T-61.5020 Luonnollisen kielen tilastollinen käsittely

Ratkaisut 1, ke 24.1.2007, 12:15–14:00 — Todennäköisyyslaskennan perusteita
Versio 1.0

1. Todennäköisyyksistä ensimmäinen $P(\text{ sana=lyhenne} \mid \text{sana=kolmikirjaiminen}) = 0.8$ kertoo, että jos me näemme kolmikirjaimisen sanan, se on todennäköisyydellä 0.8 lyhenne ja todennäköisyydellä 0.2 jotain muuta.

   Toinen kaava $P(\text{ sana=kolmikirjaiminen }) = 0.0003$ kertoo, että satunnainen sana on todennäköisyydellä 0.0003 kolmikirjaiminen ja todennäköisyydellä 0.9997 jotain muuta.

   Todennäköisyys, että satunnainen sana on kolmikirjaiminen lyhenne saadaan kertomalla edellä annetut todennäköisyydet keskenään. Eli ensin katsotaan, kuinka todennäköistä on, että sana on kolmikirjaiminen ja sitten vielä kuinka todennäköistä on, että kolmikirjaiminen sana olisi lyhenne:

   $$\begin{aligned} & P(\text{ sana=lyhenne}, \text{sana=kolmikirjaiminen }) \\ = \; & P(\text{ sana=kolmikirjaiminen }) \cdot P(\text{ sana=lyhenne} \mid \text{sana=kolmikirj. }) \\ = \; & 0.0003 * 0.8 = 0.00024 \end{aligned}$$

   Sivuhuomautuksena sanottakoon, että annetut todennäköisyydet eivät varmaankan päde todelliselle englannin kielelle.

2. Merkitään kantamuotoa *"se"* C_1:llä ja kantamuotoa *"siittää"* C_2:lla. Tunnistustulos olkoon T ja oikea luokka O. Kirjoitetaan tehtävässä annetut todennäköisyydet:

   $$\begin{aligned} P(T = C_1 | O = C_1) &= 0.95 \\ P(T = C_1 | O = C_2) &= 0.05 \\ P(T = C_2 | O = C_1) &= 0.05 \\ P(T = C_2 | O = C_2) &= 0.95 \\ P(O = C_1) &= 0.999 \\ P(O = C_2) &= 0.001 \end{aligned}$$

   Nyt voimme laskea Bayesin kaavan

   $$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}$$

   avulla todennäköisyyden, että laiteen väittäessä sanan perusmuodoksi *"siittää"* se on myös oikeassa.

   $$\begin{aligned} P(O &= C_2 | T = C_2) \\ &= \frac{P(T{=}C_2|O{=}C_2)P(O{=}C_2)}{P(T{=}C_2|O{=}C_2)P(O{=}C_2) + P(T{=}C_2|O{=}C_1)P(O{=}C_1)} \\ &= \frac{0.95 \cdot 0.001}{0.95 \cdot 0.001 + 0.05 * 0.999} \approx 0.019 \end{aligned}$$

Sanoista, joiden perusmuodoksi laite on ehdottanut *"siittää"* vain joka viideskymmenes on oikein jäsennetty. Vaikka Åke olikin saanut ihan hyvät tunnistustulokset sinänsä, käytännön testejen jälkeen hän päätti romuttaa tunnistimensa ja ryhtyä jazz-muusikoksi.

3. Jotta tällainen satunnainen kieli generoisi yksikirjaimisen sanan, sen pitää generoida kaksi merkkiä (joku muu kuin sanaväli ja sanaväli).

$$P(s = t_1) = \frac{1}{30} \cdot \frac{1}{30}$$

Tällaisia sanoja on 29 kappaletta.

Vastaavasti, tietyn kahden merkin pituisen sanan todennäköisyys on

$$P(s = t_1, t_1) = \frac{1}{30} \cdot \frac{1}{30} \cdot \frac{1}{30}$$

Tällaisia sanoja on 29^2 kappaletta. Kolmikirjaimiset sanat

$$P(s = 3) = \frac{1}{30} \cdot \frac{1}{30} \cdot \frac{1}{30} \cdot \frac{1}{30}$$

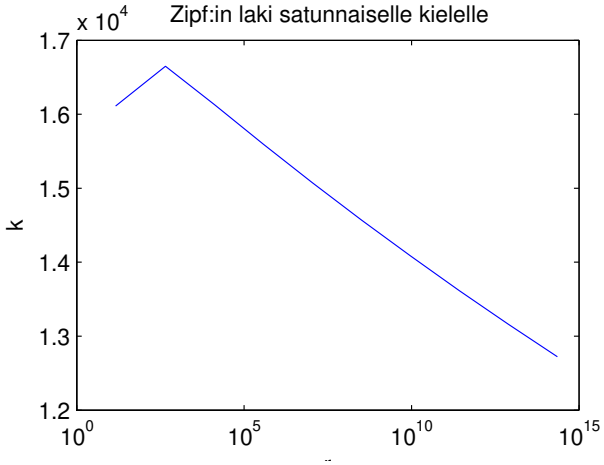
ja näitä sanoja on siis 29^3 kappaletta.

