Insinöörimatematiikka: Todennäköisyyslaskenta

Mika Hirvensalo
mikhirve@utu.fi

Matematiikan ja tilastotieteen laitos
Turun yliopisto

2024

Entropia

Määritelmä

Lähtökohtana on diskreetti satunnaismuuttuja X, joka voi saada n arvoa $x_1, \ldots, x_n$ todennäköisyyksillä $p_1, \ldots, p_n$.
Jakauman $p_1, \ldots, p_n$ *Shannon-entropia* on

$$H(p_1, \ldots, p_n) = -K(p_1 \ln p_1 + \ldots + p_n \ln p_n),$$

missä $K > 0$ on vakio. Jos $K = \frac{1}{\ln 2}$, sanotaan entropiaa binääriseksi. Tällöin $\frac{\ln p_i}{\ln 2} = \log_2 p_i$. Määritellään lisäksi $0 \cdot \ln 0 = 0$.
Jos diskreetillä satunnaismuuttujalla X on jakauma $p_1, \ldots, p_n$, merkitään ylläolevaa entropiaa myös $H(X)$:llä.

Entropia

Esimerkki

Jos kaikki arvot x_i ilmaantuvat samalla todennäköisyydellä $\frac{1}{n}$, on

$$H(\frac{1}{n}, \ldots, \frac{1}{n}) = -(\frac{1}{n} \log_2 \frac{1}{n} + \ldots + \frac{1}{n} \log_2 \frac{1}{n}) = \log_2 n.$$

Tämä merkitsee sitä, että satunnaismuuttujan X esittämän viestin välittämiseen tarvitaan keskimäärin $\log_2 n$ bittiä.

Esimerkki

Jos satunnaismuuttuja X saa arvon x_1 todennäköisyydellä 1 ja muut arvot $x_2, \ldots, x_n$ todennäköisyydellä 0, on

$$H(X) = -(1 \cdot \log_2 1 + 0 \cdot \log_2 0 + \ldots + 0 \cdot \log_2 0) = 0$$

Entropia

Esimerkki

Satunnaismuuttuja X saa kaksi arvoa x_0 ja x_1, molemmat todennäköisyydellä $\frac{1}{2}$. Tällöin

$$H(X) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) = 1.$$

Esimerkki

Satunnaismuuttuja X saa arvon x_0 todennäköisyydellä $\frac{1}{4}$ ja arvon x_1 todennäköisyydellä $\frac{3}{4}$. Tällöin

$$H(X) = -(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}) = 0,811278\ldots$$

Entropia

Lause

Oletetaan että reaaliarvoinen funktio $H(p_1, \ldots, p_n)$ toteuttaa seuraavat ehdot:

- $H(p_1, \ldots, p_n)$ on symmetrinen ja jatkuva
- $H(\frac{1}{n}, \ldots, \frac{1}{n})$ on ei-negatiivinen, aidosti kasvava
- Todennäköisyysjakaumille $(p_1, \ldots, p_n)$ ja $(q_1, \ldots, q_m)$ on voimassa

$$\begin{aligned} & H(p_1, \ldots, p_n) + p_n H(q_1, \ldots, q_m) \\ = \quad & H(p_1, \ldots, p_{n-1}, p_n q_1, \ldots, p_n q_m). \end{aligned}$$

Tällöin

$$H(p_1, \ldots, p_n) = -K(p_1 \ln p_1 + \ldots + p_n \ln p_n),$$

missä $K > 0$ on vakio.

Entropia

Yhdistetty entropia

$$H(X, Y) = -\sum p(x, y) \log_2 p(x, y)$$

Ehdollinen entropia

$$H(X \mid y_j) = -\sum p(x \mid y_j) \log_2 p(x \mid y_j)$$

$$H(X \mid Y) = \sum p(y_j) H(X \mid y_j)$$

Informaatio

$$I(X \mid Y) = H(X) - H(X \mid Y)$$

Määritelmä

Mikäli n-pituinen jono satunnaismuuttujan X arvoista muodostuvia viestejä voidaan koodata $m = \lfloor rn \rfloor$-pituisilla bittijonoilla, sanotaan, että koodaus onnistuu tahdilla (rate) r.

Shannonin lause

Jos $r > H(X)$, on mahdollista koodata X:n arvot binääriseen aakkostoon tahdilla r siten että dekoodausvirheen todennäköisyys lähenee nollaa.
Jos $r < H(X)$, edellämainittu koodaus ei ole mahdollista, vaan dekoodausvirheen todennäköisyys lähenee ykköstä.

