T-61.5020 Luonnollisen kielen tilastollinen käsittely

Vastaukset 2, ke 31.1.2007, 12:15–14:00 — Entropia ja hämmentyneisyys
Versio 1.0

1. a) Sijoitetaan entropian kaavaan

$$H(X) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

tehtävässä annetut arvot:

$$\begin{aligned} H(X) &= \frac{3}{32} \log_2 \frac{32}{3} + \frac{3}{16} \log_2 \frac{16}{3} + \frac{7}{32} \log_2 \frac{32}{7} \\ &\quad + \frac{1}{8} \log_2 8 + \frac{1}{8} \log_2 8 + \frac{1}{4} \log_2 4 \\ &= 2.50 \text{ bittiä} \end{aligned}$$

b) Tehtävän ratkaisuun tarvitaan todennäköisyyttä $P(S = s)$ (eli satunnainen substantiivi on s). Tämä todennäköisyys saadaan tehtävänannon taulukon oikeasta marginaalista. Lisäksi tarvitaan todennäköisyyttä

$$P(V = v|S = s) = \frac{P(S = s, V = v)}{P(S = s)}$$

Lähteen entropia, kun tiedetään, että edellinen symboli kuului joukkoon S on

$$H(X_i|X_{i-1} \in S) = \sum_{S=\{\text{'kissa'},\text{'tuuli'},\text{'kiipeilijä'}\}} p(s = S) H(V|s = S)$$

Tämän laskemiseksi meidän pitää osata laskea ehdollinen entropia $H(V|s = S)$. Lasketaan tämä sanalle 'kissa':

$$\begin{aligned} H(V|s = \text{'kissa'}) &= \sum_{V=\text{'naukaisi'},\text{'tuivertaa'},\text{'katosi'}} p(v = V|s = \text{'kissa'}) \log_2(p(v = V|s = \text{'kissa'})^{-1}) \\ &= \sum_{V=\text{'naukaisi'},\text{'tuivertaa'},\text{'katosi'}} \frac{p(s = \text{'kissa'}, v = V)}{P(s = \text{'kissa'})} \log_2 \frac{P(s = \text{'kissa'})}{p(s = \text{'kissa'}, v = V)} \\ &= \frac{1}{8}\frac{16}{3} \log_2(8\frac{3}{16}) + \frac{1}{16}\frac{16}{3} \log_2(16\frac{3}{16}) \\ &= \frac{2}{3} \log_2 \frac{3}{2} + \frac{1}{3} \log_2 3 \end{aligned}$$

Kun sijoitamme jokaista joukon S sanaa vastaavat todennäköisyydet, saamme

$$\begin{aligned} H(X_i|X_{i-1} \in S) &= \frac{3}{16}(\frac{2}{3} \log_2 \frac{3}{2} + \frac{1}{3} \log_2 3) + \frac{3}{8}(\frac{1}{6} \log_2 6 + \frac{4}{6} \log_2 \frac{6}{4} + \frac{1}{6} \log_2 6) \\ &\quad + \frac{7}{16}(\frac{1}{7} \log_2 7 + \frac{6}{7} \log_2 \frac{7}{6}) \\ &= 0.90 \text{ bittiä} \end{aligned}$$

Mikä on sitten todennäköisyys, että satunnainen sana on 'kissa'? Koska molemmat luokat S ja V ovat yhtä todennäköiset, tulos on

$$P(x = \text{'kissa'}) = P(x \in S)P(S = x) = 0.5 \cdot \frac{3}{16} = \frac{3}{32}$$

Huomaamme, että a)-kohta on itseasiassa b)-kohdan marginaalijakauma.

Tästä voimme päätellä, että kun tunnemme lähteen toiminnan paremmin, sen tuottamat sanat ovat vähemmän yllättäviä ja voimme koodata ne vähemmällä määrällä bittejä (0.9 bittiä $<$ 2.5 bittiä).

c) Kielen lauseissa ensimmäinen sana on aina substantiivi ja toinen verbi. Substantiivi ei riipu edellisistä sanoista, mutta verbi riippuu edellisestä substantiivista. Merkitään kielen lausetodennäköisyyksiä $P(S, V)$ ja mallin antamia todennäköisyyksiä $P_M(S, V)$. Haluamme laskea odotusarvon kielen lauseen koodauspituudelle mallin antamilla todennäköisyyksillä:

$$E(-\log P_M(S, V)) = -\sum_{s \in S, v \in V} P(S = s, V = v) \log P_M(S = s, V = v).$$

Tätä mittaa kutsutaan risti-entropiaksi (cross-entropy).