Koska sanan esiintymistodennäköisyys on suoraan verrannollinen sen odotettuun esiintymistiheyteen testiaineistossa, voimme tehdä kirjan taulukon 1.3 kaltaisen taulukon suoraan laskemalla todennäköisyyksiä. Koska samanpituiset sanat ovat yhtä todennäköisiä eikä niitä voi asettaa yleisyysjärjestykseen, laskemme k:n arvon vain yhdelle samanpituisista sanoista. Tulokset on esitetty taulukossa 1 ja piirretty kuvaan 1.

Taulukko 1: *Zipfin vakio. Taulukon vasempaan sarakkeeseen on merkitty kuinka monenneksi yleisin sana on kyseessä. Keskellä lukee, kuinka monta kertaa voimme odottaa näkevämme sanan 1000000 sanan pitusessa aineistossa. Oikealla on laskettu vakio k, kahden ensimmäisen sarakkeen tulo.*

r	f	k
15	1111	16111
450	37.04	16648
13064	1.235	16129
378900	0.0412	15593
1098800	0.00137	15073
318660000	0.0000457	14570

Huomataan, että satunnaisellakin kielellä k pysyttelee melko samansuuruisena hyvin suurella r:n vaihteluvälilläkin. Zipfin löytö ei ehkä tunnu tämän faktan valossa aivan niin hämmästyttävltä.



Kuva 1: *k r:n funktiona*

4. Tehtävän ratkaisussa oletetaan tunnetuksi seuraavat kaavat:

$$\begin{aligned} E(x) &= \int_{-\infty}^{\infty} xp(x)dx \\ Var(x) &= \int_{-\infty}^{\infty} (x - E(x))^2 p(x)dx \end{aligned}$$

a) Lasketaan odotusarvo yhden heiton silmäluvuksi. Noppa laskeutuu jokaiselle 101:lle sivustaan yhtä todennäköisesti, eli jokaisen tapahtuman todennäköisyys $p(x) = \frac{1}{101}$.
Odotusarvo:

$$\begin{aligned} E(x) &= \sum_{i=0}^{100} ip(x = i) \\ &= \frac{1}{101}(1 + 2 + 3 + 4 + \cdots + 100) \\ &= \frac{1}{101}\left((1 + 100) + (2 + 99) + (3 + 98) + \cdots + (50 + 51)\right) \\ &= \frac{50 * 101}{101} = 50 \end{aligned}$$

Varianssi voidaan laskea kaavalla:

$$\begin{aligned} Var(x) &= \sum_{i=0}^{100} (i - E(x))^2 p(x = i) \\ &= \frac{1}{101}(50^2 + 49^2 + \cdots + 1 + 0 + 1 + 2^2 + \cdots + 49^2 + 50^2) \\ &= \frac{2}{101}(1 + 2^2 + \cdots + 49^2 + 50^2) \end{aligned}$$

Nyt voimme käyttää avuksemme seuraava kaavaa

$$1 + 2^2 + 3^2 + 4^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

jolloin saamme tulokseksi

$$Var(x) = \frac{2}{101}\frac{50 \cdot 51 \cdot 101}{6} = 850$$

b) Ratkaistaksemme tämän tehtävä, tarvitsemme muutamia todennäköisyyslaskun peruskaavoja. Kaavat on tässä johdettu, mutta niiden johtamisen osaaminen ei ole olennaista kurssin kannalta.

Riippumattomien satunnaismuuttujien summan oletusarvo

Olkoon satunnaismuuttujat x ja y riippumattomia. Lasketaan näiden satunnaismuuttujien summan oletusarvo.

$$\begin{aligned}
E(x+y) &= \int (x+y)p(x,y)dxdy \\
&= \int (x+y)p(x)p(y)dxdy \\
&= \int xp(x)p(y)dxdy + \int yp(x)p(y)dxdy \\
&= \int p(y)dy \int xp(x)dx + \int p(x)dx \int yp(y)dy \\
&= 1 \cdot \int xp(x)dx + 1 \cdot \int yp(y)dy \\
&= E(x) + E(y)
\end{aligned}$$

Vakiolla kerrotun satunnaismuuttujan varianssi

$$\begin{aligned}
Var(ax) &= \int (ax - E(ax))^2 p(x)dx \\
&= \int (ax - aE(x))^2 p(x)dx \\
&= a^2 \int (x - E(x))^2 p(x) \\
&= a^2 Var(x)
\end{aligned}$$

