T-61.5020 Luonnollisten kielten tilastollinen käsittely

Vastaukset 6, to 28.2.2007, 12:15–14:00 — Samankaltaisuusmitat
Versio 1.1

1. **Euklidinen etäisyys (eli L_2-normi)**

   Euklidinen etäisyys vektorin $\mathbf{x} = [x_1\ x_2 \ldots x_n]$ ja $\mathbf{y} = [y_1\ y_2 \ldots y_n]$ välillä määritellään

   $$Euc(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

   Lasketaan euklidinen etäisyys esimerkin vuoksi Tintuksen ja Korvalääkkeen välillä:

   $$\begin{aligned} Euc(Ti, Ko) &= \sqrt{(0-10)^2 + (0-6)^2 + (5-2)^2 + (1-1)^2 + (4-0)^2} \\ &= 12.7 \\ Euc(Ko, Te) &= 9.9 \\ Euc(Ti, Te) &= 5.1 \end{aligned}$$

   L_1-normi

   L_1-normin mukainen etäisyys määritellään

   $$L_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i| \tag{2}$$

   Lasketaan etäisyydet:

   $$\begin{aligned} L_1(Ti, Ko) &= |0-10| + |0-6| + |5-2| + |1-1| + |4-0| \\ &= 23.0 \\ L_1(Ko, Te) &= 17.0 \\ L_1(Ti, Te) &= 10.0 \end{aligned}$$

   Kosini

   Kosini onkin sitten hieman erilainen tapaus, se on samankaltaisuusmitta. Se määritellään vaikkapa

   $$\cos(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{3}$$

	raikas	hapokas	makea	hedelmäinen	pehmeä
Tintus	0	0	0.50	0.10	0.40
Korvalääke	0.53	0.32	0.11	0.05	0
Termiitti	0.07	0.29	0.21	0.21	0.21

Taulukko 1: ML-estimaatti sanatodennäköisyyksille

Lasketaan etäisyydet:

$$\begin{aligned} cos(Ti, Ko) &= \frac{0 \cdot 10 + 0 \cdot 6 + 5 \cdot 2 + 1 \cdot 1 + 4 \cdot 0}{\sqrt{5^2 + 1 + 4^2}\sqrt{10^2 + 6^2 + 2^2 + 1^2}} \\ &= 0.14 \\ cos(Ko, Te) &= 0.55 \\ cos(Ti, Te) &= 0.70 \end{aligned}$$

Tässä siis suurempi luku vastaa suurempaa samankaltaisuutta ja etäisyydet / samankaltaisuudet ovat samassa järjestyksessä kuin edelläkin.

Informaatiosäde

Informaatiosäteen laskemista varten muodostetaan suurimman uskottavuuden estimaatit sille, että seuraava lähteen l_i (Tintus, Korvalääke, Termiitti) tuottama tunnettu sana on w_i. Tämä voidaan laskea jakamalla jokainen annetun matriisin rivin alkio rivin alkioiden summalla (Taulukko 1). Määritellään vielä, että

$$0 \log \frac{0}{x} = 0, \ \forall x \in \Re$$

Informaatiosäde voidaan laskea kaavasta

$$\begin{aligned} Irad(p, q) &= D(p||\frac{p+q}{2}) + D(q||\frac{p+q}{2}) \\ &= \sum_i p_i \log \frac{p_i}{\frac{p_i+q_i}{2}} + \sum_i q_i \log \frac{q_i}{\frac{p_i+q_i}{2}} \end{aligned}$$

Lasketaan informaatiosäde annetuille lähteillä:

$$\begin{aligned} Irad(Ti, Ko) &= 0 \cdot \log \frac{2 \cdot 0}{0.53} + 0 \cdot \log \frac{2 \cdot 0}{0.32} + 0.50 \cdot \log \frac{2 \cdot 0.50}{0.61} + 0.10 \cdot \log \frac{2 \cdot 0.10}{0.15} \\ &\quad + 0.40 \cdot \log \frac{2 \cdot 0.40}{0.40} + 0.53 \cdot \log \frac{2 \cdot 0.53}{0.53} + 0.32 \cdot \log \frac{2 \cdot 0.32}{0.32} \\ &\quad + 0.11 \cdot \log \frac{2 \cdot 0.11}{0.61} + 0.05 \cdot \log \frac{2 \cdot 0.05}{0.15} + 0 \cdot \log \frac{2 \cdot 0}{0.40} \\ &= 1.5 \\ Irad(Ko, Te) &= 0.6 \\ Irad(Ti, Te) &= 0.5 \end{aligned}$$

Huomataan, että kaikki mitat asettavat lääkeet asettavat lääkkeet samankaltaisuuksin mukaan samaan järjestykseen: Tintus ja Termiitti ovat samankaltaisimmat, Tintus ja Korvalääke erilaisimmat.