Mallin antamat verbien ja substantiivien todennäköisyydet ovat toisistaan riippumattomia, joten $P_M(S = s, V = v) = P_M(S = s)P_M(V = v)$. Sijoitetaan se lausekkeeseen, ja kirjoitetaan summaus auki ensin substantiivien ja sitten verbien osalta:

$$\begin{aligned}
& E(-\log P_M(S, V)) \\
= & -\sum_{s \in S, v \in V} P(S = s, V = v) \log P_M(S = s, V = v) \\
= & -\sum_{s \in S} \sum_{v \in V} P(S = s)P(V = v | S = s) \log(P_M(s)P_M(v)) \\
= & -P(S = \text{kissa}) \sum_{v \in V} P(V = v | S = \text{kissa}) \log(P_M(\text{kissa})P_M(v)) \\
& -P(S = \text{tuuli}) \sum_{v \in V} P(V = v | S = \text{tuuli}) \log(P_M(\text{tuuli})P_M(v)) \\
& -P(S = \text{kiipelijä}) \sum_{v \in V} P(V = v | S = \text{kiipelijä}) \log(P_M(\text{kiipelijä})P_M(v)) \\
= & -\frac{3}{16} \cdot [\frac{1}{8}\frac{16}{3} \log(\frac{3}{32}\frac{1}{8}) + \frac{1}{16}\frac{16}{3} \log(\frac{3}{32}\frac{1}{4})] \\
& -\frac{3}{8} \cdot [\frac{1}{16}\frac{8}{3} \log(\frac{3}{16}\frac{1}{8}) + \frac{1}{4}\frac{8}{3} \log(\frac{3}{16}\frac{1}{8}) + \frac{1}{16}\frac{8}{3} \log(\frac{3}{16}\frac{1}{4})] \\
& -\frac{7}{16} \cdot [\frac{1}{16}\frac{16}{7} \log(\frac{7}{32}\frac{1}{8}) + \frac{3}{8}\frac{16}{7} \log(\frac{7}{32}\frac{1}{4})] \\
= & \; 5.01
\end{aligned}$$

Lauseden keskimääräinen koodauspituus (ts. risti-entropia lausetta kohti) on siis 5.01 bittiä.

Jokaisessa lauseessa on kaksi sanaa, joten keskimääräinen yhden sanan koodauspituus on 2.50 bittiä. Tulos on sama kuin a)-kohdassa, eikä se ole sattumaa: Kummassakin tapauksessa jakauma, jonka yli odotusarvo mallin antamille koodauspituuksille lasketaan, on sama.

2. a) Kunkin alkeistapauksen todennäköisyys on $\frac{1}{30}$. Alkeistapauksia on 30. Sijoitetaan entropian kaavaan:

$$\begin{aligned} H(X) &= \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)} \\ &= \sum_{i=1}^{30} \frac{1}{30} \log_2(30) \\ &= \log_2(30) \approx 4.91 \text{ bittiä} \end{aligned}$$

b) Sanan, jossa on vain yksi merkki, sanotaan vaikka joukon ensimmäinen merkki, todennäköisyys on

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

sillä ensimmäisen merkin pitää olla joukon ensimmäinen ja sitten pitää tulla sanaväli. 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. Homma jatkuu samalla tavalla useammille sanoille.

Lasketaan tällaisen lähteen entropia:

$$\begin{aligned} H(X) &= \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)} \\ &= 29 * (\frac{1}{30})^2 \log_2(30^2) + 29^2 * (\frac{1}{30})^3 \log_2(30^3) + 29^3 * (\frac{1}{30})^4 \log_2(30^4) + \dots \\ &= \frac{1}{29} \left(\left(\frac{29}{30}\right)^2 \cdot 2 \cdot \log_2(30) + \left(\frac{29}{30}\right)^3 \cdot 3 \cdot \log_2(30) + \left(\frac{29}{30}\right)^4 \cdot 4 \cdot \log_2(30) + \dots \right) \\ &= \frac{\log_2(30)}{29} \left(-\frac{29}{30} + \sum_{i=0}^{\infty} i \cdot \left(\frac{29}{30}\right)^i \right) \end{aligned}$$

