

Insinöörimatematiikka: Todennäköisyyslaskenta

Mika Hirvensalo
mikhirve@utu.fi

Matematiikan ja tilastotieteen laitos
Turun yliopisto

2026

Lähtökohdat

- Havainnot $x_1, \dots, x_n$ 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: $X_1, X_2, \dots, X_n$.
- Pyrkimys: Hyvät arviot ainakin odotusarvolle $\mathbb{E}(X_i) = \mu$ ja varianssille $\text{Var}(X_i) = \sigma^2$.

Otoskeskiarvo

Oletetaan, että satunnaismuuttujat X_i ovat jakautuneet samoin, odotusarvo μ ja varianssi σ^2 . Tällöin otoskeskiarvo

$$\overline{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$$

on odotusarvon μ harhaton ja tarkentuva estimaattori.

Tämä tarkoittaa, että $\mathbb{E}(\overline{X}_n) = \mu$ ja $\text{Var}(\overline{X}_n) \xrightarrow{n \rightarrow \infty} 0$.

Otosvarianssi

Oletetaan, että satunnaismuuttujat X_i ovat jakautuneet samoin, odotusarvo μ ja varianssi σ^2 . Tällöin otosvarianssi

$$S_n^2 = \frac{1}{n-1} \left((X_1 - \bar{X}_n)^2 + \dots + (X_n - \bar{X}_n)^2 \right)$$

on varianssin σ^2 harhaton ja tarkentuva estimaattori.

Tämä tarkoittaa, että $\mathbb{E}(S_n^2) = \sigma^2$ ja $\text{Var}(S_n^2) \xrightarrow{n \rightarrow \infty} 0$.

Lause

Jos satunnaismuuttujat $X_1, \dots, X_n$ 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

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

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

Lause

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

$$W = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$$

jakautunut jakauman $\chi^2(n-1)$ mukaisesti.

Lähtökohta

- Havainnot $X_1, \dots, X_n$.
- Oletus: $X_i \sim N(\mu, \sigma^2)$.
- $\overline{X}_n$ ja S_n havainnoista.
- Onko $\mu = \mu_0$ (nollahypoteesi) vai $\mu \neq \mu_0$ (vastahypoteesi)?
- Varmuutta ei voi yleensä saavuttaa: p -arvo
- p -arvo: Todennäköisyys sille, että nollahypoteesi pitää paikkansa.

“Elementary” entropy

$H(n)$ = number of elementary units (bits, trits, etc.) to encode n (uniform) conditions.

- $\{1, 2\} \mapsto \{0, 1\}$,
- $\{1, 2, 3\} \mapsto \{0, 11, 10\}$,
- $\{1, 2, 3, 4\} \mapsto \{00, 01, 10, 11\}$, etc.

Binary (elementary) entropy

$$H_2(n) = \log_2 n = \frac{1}{\log 2} \log n$$

Entropy measures uncertainty

Ternary (elementary) entropy

$$H_3(n) = \frac{1}{\log 3} \log n$$

q -ary (elementary) entropy

$$H_q(n) = \frac{1}{\log q} \log n$$

Boltzmann:

Identical particles with same internal condition are indistinguishable.

- Let l be the number of particles, each having n potential conditions $\Sigma = \{1, 2, \dots, n\}$, $l \gg n$.
- List the conditions of all particles: $c_1 c_2 \dots c_l \in \{1, 2, \dots, n\}^l$
- Assume condition (letter) i occurs k_i times, so $k_1 + \dots + k_n = l$ and $p_i = \frac{k_i}{l}$ is the probability (frequency) of condition i

Combinatorics:

There are

$$\frac{I!}{k_1! \dots k_n!}$$

such lists (strings of conditions)

Elementary entropy:

$$K \log \frac{I!}{k_1! \dots k_n!}$$

Per particle:

$$\frac{K}{I} \log \frac{I!}{k_1! \dots k_n!}$$

Stirling: $\log k! = k \log k - k + O(\log k)$, so

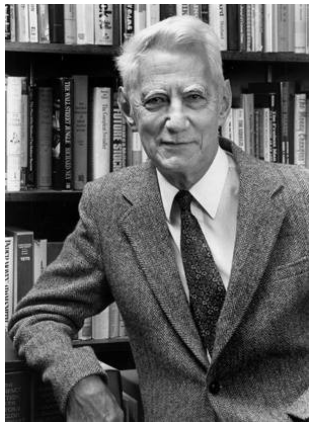
$$\begin{aligned} & \frac{K}{I} \log \frac{I!}{k_1! \dots k_n!} \\ &= \frac{K}{I} (I \log I - I + O(\log I)) \\ & \quad - \sum_{i=1}^n (k_i \log k_i - k_i + O(\log k_i)) \\ &= -K \sum_{i=1}^n p_i \log p_i + O\left(\frac{\log I}{I}\right). \end{aligned}$$

