T-61.5020 Luonnollisen kielen tilastollinen käsittely

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

1. Otetaan pieni kieli johon kuuluu kuusi sanaa:

W	P(W)
'kissa'	$\frac{3}{32}$
'tuuli'	$\frac{3}{16}$
'kiipeilijä'	$\frac{7}{32}$
'naukaisi'	$\frac{1}{8}$
'tuivertaa'	$\frac{1}{8}$
'katosi'	$\frac{1}{4}$

   a) Oletetaan että lähde tuottaa satunnaismuuttujan X arvoja yo. taulukon todennäköisyyksien mukaan. Mikä on lähteen entropia $H(X)$?

   b) Tarkemmin asiaa tutkittaessa käykin ilmi että kielessä on lauserakenne 'SV' jossa kategoriat S $\in$ {'kissa', 'tuuli', 'kiipeilijä'} ja V $\in$ {'naukaisi', 'tuivertaa', 'katosi'}. Satunaismuuttujien yhteistodennäköisyysjakauma P(S,V) on:

	'naukaisi'	'tuivertaa'	'katosi'	
'kissa'	$\frac{1}{8}$	0	$\frac{1}{16}$	$\frac{3}{16}$
'tuuli'	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{3}{8}$
'kiipeilijä'	$\frac{1}{16}$	0	$\frac{3}{8}$	$\frac{7}{16}$
	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	

   Mikä on lähteen entropia, kun tiedetään, että edellinen symboli kuului joukkoon S, eli mikä on $H(X_i|X_{i-1} \in S)$?

   c) Oletetaan että kieli noudattaa b)-kohdan todennäköisyyksiä. Mallinnetaan tämän kielen substantiivilla alkavia kaksisanaisia lauseita a)-kohdan mallilla. Laske kuinka monta bittiä keskimäärin lausetta kohti on käytettävä lauseiden koodamiseen mallin antamilla optimaalisilla koodinpituuksilla.

2. Tarkastellaan edellisesta laskarista tuttua satunnaista kieltä: 30 symbolia, jotka kukin ovat yhtä todennäköisiä. Yksi symboleista on sanaväli.

   a) Lähde generoi merkkejä yhden kerrallaan. Mikä on lähteen entropia?

   b) Lähde generoi yhden sanan kerrallaan (eli merkki/merkkejä + sanaväli). Kutakin sanaa kohdellaan yhtenä omana kokonaisuutenaan. Mikä on tällaisen lähteen entropia?

3. Meillä on kolme kielioppia, jotka on esitetty seuraavan sivun taulukoissa. Testimateriaalina meillä on kaksi lausetta:

   a) Kissa menee puuhun.

Malli 1	Malli 2
P(sana='kissa')=0.1	P(sana=subjekti)=0.33
P(sana='koira')=0.1	P(sana=verbi)=0.33
P(sana='valas')=0.1	P(sana=kohde)=0.33
P(sana='kala')=0.1	
P(sana='istui')=0.1	
P(sana='menee')=0.1	
P(sana='on')=0.1	
P(sana='puuhun')=0.1	
P(sana='kuuhun')=0.1	
P(sana='suuhun')=0.1	

Malli 3	
P(sana='kissa' | sana=ensimmäinen)	=0.25
P(sana='koira' | sana=ensimmäinen)	=0.25
P(sana='valas' | sana=ensimmäinen)	=0.25
P(sana='kala' | sana=ensimmäinen)	=0.25
P(sana='istui' | edellinen_sana $\in$ {'kissa','koira','valas','kala'})	=0.33
P(sana='menee' | edellinen_sana $\in$ {'kissa','koira','valas','kala'})	=0.33
P(sana='on' | edellinen_sana $\in$ {'kissa','koira','valas','kala'})	=0.33
P(sana='puuhun' | edellinen_sana $\in$ {'istui','menee','on'})	=0.33
P(sana='kuuhun' | edellinen_sana $\in$ {'istui','menee','on'})	=0.33
P(sana='suuhun' | edellinen_sana $\in$ {'istui','menee','on'})	=0.33

b) Valas on kala paitsi ettei.

Laske kunkin mallin hämmentyneisyys (perplexity) molemmille testilauseille. Ovatko tulokset keskenään vertailukelpoisia?

Hämmentyneisyys voidaan määritellä testijoukon sanojen todennäköisyyksien geometrisen keskiarvon käänteislukuna:

$$Perp(w_1, w_2, \ldots, w_n) = P(w_1, w_2, \ldots, w_n)^{-\frac{1}{n}}$$