Nyt tarvitaan suluissa olevan summan arvoa. Ratkaistaan annettu sarja seuraavasti:

$$\sum_{i=0}^{\infty} i q^i = q + 2q^2 + 3q^3 + 4q^4 + \dots \tag{1}$$

Kerrotaan yhtälö q:lla.

$$q\sum_{i=0}^{\infty} iq^i = q^2 + 2q^3 + 3q^4 + 4q^5 + \dots \tag{2}$$

Vähennetään yhtälö 2 puolittain yhtälöstä 1

$$\begin{aligned}(1-q)\sum_{i=0}^{\infty} iq^i &= q + q^2 + q^3 + q^4 + \dots &(3)\\ \sum_{i=0}^{\infty} iq^i &= \frac{q + q^2 + q^3 + q^4 + \dots}{1-q} &(4)\end{aligned}$$

Kerrotaan yhtälö 4 vielä kerran q:lla

$$q\sum_{i=0}^{\infty} iq^i = \frac{q^2 + q^3 + q^4 + q^5 + \dots}{1-q} \tag{5}$$

Nyt vähennetään yhtälöt 4 ja 5 toisistaan ja saadaan ratkaisu:

$$\begin{aligned}(1-q)\sum_{i=0}^{\infty} iq^i &= \frac{q}{1-q} &(6)\\ \sum_{i=0}^{\infty} iq^i &= \frac{q}{(1-q)^2} &(7)\end{aligned}$$

Tässä ratkaisussa pitää vielä huomioda, että jotta yhtälö 2 voidaan vähentää yhtälöstä 1, pitää sarjojen olla suppenevia, eli $|q| < 1$.

Kun tämä hässäkkä sijoitetaan alkuperäiseen ongelmaan, saadaan

$$\begin{aligned}&\frac{\log_2(30)}{29}\left(-\frac{29}{30} + \frac{\frac{29}{30}}{(1-\frac{29}{30})^2}\right)\\ =\ &\log_2(30)(30 - \frac{1}{30})\\ =\ &147 \text{ bittiä}\end{aligned}$$

Ensi silmäyksellä tämä tulos saattaa tuntua hämmentävältä, eikä tuloksen pitäis olla sama kuin a)-kohdassa? Pikainen tarkistuslasku ehkä hälventää hieman epäluuloja: Sanan pituuden odotusarvo on 29, eli entropia per merkki on n. $147/(29+1) = 4.90$ bittiä.

On myös syy, miksi tulosten ei pitäisi olla aivan samat: Ensimmäinen lähde voi tuottaa sanan, jossa on kaksi välilyöntiä peräkkäin, kun taas toinen lähde ei voi annetun formuloinnin mukaan sitä tuottaa. Tästä johtuen pitäisi toisen lähteen entropia per merkki olla hieman alempi.

3. Ensin hieman johdatusta aiheeseen: Haluamme tutkia kuinka hyvin luomamme malli ennustaa mallinnettavaa ilmiötä, tässä tapauksessa kieltä. Jos mallinnettavaan ilmiöön liittyvä todennäköisyysjakauma on tiedossa, voisimme laskea esimerkiksi mallin jakauman ja sen välisen risti-entropian (tehtävä 1c). Yleensä mallinnettavaa jakaumaa ei tietenkään tunneta. (Muutenhan olisi monesti turhaa edes tehdä toista mallia.) Tällöin evaluointimitan täytyy perustua johonkin ilmiön tuottamaan datajoukkoon. Datajoukolle D voidaan esimerkiksi laskea luodun mallin M antama uskottavuus (likelihood), $P(D \mid M)$. Helpommin käsiteltävä luku on yleensä uskottavuuden logaritmi normalisoituna datajoukon koolla:

$$\text{Average-Log-Likelihood}(D \mid M) = \frac{1}{n}\sum_{i=1}^{n} \log P(D_i \mid M)$$

Kun logaritmin kantaluku on 2, tämä on itse asiassa (etumerkkiä vaille) risti-entropian suurimman uskottavuuden estimaatti. Kun muokkaamme lauseketta hieman, huomaamme vielä että tämä risti-entropian estimaatti on hämmentyneisyyden (perplexity) logaritmi:

$$-\frac{1}{n}\sum_{i=1}^{n} \log P(D_i \mid M) = \sum_{i=1}^{n} \log P(D_i \mid M)^{-\frac{1}{n}} = \log\big[(\prod_{i=1}^{n} P(D_i \mid M))^{-\frac{1}{n}}\big].$$

Huomattakoon, että näissä kaavoissa oletetaan että datapisteiden todennäköisyydet ovat toisistaan riippumattomia. Kielen mallinnuksessa näin ei yleensä ole, vaan tarkka lauseke todennäköisyyksille olisi tekstiä ennustavassa mallissa $P(D_i \mid D_1, \ldots, D_{i-1}, M)$.