For a probability distribution $(p_1, \dots, p_n)$ of events $\{e_1, \dots, e_n\}$, define

$$H(p_1, \dots, p_n) = -K \sum_{i=1}^n p_i \log p_i$$

For $p_1 = \dots = p_n = \frac{1}{n}$

$$H\left(\frac{1}{n}, \dots, \frac{1}{n}\right) = -K \cdot n \cdot \frac{1}{n} \log \frac{1}{n} = K \log n$$



Claude Shannon (1916-2001)

- $H(p_1, \dots, p_n)$ symmetric, continuous
 - $H(\frac{1}{n}, \dots, \frac{1}{n})$ non-negative, strictly increasing in n
 - $H(p_1, \dots, p_n) + p_n H(q_1, \dots, q_m) = H(p_1, \dots, p_{n-1}, p_n q_1, \dots, p_n q_m)$
- $$\Rightarrow H(p_1, \dots, p_m) = -K \sum_{i=1}^n p_i \log p_i$$

Esimerkki

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

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

Esimerkki

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

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

Joint Entropy

Let

$$X = \{x_1, \dots, x_n\}$$

be a random variable with distribution $p(x_1), \dots, p(x_n)$.

Then

$$H(X) = - \sum_{i=1}^n p(x_i) \log p(x_i).$$

If also $Y = \{y_1, \dots, y_m\}$ is a random variable,

the *joint entropy* is

$$H(X, Y) = - \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log p(x_i, y_j)$$

Conditional Entropy

Joint entropy

$$H(X, Y) = - \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log p(x_i, y_j)$$

If we know $Y = y_j$, then

$$H(X | y_j) = - \sum_{i=1}^n p(x_i | y_j) \log p(x_i | y_j)$$

and define

$$H(X | Y) = \sum_{j=1}^m p(y_j) H(X | y_j).$$

“Uncertainty of X when Y is known”

Lemma

$$H(X | Y) \leq H(X)$$

Lemma

$$H(X, Y) \leq H(X) + H(Y)$$

Lemma

$$H(X | Y) = H(X, Y) - H(Y)$$

(Uncertainty of X when Y is known)

Example

Team *A* wins with probability $\frac{1}{2}$, $X = \{\text{win}, \text{loss}\}$. Then $H(X) = -(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2}) = 1$.

Example

As a *home team*, *A* wins with $\frac{3}{4}$ probability, but as *visitor*, *A* wins only with $\frac{1}{3}$ probability.

$$H(X | h) = -\left(\frac{3}{4} \log_2 \frac{3}{4} + \frac{1}{4} \log_2 \frac{1}{4}\right) = 0.811278 \dots,$$

$$H(X | v) = -\left(\frac{1}{3} \log_2 \frac{1}{3} + \frac{2}{3} \log_2 \frac{2}{3}\right) = 0.918296 \dots$$

Example (Continued)

$$H(X | h) = -\left(\frac{3}{4} \log_2 \frac{3}{4} + \frac{1}{4} \log_2 \frac{1}{4}\right) = 0.811278 \dots,$$

$$H(X | v) = -\left(\frac{1}{3} \log_2 \frac{1}{3} + \frac{2}{3} \log_2 \frac{2}{3}\right) = 0.918296 \dots$$

Let $Y = \{0, 1\}$ be a fair coin toss for deciding if team A plays home. Then

$$H(X | Y) = \frac{1}{2}H(X | h) + \frac{1}{2}H(X | v) = 0.864787 \dots$$

Definition (Mutual information of X and Y)

$$I(X : Y) = H(X) - H(X | Y)$$

$$\begin{aligned} I(X : Y) &= H(X) - H(X | Y) \\ &= H(X) - (H(X, Y) - H(Y)) \\ &= H(X) + H(Y) - H(X, Y) \\ &= I(Y : X) \end{aligned}$$

$$I(X : Y)$$

“Uncertainty of X minus uncertainty of X when Y known”

Example

X is the team A result, Y is the coin toss outcome. Then

$$I(X : Y) = 1 - 0.864787 \dots = 0.135213 \dots$$

Example

Team B wins with $1/2$ probability, but with 99 probability as home team and only with 5 probability as visitor. Then

$$H(X | h) = -\left(\frac{99}{100} \log_2 \frac{99}{100} + \frac{1}{100} \log_2 \frac{1}{100}\right) = 0.0807931 \dots,$$

$$H(X | v) = -\left(\frac{5}{100} \log_2 \frac{5}{100} + \frac{95}{100} \log_2 \frac{95}{100}\right) = 0.286397 \dots$$

Example (Continued)

$$H(X | Y) = \frac{1}{2}H(X | h) + \frac{1}{2}H(X | v) = 0.274894 \dots$$

and

$$I(X : Y) = 1 - 0.274894 \dots = 0.725106 \dots$$

Määritelmä

Mikäli n -pituisen 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 dekodeausvirheen todennäköisyys lähenee nollaa.

Jos $r < H(X)$, edellämainittu koodaus ei ole mahdollista, vaan dekodeausvirheen todennäköisyys lähenee ykköstä.

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 $\mathbb{P}_{er}(C_n) < \epsilon$.