T-61.5020 Luonnollisen kielen tilastollinen käsittely

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

1. Englannin kielessä voisi päteä seuraavanlaiset todennäköisyydet:

   $$P(\text{ sana=lyhenne} \mid \text{sana=kolmikirjaiminen }) = 0.8$$

   $$P(\text{ sana=kolmikirjaiminen }) = 0.0003$$

   Millä todennäköisyydellä satunnainen havaittu sana on kolmikirjaiminen lyhenne?

2. Tarkastellaan lingvisti Å. Lindquistin kehittämää sanan perusmuotoistuskonetta. Kontekstin perusteella se osaa johtaa sanan *"siitä"* perusmuodoksi joko sanan *"se"* tai *"siittää"*. Laite osaa päätellä perusmuodosta *"se"* taivutetun sanan oikean perusmuodon todennäköisyydellä 0.95 ja väärä perusmuoto lipsahtaa todennäköisyydellä 0.05. Samoin käy perusmuodosta *"siittää"* taivutetuille sanoille. Koska perusmuoto *"se"* on paljon yleisempi, vain joka tuhannes *"siitä"* pitää perusmuotoistaa sanaksi *"siittää"*.

   Laite kertoo meille, että erään sanan *"siitä"* perusmuoto on *"siittää"*. Millä todennäköisyydellä laite on oikeassa?

3. Kun lasketaan kielestä yksinkertaisia tilastoja, viitataan usein Zipfin lakiin. Sanat taulukoidaan niin, että yleisin laitetaan ensimmäiseksi ($r = 1$) ja muut järjestyksessä sen perään ($r = 2, 3, \ldots$). Kunkin sanan viereen kirjoitetaan kuinka monta kertaa se esiintyi tekstissä (f). Zipf väittää että

   $$f \propto \frac{1}{r}$$

   Sanallisesti sanottuna siis f on verrannollinen $\frac{1}{r}$:ään tai $f * r$=vakio.

   Päteekö Zipfin laki satunnaisesti generoidulle kielelle, jossa on 30 kirjainta, joista yksi on sanaväli?

4. a) Heitetään 101-sivuista noppaa, jonka sivuilla on luvut 0 – 100. Laske saadun silmäluvun odotusarvo ja varianssi. Hahmottele todennäköisyyden p(X) kuvaaja jossa X on heiton silmäluku.

   b) Heitetään kahta 101-sivuista noppaa ja jaetaan silmälukujen summa kahdella. Laske tuloksen odotusarvo ja varianssi. Hahmottele todennäköisyyden p(X) kuvaaja jossa X on heiton silmälukujen summa jaettuna kahdella.

   c) Ja vielä heitetään kymmentä noppaa, jaetaan tulos kymmennellä. Hahmottele kuvaaja.

d) Etsitään käsiimme kaikki maailman nopat ($n \to \infty$). Millainen jakauma meillä nyt mahtaa olla? Hahmottele kuvaaja.

5. *Pienimmän kuvauspituuden periaate* (Minimum Description Length principle, MDL) on eräs tapa etsiä parasta mallia havaittavalle datajoukolle. Siinä pyritään löytämään datalle kuvaus, jonka avulla se voidaan tallentaa pienimmällä mahdollisella tilamäärällä. Tilankäyttöä mitataan bitteinä. Ideaalisessa pienimmän kuvauspituuden periaatteessa etsitään pienintä mahdollista Turingin konetta (toisin sanoen tietokoneohjelmaa), joka tuottaa halutun datajoukon.

   Ideaalinen pienin kuvauspituus on osoitettu mahdottomaksi löytää, ja siksi menetelmästä on vähemmän objektiivisia mutta käyttökelpoisempia versioita. Kaksiosaisessa koodausmenetelmässä (two-part coding scheme) valitaan ensin mallien luokka, joka kuvaa dataa annetulla parametrijoukolla θ. Tarkoitus on kuvata ja lähettää pienimmällä mahdollisella bittimäärällä datajoukko x, jonka oletetaan olevan generoitu jollain luokan malleista. Vastaanottaja tietää mallien luokan, muttei sen parametrien arvoja, joten myös ne pitää lähettää. Merkitään parametrien kuvauspituutta $L(\theta)$:lla ja data kuvauspituutta, kun mallin parametrit tiedetään, $L(x \mid \theta)$:lla. Tarkoitus on minimoida kokonaiskuvauspituus $L(x, \theta) = L(\theta) + L(x \mid \theta)$.

   Tilastollisessa mallinnuksessa malliluokan koodausta vastaa todennäköisyysjakauma $p(X \mid \theta)$ ja parametrien koodausta jakauma $p(\theta)$. Informaatioteoriasta tiedämme, että jos viestin todennäköisyys on $p(i)$, sen optimaalinen koodauspituus on $-\log p(i)$ bittiä. Näytä, että parametrien valinta kaksiosaisessa koodausmenetelmässä vastaa mallin posterioiritodennäköisyyden valintaa *Maximum A Posteriori* (MAP) -estimoinnissa.