KL-divergenssi

KL-divergenssin määritelmästä voimme suoraan nähdä muutaman siihen liittyvän ongelman:

$$D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$$

KL-divergenssi ei ole symmetrinen, vaan pitäisi aina päättää kumpi lääke on referenssilääke, mihin toista verrataan. Toinen ongelma on siinä, että jos vertailtavalla jakaumalla q on nollatodennäköisyys jossain, missä referenssijakauma p ei ole nolla, niin KL-divergenssi menee äärettömyyksiin.

2. **Kullback-Leibler -divergenssi**

KL-divergenssin määritelmä on

$$D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$$

Etsitään jakauma, joka minimoi KL-divergenssin. Lisätään Lagrange-kerroin λ_1 pitämään huolta siitä, että p pysyy todennäköisyysjakaumana (eli $\sum_i p_i = 1$) ja λ_2 q:lle.

$$E = D(p||q) + \lambda_1(1 - \sum_i p_i) + \lambda_2(1 - \sum_i q_i) = \sum_i p_i \log \frac{p_i}{q_i} + \lambda_1(1 - \sum_i p_i) + \lambda_2(1 - \sum_i q_i)$$

Merkitään osittaisderivaatta p_i:n suhteen nollaksi:

$$\begin{aligned} \frac{\partial E}{\partial p_i} &= p_i \cdot \frac{1}{\frac{p_i}{q_i}} \cdot \frac{1}{q_i} + \log \frac{p_i}{q_i} - \lambda_1 \\ &= \log p_i - \log q_i + 1 - \lambda_1 = 0 \end{aligned}$$

Ratkaistaan p_i:

$$p_i = q_i \cdot e^{\lambda_1 - 1}$$

Lasketaan osittaisderivaatta λ_1 suhteen:

$$\begin{aligned} \frac{\partial E}{\partial \lambda_1} &= 1 - \sum_i p_i = 0 \\ \Rightarrow \sum_i p_i &= 1 \end{aligned}$$

Vastaava ehto q_i:lle saadaan λ_2 suhteen derivoimalla, mikä oli tarkoituskin. Lasketaan vielä nollakohta q_i:n suhteen:

$$\begin{aligned} \frac{\partial E}{\partial q_i} &= p_i \cdot \frac{1}{\frac{p_i}{q_i}} \cdot p_i \cdot (-\frac{1}{q_i^2}) - \lambda_2 = -\frac{p_i}{q_i} - \lambda_2 = 0 \\ \Leftrightarrow p_i &= -\lambda_2 q_i \end{aligned}$$

Koska sekä q:n että p:n tulee siis summatua yhteen, saamme:

$$\begin{aligned} 1 &= \sum_i p_i = \sum_i (-\lambda_2 q_i) = -\lambda_2 \sum_i q_i = -\lambda_2 \\ \Rightarrow p_i &= -\lambda_2 q_i = q_i \end{aligned}$$

Toisen asteen derivaattoja tarkastelemalla voimme vielä varmistua siitä että tämä todellakin on minimi eikä maksimi:

$$\begin{aligned} \frac{\partial^2 E}{\partial p_i \partial p_i} &= \frac{1}{p_i} > 0 \\ \frac{\partial^2 E}{\partial q_i \partial q_i} &= \frac{p_i}{q_i^2} > 0 \\ \frac{\partial^2 E}{\partial p_i \partial p_j} = \frac{\partial^2 E}{\partial q_i \partial q_j} &= 0 \end{aligned}$$

Jos sijoitamme KL-divergenssin kaavaan $q_i = p_i$ saamme divergenssiksi nolla. Eli *KL-divergenssi on nolla jos ja vain jos jakaumat q ja p ovat samoja, muuten nollaa suurempi.*