Riippumattomien satunnaismuuttujien summan varianssi

Olkoon satunnaismuuttujat x ja y riippumattomia. Lasketaan näiden satunnais-

muuttujien summan varianssi.

$$\begin{aligned}
Var(x+y) &= \int\int (x+y-E(x+y))^2 p(x,y)dxdy \\
&= \int\int (x+y)^2 p(x,y)dxdy - 2\int\int (x+y)E(x+y)p(x,y)dxdy \\
&\quad + \int\int E(x+y)^2 p(x,y)dxdy \\
&= E((x+y)^2) - 2E(x+y)^2 + E(x+y)^2 \\
&= E((x+y)^2) - E(x+y)^2 \\
&= E(x^2+2xy+y^2) - (E(x)+E(y))^2 \\
&= E(x^2) + E(2xy) + E(y^2) - E(x)^2 - 2E(x)E(y) - E(y)^2 \\
&= E(x^2) - E(x)^2 + E(y^2) - E(y)^2 \\
&\quad + \int\int 2xyp(x)p(y)dxdy - 2\int xp(x)dx \int yp(y)dy \\
&= E(x^2) - E(x)^2 + E(y^2) - E(y)^2 \\
&= Var(x) + Var(y)
\end{aligned}$$

Tämän pakerruksen jälkeen päästään itse asiaan. Nyt halutaan laskea oletusarvo lauseelle $(x+y)/2$, missä x on ensimmäiseen heittoon liittyvä satunnaismuuttuja ja y on toiseen heittoon liittyvä satunnaismuuttuja.

$$E(\frac{x+y}{2}) = \frac{1}{2}(E(x)+E(y)) = \frac{1}{2}(50+50) = 50$$

Huomaamme siis, että odotusarvo ei muutu. Entä miten käykään varianssin?

$$\begin{aligned}
&Var(\frac{x+y}{2}) = Var(\frac{x}{2}) + Var(\frac{y}{2}) = \frac{1}{4}Var(x) + \frac{1}{4}Var(y) \\
&= \frac{1}{4}(850+850) = 425
\end{aligned}$$

c) Heitämme kymmentä noppaa, sovellamme edelle opittuja tuloksia. Odotusarvo

$$E(\frac{x_1+x_2+\cdots+x_{10}}{10}) = \frac{1}{10}\cdot 10\cdot 50 = 50$$

Varianssi

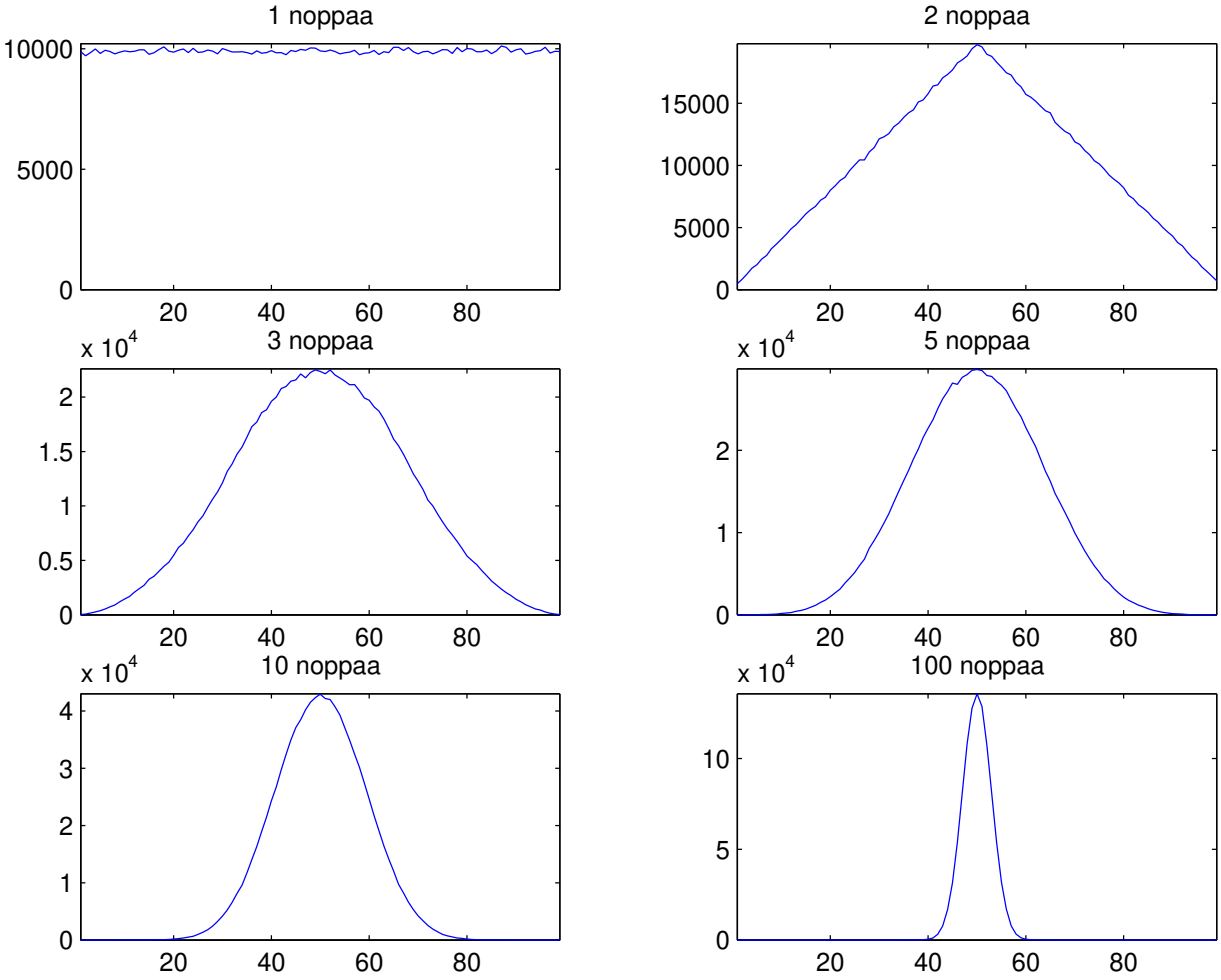
$$Var(\frac{x_1+x_2+\cdots+x_{10}}{10}) = \frac{1}{100}\cdot 10\cdot 850 = 85$$