Seuraavaksi itse laskuihin:

a) Merkitään mallin yksi antamaa hämmentyneisyyttä $Perp_1$, mallin 2 puolestaan $Perp_2$ ja niin edelleen.

$$\begin{aligned}
&Perp_1(\text{'kissa'}, \text{'menee'}, \text{'puuhun'})\\
&= P_1(\text{sana}_1{=}\text{'kissa'}, \text{sana}_2{=}\text{'menee'}, \text{sana}_3{=}\text{'puuhun'})^{-\frac{1}{3}}\\
&= (P_1(\text{sana}{=}\text{'kissa'})P_1(\text{sana}{=}\text{'menee'})P_1(\text{sana}{=}\text{'puuhun'}))^{-\frac{1}{3}}\\
&= (0.1 \cdot 0.1 \cdot 0.1)^{-\frac{1}{3}} = 10
\end{aligned}$$

Malli 1 siis valitsee koko ajan keskimäärin kymmenestä eri sanasta. Tulos vaikuttaa oikealta. Entäpä malli 2?

$$\begin{aligned}
& Perp_2(\text{'kissa'}, \text{'menee'}, \text{'puuhun'})\\
= \;& P_2(\text{sana}_1{=}\text{subjekti}, \text{sana}_2{=}\text{verbi}, \text{sana}_3{=}\text{kohde})^{-\frac{1}{3}}\\
= \;& (P_2(\text{sana}{=}\text{subjekti})P_2(\text{sana}{=}\text{verbi})P_2(\text{sana}{=}\text{kohde}))^{-\frac{1}{3}}\\
= \;& (0.33 \cdot 0.33 \cdot 0.33)^{-\frac{1}{3}} = 3
\end{aligned}$$

Malli 2 valitsee keskimäärin 3:sta eri vaihtoehdosta, tulos vaikuttaa järkevältä.

$$\begin{aligned}
& Perp_3(\text{'kissa'}, \text{'menee'}, \text{'puuhun'}) \\
= \ & P_3(\text{sana}_1\text{='kissa'}, \text{sana}_2\text{='menee'}, \text{sana}_3\text{='puuhun'})^{-\frac{1}{3}} \\
= \ & (P_3(\text{sana='kissa'}|\text{sana=ensimmäinen}) \\
& \cdot P_3(\text{sana='menee'} \mid \text{edellinen_sana = 'kissa'}) \\
& \cdot P_3(\text{sana='puuhun'} \mid \text{edellinen_sana = 'menee'}))^{-\frac{1}{3}} \\
= \ & (0.25 \cdot 0.33 \cdot 0.33)^{-\frac{1}{3}} = 3.32
\end{aligned}$$

Tämä malli valitsee siis keskimäärin 3.32 sanasta koko ajan.

Tämän esimerkin valossa kielimallit 1 ja 3 ovat vertailukelpoiset. Kielimalli 3 vaikuttaa näistä selvästi paremmalta. Kielimalli 2 ei voi verrata muihin, sillä se operoi selvästi pienemmällä symbolijoukolla. Selvempi esimerkki olisi ehkä kielimalli, jonka mielestä kaikki sanat kuuluvat ryhmään 1 ja tämän ryhmän todennäköisyys on siis 1. Tämä kielimalli siis hämmentyneisyyden mukaan täydellinen, sillä se ei ole yhtään yllättynyt mistään sanasta.

b) Tarkastellaanpa vielä toista testilausetta. Mallille 1

$$\begin{aligned}
& Perp_1(\text{'valas'}, \text{'on'}, \text{'kala'}, \text{'paitsi'}, \text{'ettei'}) \\
= \ & (P_1(\text{sana='valas'}) P_1(\text{sana='on'}) P_1(\text{sana='kala'}) \\
& \cdot P_1(\text{sana='paitsi'}) P_1(\text{sana='ettei'}))^{-\frac{1}{5}} \\
= \ & (0.1 \cdot 0.1 \cdot 0.1 \cdot 0 \cdot 0)^{-\frac{1}{5}} \\
= \ & \frac{1}{0^{\frac{1}{5}}} = \infty
\end{aligned}$$