Informaatiosäde

Informaatiosäteen määritelmä on

$$IRad(p,q) = D(p||\frac{p+q}{2}) + D(q||\frac{p+q}{2})$$

Laskimme juuri, että KL-divergenssi on nolla, kun jakaumat ovat samat ja muuten tätä enemmän. Informaatiosäteen tapauksessa nolladivergenssiin siis päästään myös vain kun $q_i = p_i$:

$$IRad(p,q) = \sum_i p_i \log \frac{p_i}{\frac{p_i+p_i}{2}} + \sum_i p_i \log \frac{p_i}{\frac{p_i+p_i}{2}} = 0$$

Ehto on siis sama kuin KL-divergenssillä.

L_1-normi

L_1-normin määritelmä on

$$L_1(p,q) = \sum_i |p_i - q_i|$$

Tämähän on selvästi pienimmillään nolla. Se tapahtuu kun $q_i = p_i$.

Huomataan siis, että kaikki mitat antavat pienimmän etäisyyden samalla ehdolla — jakaumien on oltava samat — ja tämä pienin arvo on nolla.

3. **Kullback-Leibler -divergenssi**

Katsotaan vielä KL-divergenssin määritelmää:

$$D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$$

Huomataan, että jos $q_i = 0$ kun $p_i \neq 0$, saadaan etäisyydeksi ∞.

Informaatiosäde

Kirjoitetaan informaatiosäteen määritelmä auki:

$$IRad(p,q) = D(p||\frac{p+q}{2}) + D(q||\frac{p+q}{2}) = \sum_i p_i \log \frac{2p_i}{p_i+q_i} + \sum_i q_i \log \frac{2q_i}{p_i+q_i}$$

Intuition avulla arvataan sopivaksi jakaumaksi sellainen, missä jakaumat sijaitsevat täysin eri alueilla:

$$\text{jos } p_i > 0 \Rightarrow q_i = 0$$
$$\text{jos } q_i > 0 \Rightarrow p_i = 0$$

Sijoitetaan tällaiset jakaumat informaatiosäteen lausekkeeseen:

$$\begin{aligned} IRad(p,q) &= \sum_i p_i \log \frac{2p_i}{p_i} + \sum_i q_i \log \frac{2q_i}{q_i} \\ &= \log 2 \sum_i p_i + \log 2 \sum_i q_i = 2 \log 2 \end{aligned}$$

Huomataan, että ehdot täyttävä jakauma antaa suurimman etäisyyden. Todistus siitä, että $2 \log 2$ on suurin mahdollinen informaatiosäde ja että ylläarvatut ehdot vaaditaan tämän etäisyyden saavuttamiseksi olisi sitten jonkin verran hankalampi.

L_1-normi

L_1-normin määritelmähän oli

$$L_1(p, q) = \sum_i |p_i - q_i|$$

Intuitiollahan voisi jo päätellä, että vastaus on sama kuin informaatiosäteen tapauksessa, mutta yritetään perustella asiaa vielä matemaattisesti. Jaetaan alkeistapaukset I kahteen osaan. Osassa $j \in I$ on tapaukset, joissa $p_j > q_j$ ja osassa $k \in I$ tapaukset, joissa $q_k > p_k$. Kirjoitetaan itseisarvot auki:

$$\begin{aligned} L_1(p, q) &= \sum_j (p_j - q_j) + \sum_k (q_k - p_k) \\ &= \sum_j p_j - \sum_k p_k + \sum_k q_k - \sum_j q_j \end{aligned}$$

Koska todennäköisyydet ovat positiivisia ja summautuvat 1:een, suurin etäisyys saadaan kun

$$\begin{aligned} &\text{jos } p_i > 0 \Rightarrow q_i = 0 \\ &\text{jos } q_i > 0 \Rightarrow p_i = 0 \end{aligned}$$

eli etäisyys on

$$L_1(p, q) = \sum_i p_i + \sum_i q_i = 2$$

Yhteenveto

Informaatiosäteen ja L_1-normin tapauksessa kahden todennäköisyysjakauman välisen suurimman etäisyyden saavuttamiseen vaaditaan samat ehdot. Sen sijaan KL-divergenssi menee äärettömyyksiin jo, kun vertailujakauma q on nolla jossain missä p ei ole nolla.