Shannonin lause

Määritelmä

Binäärisen symmetrisen kanavan *kapasiteetti* on

$$C(p) = 1 - H_2(p),$$

missä p on kanavan virhetodennäköisyys.

Koodin C_n *informaatiosuhde* on

$$R(C_n) = \frac{\log_2 |C_n|}{n}.$$

Shannonin lause 2

Jos $p < \frac{1}{2}$ ja $R < C(p)$ ja $\epsilon > 0$, niin on sellainen rajaluku $N = N(p, R, \epsilon)$, että aina kun $n \geq N$, niin on olemassa n-pituinen binäärikoodi C_n jolle $R(C_n) \geq R$ ja $\mathbb{P}_{er}(C_n) < \epsilon$.

Estimointi

Lähtökohdat

- Havainnot $x_1, \ldots x_n$ jostakin ilmiöstä.
- Tyypillisesti ei tiedetä miten havainnoitava ilmiö on jakautunut.
- i:s havainto x_i on satunnaismuuttujan X_i arvo.
- Satunnaismuuttujat X_i riippumattomat.
- Satunnaisotos: X_1, X_2, ..., X_n.
- Pyrkimys: Hyvät arviot ainakin odotusarvolle $\mathbb{E}(X_i) = \mu$ ja varianssille $\mathrm{Var}(X_i) = \sigma^2$.

Estimointi

Odotusarvo

Otoskeskiarvo

$$\overline{X_n} = \frac{1}{n}(X_1 + \ldots + X_n).$$

Tällöin

$$\mathbb{E}(\overline{X_n}) = \frac{1}{n} \cdot (\mathbb{E}(X_1) + \ldots + \mathbb{E}(X_n)) = \mu.$$

Tätä ominaisuutta sanotaan *harhattomuudeksi*

Estimointi

Tarkentuvuus

$$\begin{aligned}
\mathrm{Var}(\overline{X_n}) &= \mathrm{Var}(\frac{1}{n}X_1) + \ldots + \mathrm{Var}(\frac{1}{n}X_n) \\
&= \frac{1}{n^2}\,\mathrm{Var}(X_1) + \ldots + \frac{1}{n^2}\,\mathrm{Var}(X_n) \\
&= n \cdot \frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n} \xrightarrow{n\to\infty} 0.
\end{aligned}$$

Tällöin sanotaan, että $\overline{X_n}$ on *tarkentuva* estimaattori.

Varianssi

Koska

$$\sigma^2 = \mathrm{Var}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = \mathbb{E}(X_i^2) - \mu^2,$$

on

$$\mathbb{E}(X_i^2) = \sigma^2 + \mu^2$$

Samoin

$$\frac{\sigma^2}{n} = \mathrm{Var}(\overline{X_n}) = \mathbb{E}(\overline{X_n}^2) - \mathbb{E}(\overline{X_n})^2 = \mathbb{E}(\overline{X_n}^2) - \mu^2,$$

josta

$$\mathbb{E}(\overline{X_n}^2) = \frac{\sigma^2}{n} + \mu^2.$$

Varianssi

$$\mathbb{E}(\sum_{k=1}^{n}(X_k - \overline{X_n})^2) = (n-1)\sigma^2.$$

Tämän vuoksi määritellään *otosvarianssi*

$$S_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X_n})^2,$$

jolloin

$$\mathbb{E}(S_n^2) = \sigma^2.$$

S_n^2 on siis harhaton. Myös tarkentuva: $\mathrm{Var}(S_n^2) = \frac{\mu_4}{n} + \frac{3-n}{n(n-1)}\sigma^4$.

Luottamusvälit

Lause

Jos satunnaismuuttujat $X_1, \ldots, X_n$ ovat jakautuneet $N(\mu, \sigma^2)$ mukaisesti, on

$$T = \frac{\overline{X_n} - \mu}{S_n/\sqrt{n}}$$

jakautunut jakauman $t(n-1)$ (Student's t-distribution) mukaisesti.

Seuraus

Olkoon $\alpha \in (0,1)$ ja a sellainen, että $\mathbb{P}(-a < T < a) = 1 - \alpha$. Tällöin

$$\mathbb{P}(\overline{X_n} - a\frac{S_n}{\sqrt{n}} \leq \mu \leq \overline{X_n} + a\frac{S_n}{\sqrt{n}}) = 1 - \alpha.$$

Jos esim. $\alpha = 0,05$, on kyseessä 95% luottamusväli. Huom. Tyypillisesti merkitään $a = t_{n-1;\alpha/2}$.