d) Kun heitämme yhä useampaa noppaa, tarkentuu jakauma odotusarvon ympärille. Rajalla odotusarvo on 50 ja varianssi 0 eli saamme aina varmasti tulokseksi 50.

Odotusarvo ja varianssi eivät suinkaan kerro kaikkea jakaumasta. Kuvassa 2 on simuloitu matlabilla erilaisia määriä nopanheittoa. Huomaamme että jakauman muoto

muuttuu, mitä useampaa nopaa heitetään. Muoto tulee lähemmäksi ja lähemmäksi normaalijakaumaa. Tämän takia useita luonnollisia ilmiöitä mallinnetaan normaalijakaumalla: Jos tulokseen vaikuttaa monta pientä satunnaista asiaa, tulos on normaalisti jakautunut. Tämä on myös hyvä tekosyy käyttää normaalijakaumaa, jolla saadaan laskut usein helppoon muotoon.

Formaalimpi todistelu siitä, että jakauma lähestyy normaalijakaumaa löytyy http://mathworld.wolfram.com/CentralLimitTheorem.html



Kuva 2: *Nopanheittoa. Kutakin kuvaa varten on koe toistettu miljoona kertaa.*

5. Tarkoitus on siis minimoida kokonaiskuvauspituutta

$$L(x, \theta) = L(\theta) + L(x \,|\, \theta).$$

Merkitään lausekkeen minimoivaa parametrijoukkoa $\hat{\theta}$:lla. Saadaan

$$\hat{\theta} = \arg\min_{\theta} L(x, \theta) = \arg\min_{\theta} \{L(\theta) + L(x \,|\, \theta)\}.$$

Sijoitetaan tähän optimaaliset kuvauspituudet $L(\theta) = -\log p(\theta)$ ja $L(x \mid \theta) = -\log p(x \mid \theta)$:

$$\hat{\theta} = \arg\min_{\theta}\{-\log p(\theta) - \log p(x \mid \theta)\}$$

Yhdistetään termit logaritmien laskusääntöä käyttäen:

$$\hat{\theta} = \arg\min_{\theta}\{-\log(p(\theta)p(x \mid \theta))\}$$

Logaritmi on monotonisesti kasvava funktio, ja sen vastaluku siten monotonisesti laskeva, joten sama arvo saadaan maksimoimalla todennäköisyyksien tuloa:

$$\hat{\theta} = \arg\max_{\theta}\{p(\theta)p(x \mid \theta)\}$$

Lopuksi muistetaan Bayesin kaavasta $p(x, \theta) = p(x)p(\theta \mid x) = p(\theta)p(x \mid \theta)$:

$$\hat{\theta} = \arg\max_{\theta}\{p(x)p(\theta \mid x)\}$$

Jakauma $p(x)$ ei riipu parametreista, joten se voidaan tiputtaa pois. Näin ollen samaan lopputulokseen päästään mallin posteriorijakauman maksimoinnilla:

$$\hat{\theta} = \arg\max_{\theta} p(\theta \mid x)$$