Huomataan, ettei hämmentyneisyyttä voida laskea, jos malli asettaa testijoukon sanalle todennäköisyyden nolla. Usein nämä sanat jätetään huomiotta ja saadaan siis

$$\begin{aligned}
& Perp_1(\text{'valas'}, \text{'on'}, \text{'kala'}) \\
= \ & (P_1(\text{sana='valas'}) P_1(\text{sana='on'}) P_1(\text{sana='kala'}))^{-\frac{1}{3}} = 10
\end{aligned}$$

Jotta tulos olisi mielekäs, on nyt myös ilmoitettava ohi kieliopin menneet sanat, tässä tapauksessa siis $\frac{2}{5} \cdot 100\% = 40\%$ sanoista ei osunut kielioppiin. Mallille 2 sadaan vastaavasti

$$\begin{aligned}
& Perp_2(\text{'valas'}, \text{'on'}) \\
= \ & (P_2(\text{sana=subjekti}) P_2(\text{sana=verbi}))^{-\frac{1}{3}} \\
= \ & (0.33 \cdot 0.33)^{-\frac{1}{2}} = 3
\end{aligned}$$

Ohi kieliopin menee myös 60% sanoista.

Malliin kolme sopii vain kaksi ensimmäistä sanaa:

$$\begin{aligned}
& Perp_3(\text{'valas'}, \text{'on'}) \\
= \; & (P_3(\text{sana='valas'|sana=ensimmäinen}) \\
& \cdot P_3(\text{sana='on'} \mid \text{edellinen_sana} = \text{'valas'})))^{-\frac{1}{3}} \\
= \; & (0.25 \cdot 0.33)^{-\frac{1}{2}} = 3.5
\end{aligned}$$

Tässä siis 60% sanoista menee ohi kieliopin.

Ovatko b)-kohdan tulokset vertailukelpoisia? Malli 2 voidaan diskata samoilla perusteilla kuin a)-kohdassakin. Malleja 1 ja 3 voidaan vertailla, kun otetaan myös huomioon ohi kieliopin menneet sanat. Malli 1 kattaa sanaston paremmin, mutta malli 3 antaa paremman hämmentyneisyyden. Usein kielimallin laatiminen on tasapainottelua näiden kahden ominaisuuden välillä.

Mikä siis on tarinan opetus? Hämmentyneisyydellä voidaan verrata kahta kielimallia, jos tulokset lasketaan samalla tavalla ja myös ohi kieliopin menneitten sanojen osuus ilmoitetaan. Eri lähteissä olevia tuloksia verratessa kannattaa kuitenkin kiinnittää huomiota siihen, miten laskut tarkalleen ottaen on tehty, jottei vedä vääriä johtopäätöksiä.

Loppuyhteenvetona lista erilaisista entropiamitoista:

- Entropia (entropy)
  $H(X) = E(-\log P(X)) = \sum_x P(x) \log \frac{1}{P(x)}$
  Tulkinta: Lähteen itseinformaatio eli keskimääräinen koodauspituus, joka tarvitaan viestin lähettämiseen optimaalisella koodauksella

- Risti-entropia (cross-entropy)
  $H_M(X) = E(-\log P_M(X)) = \sum_x P(x) \log \frac{1}{P_M(x)}$
  Tulkinta: Keskimääräinen koodauspituus, joka tarvitaan viestin lähettämiseen mallilla M

- Suhteellinen entropia (relative entropy) / Kullback-Leibler divergenssi
  $D(P(X)||P_M(X)) = E(-\log \frac{P(X)}{P_M(X)}) = \sum_x P(x) \log \frac{P(x)}{P_M(x)} = H_M(X) - H(X)$
  Tulkinta: Kuinka monta bittiä hävitään, jos koodataan lähteen viestejä mallilla M

- Hämmentyneisyys (perplexity)
  $Perp_M(X) = 2^{H_M(X)} = \prod_x (\frac{1}{P_M(x)})^{P(x)}$
  Tulkinta: Keskimääräinen mallin M haarautumiskerroin lähteen datalle