T-61.281 Luonnollisten kielten tilastollinen käsittely

Harjoitus 10, ti 1.4.2003, 16:15-18:00 – Sanojen merkitysten erottelu, Versio 1.0

Varsinaiset laskaritehtävät

1. Bayesin kaavan mukaan merkityksen s_k todennäköisyys saadaan kun tiedetään konteksti c kaavasta
$$P(s_k|c) = \frac{P(c|s_k)P(s_k)}{P(c)}$$
Naiivi Bayes-menetelmä perustuu oletukseen, että kontekstin sanojen esiintymistodennäköisyydet eivät riipu toisistaan. Johda naiivin Bayesluokittimen kaava.

2. Sinulle on annettu kaksi lauseryhmää. Käytä näillä lauseilla opetettua naiivia Bayes-tunnistinta erottamaan testilauseista sanan "sataa" eri merkitykset. Käytössäsi on laite, jolla voi kaikki muut numerot palauttaa muotoon "num", mutta joka ei tietenkään osaa erotella "sataa" sanan merkityksiä).
Vinkki: Kannattaa käyttää estimaateissa prioria, joka sanoo että kaikki sanat esiintyvät yhtä todennäköisesti kaikissa konteksteissa. Tätä voidaan kuvata tässä esim. lisäämällä kaikkiin tarvittaviin laskureihin hihasta ravistettu λ, joka kuvaa priorius-komuksen vahvuutta. Voit olettaa, että malli tuntee vain sekä opetus- että testijoukossa olevat sanat (yhteensä 85 erilaista).
Vinkki2: Kannattaa laskea vain testijoukon tunnistamiseen tarvittavat tiedot.

   Ryhmä 1.

   Tuulet puhaltavat edelleen noin 100 kilometrin tuntivauhtia ja paikoin **sataa** rankasti

   Washingtonin alueella **sataa** lunta tai keli on muuten huono

   Meille **sataa** mannaa taivaasta

   Jos kilpailun aikana **sataa** vettä

   Joulun aikana täytyy **sataa** lunta 20-30 senttiä

   Joulunpyhinä **sataa** runsaammin lunta, Lounais-Suomessa mahdollisesti räntää sekä vettä

   Ryhmä 2.

   Kapinoivat vangit pitivät edelleen maanantaina noin **sataa** vanginvartijaa panttivankeinaa

   Kaikkiaan paikalla oli noin pari **sataa** virkavallan edustajaa

   Poliisi pidätti pari **sataa** ihmistä

   Räjähdyksissä kuoli noin kaksi **sataa** ihmistä

   Lusin pohjoispuolella saa ajaa **sataa** viiden kilometrin matkalla

   Talvimyrsky irrotti Stora-Enson Kotkan sahan kattoa monta **sataa** neliömetriä

Testijoukko

Koirasusitarhassa vieraili pari **sataa** ihmistä.

Pohjois-Suomessa räntää tai lunta **sataa** keskimäärin joka kolmantena vappuna

Itse tapahtumaan odotetaan noin kuutta **sataa** vierasta

Pommeja **sataa** kaikkialla eikä kaduilla ole ketään

3. Käytettävissäsi on seuraava ote synonyymisanakirjasta (thesaurus):

   ammunta: (1) Tilanne, jossa harjoitellaan aseiden käyttöä. Esim.
   Joukkue harjoitteli konepistoolilla sarjatulen ammuntaa.
   Ammunta on aiheuttanut varusmiehille kuulovaurioita

   ammunta: (2) Nautakarjasta lähtevä äännähtely. Esim.
   Niityltä oli kuulunut lehmän ammuntaa.
   Ammunnan hälyttämä naapurin isäntä löysi nälkiintyneen vasikan.

   varusmies: asepalvelusta suorittava kansalainen

   loukkaantua (1): pahastua
   Hän pahastua kovasti siitä, mitä kuuli.

   loukkaantua (2): satuttaa itsensä
   Hän loukkaantui törmäyksessä.

   kivääri: ampuma-ase, jota usein käytetään isomman riistan, esim. hirven metsästyksessä. *rynnäkkö˜*, sotaa varten kehitetty versio, joka pystyy ampumaan sarjatulta.

   harjoitella: toistaa usein jotain harrastusta oppiakseen paremmin suoriutumaan siitä.
   Hän harjoitteli joka päivä pianon soittoa.

   Tarkastellaan lausetta "Varusmies loukkaantui harjoitellessaan kiväärillä ammuntaa niityllä." Valitse näiden tietojen perustella lauseen "ammunta"-sanan merkitys. Käytössäsi on laite, joka perusmuotoistaa sanat.

Sivuhuomautus: Vaikka kirjan menetelmissä käytetään runsaasti sanaa Bayes, mitkään estimaatit siellä eivät ole Bayesiläisiä vaan perinteisiä maksimiuskottavuusestimaatteja.

Tietokonelaskarit

4. Käytössäsi on englanninkielinen aineisto (esim. Google, http://www.google.com). Haluat tietää

   a) sanan *kallistua* merkityksen lauseessa *"Hinnat kallistuivat"*. Sanakirjasta näet taulukon 1 faktat.

   b) tarkoitetaanko lauseessa *"Haluatko potkia, sorkkia, maksaa vai kärsiä ?"* sanoilla *potkia, sorkkia maksaa* ja *kärsiä* lihakaupan tiskiltä saatavaa tavaraa vai erilaisia toimia (verbejä). Käännöstietoja löytyy taulukosta 1.

Sana	Käännös
hinta	price
kallistua	s_1 : slant, lean, lurch s_2 : go up
haluta	want, like, desire, covet
potkia	kick
potka	shin
sorkkia	poke, prod, fiddle
sorkka	hoof
maksaa	cost, pay
maksa	liver
vai	or
kärsiä	suffer, ache, sustain
kärsä	snout

Taulukko 1: *Otteita sanakirjasta.*

5. Sana "kuusi" on esiintynyt allaolevan listan konteksteissa. Tiedetään etukäteen, että sanalla on kaksi merkitystä. Luokittele EM-algoritmilla kontekstit kahteen ryhmään sen mukaan, kummassa merkityksessä sana "kuusi" esiintyi.

   yksi kaksi kolme

   kaksi kolme neljä

   neljä viisi seitsemän

   mänty leppä haapa

   leppä haapa koivu

   haapa koivu kataja

   koivu kataja leppä

   yksi mänty kaksi haapa leppä

   yksi haapa seitsemän kahdeksan

   kaksi haapa

6. Etsi käsiisi suomenkielistä tekstiä ja tee ohjelma, joka erottelee ohjaamattamasti sanan eri merkityksiä. Voit käyttää joko oikeasti monimerkityksistä sanaa tai luoda keinotekoisen monimerkityksisen sanan (esim. muuttamalla aineistosta löytyvät sanat "sade" ja "komissio" yhdeksi sanaksi "sadekomissio" ja tutkimalla tämän sanan eri merkityksiä).