$$\frac{1}{2} \cdot 1 + 0 \cdot 2 + 2 \cdot \frac{1}{8} \cdot 3 + 4 \cdot \frac{1}{16} \cdot 4 = \frac{9}{4}.$$

Koska tässä koodauksessa keskimääräinen pituus on sama kuin entropia, voidaan olla varmoja että tehokkaampaa koodausta ei ole.

Bittijono dekoodautuu jonoksi $CEABEFGCAABB$. YKsikäsitteinen dekoodaus johtuu siitä, että yksikään binäärisistä koodijonoista ei ole toisen prefiksi (etuliite).

6. Olkoon X satunnaismuuttuja, joka saa arvot $\{0,1\}$ yhtä todennäköisesti. Y on satunnaismuuttuja, joka saadaan, kun X:n arvo lähetetään sellaisen kanavan läpi, jossa 0 muuttuu arvoksi 1 todennäköisyydellä $\epsilon \in [0,1]$ 1 muuttuu arvoksi 0 todennäköisyydellä ϵ. Määritä $H(Y)$, $H(X,Y)$, $H(X \mid Y)$ ja $I(X \mid Y)$.

   Ohje: Entropiaa $H(X,Y)$ varten määritä ensin todennäköisyydet, joilla arvot $(x,y) \in \{(0,0),(0,1),(1,0),(1,1)\}$ esiintyvät. $H(X \mid Y) = H(X,Y) - H(Y)$.

   Mallivastaus: $\mathbb{P}(Y=0) = (1-\epsilon)\frac{1}{2} + \epsilon\frac{1}{2} = \frac{1}{2}$ ja samoin $\mathbb{P}(Y=1) = \frac{1}{2}$. Näin ollen $H(Y) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}$.

   $\mathbb{P}(X=0,Y=0) = \frac{1}{2}\cdot(1-\epsilon)$, $\mathbb{P}(X=0,Y=1) = \frac{1}{2}\cdot\epsilon$, $\mathbb{P}(X=1,Y=0) = \frac{1}{2}\cdot\epsilon$ ja $\mathbb{P}(X=1,Y=1) = \frac{1}{2}\cdot(1-\epsilon)$. Näiden todennäköisyyksien perusteella

$$\begin{aligned} H(X,Y) &= -2\cdot\frac{1}{2}(1-\epsilon)\log_2\frac{1}{2}(1-\epsilon) - 2\cdot\frac{\epsilon}{2}\log_2\frac{\epsilon}{2} \\ &= 1-(1-\epsilon)\log_2(1-\epsilon) - \epsilon\log_2(\epsilon). \end{aligned}$$

   Näiden perusteella edelleen

$$\begin{aligned} I(X \mid Y) &= H(X) - H(X \mid Y) = H(X) - (H(X,Y) - H(Y)) = H(X) + H(Y) - H(X,Y) \\ &= 1 + \epsilon\log_2\epsilon + (1-\epsilon)\log_2(1-\epsilon) \end{aligned}$$

7. Taulukossa

172	168	167	192	184	165	163	177	188	186	180	182	179	175	181	183

   on eräästä ryhmästä umpimähkään valittujen 16 henkilön pituus senttimetreinä. Taulukon voi ladata Excel-muodossa osoitteesta

   https://users.utu.fi/mikhirve/ins2324/TNlaskenta/Pituudet.xlsx. Excel-muotoisen tiedoston voi siirtää esim. copy-paste toiminnolla Matlabin Workspace-tilaan.

   Laske pituuksien otoskeskiarvo ja otosvarianssi. Ohje Matlabia varten: Jos `a=[172,168,..., 183]` on vektori, jonka koordinaatteina ovat pituudet, saadaan koordinaattien summa seuraavasti: `sum(a)`. Sellainen vektori, jonka koordinaatteina ovat luvut $(172-\mu)^2$, $(168-\mu)^2$, ... saadaan syötteellä `(a-`μ`).^2` (huomaa piste ennen symbolia ^). Toisin: Matlabissa on myös valmiit tilastotoiminnot, ohje on lähinnä niitä varten jotka tahtovat perehtyä asiaan kunnolla.

   Mallivastaus: Otoskeskiarvo mitatuilla arvoilla saadaan taulukossa esiintyvien pituuksien aritmeettisena keskiarvona:

$$\mu = \overline{X}_{16} = \frac{1}{16}(172+168+\ldots+183) = 177.6250$$

   Otosvarianssi puolestaan saadaan summana

$$S^2_{16} = \frac{1}{16-1}((172-\mu)^2 + (168-\mu)^2 + \ldots + (183-\mu^2)) = 73.9833$$

8. Oletetaan, että ryhmässä pituudet noudattavat normaalijakaumaa. Määritä sellainen reaalilukuväli, jonka keskelle edellisessä tehtävässä laskemasi otoskeskiarvo kuuluu vähintään todennäköisyydellä 90%. Ohje: Luento 31.1. Matlabissa Studentin t-jakauman kertymäfunktio syötetään muodossa `tcdf(x,n)`, sen käänteisfunktio `tinv(x,n)` ja tiheysfunktio muodossa `tpdf(x,n)`. Näissä n on otokseen kokoon liittyvä parametri.

Mallivastaus: Aluksi on etsittävä sellainen a, että Studentin $t(15)$-jakaumassa

$$\mathbb{P}(-a \leq T \leq a) = 0.9.$$

Koska Studentin jakauma on origokeskinen ja symmetrinen, tämä on saadaan selvittämällä millä a:n arvolla

$$\mathbb{P}(T \leq a) = 0.95,$$

ja koska vasen puoli on sama kuin Studentin jakauman $t(15)$ kertymäfunktio arvolla a, saadaan a tämän käänteisfunktiona. Matlabissa: `a=tinv(0.95,15)` $= 1.7531$ (tätä merkitään tyypillisesti $a = t_{15;0.05}$).

Luentoesimerkin mukaan odotusarvo on välillä

$$[\overline{X}_{16} - a\frac{S_{16}}{\sqrt{16}}, \overline{X}_{16} + a\frac{S_{16}}{\sqrt{16}}] = [173.8552, 181.3948].$$

Huomautus: Luottamusväli on varsin pitkä, keskipituus 177.6250 cm voidaan siis määrittää vain n. ± 3.5 cm tarkkuudella kun määritys on oikein (vähintään) 90 % todennäköisyydellä. Suuri luottamusväli johtuu siitä, että havaintojen (mittausten) määrä on vähäinen.