INF5820 – H2008 – 4.gang, 10.9

Ordmeningsentydiggjøring
Word Sense Disambiguation (WSD)

Word Net:
Noun
S: (n) bass (the lowest part of the musical range)
S: (n) bass, bass part (the lowest part in polyphonic music)
S: (n) bass, basso (an adult male singer with the lowest voice)
S: (n) sea bass, bass (the lean flesh of a saltwater fish of the family Serranidae)
S: (n) freshwater bass, bass (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
S: (n) bass, bass voice, basso (the lowest adult male singing voice)
S: (n) bass (the member with the lowest range of a family of musical instruments)
S: (n) bass (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Hvorfor?

- Dialogsystem: Dumt å få tilsendt en fisk hvis du tror du har bestilt en el-bass.
- Oversettelse: ”Saltbakt kontrabass”
- Søk: Du er interessert i å fiske i England og vil bare ha relevante treff.
  (Prøv Google: bass England)

En komponent i mange anvendelser.

Beslektede oppgaver

- WSD: *bass* – betydning en eller to
- tagging: *fisker* – verb eller substantiv, entall eller flertall
- oversettelse: eng: *bass* – norsk: *bass* eller *abbor*

Avgjøres i kontekst
Liknende, men ikke identiske teknikker

Teknikker

- En del verktøy som er felles for disse 3 oppgavene – og flere oppgaver
- Litt mer spesiell teknikker som vi også vil bruke i kap. 23
- Interessante likheter og forskjeller til de to andre oppgavene

I dag:

1. Litt sannsynlighetsregning
2. Litt statistikk
3. Bayes-klassifisering for WSD

Sannsynlighetsregning

Basis

1. Hva er sannsynligheten for at neste bil som kommer rundt hjørnet er rød?
2. Hva er sannsynligheten for å kaste kron?
3. Hva er sannsynligheten for å kaste to kron på rad?
4. Hva er sannsynligheten for å få 7 med to terninger?

X = alle biler (som kan komme rundt hjørnet)
A = alle rød biler (som kan komme rundt hjørnet)

For enhver endelig mengde Y, skriver vi |Y| for antall elementer i Y.

Sannsynligheten for at neste bil er rød: $P(A) = \frac{|A|}{|X|}$ (hvis vi ikke vet noe mer)

Mer generelt snakker vi om et sannsynlighetsmål for delmengder av X s.a.

1. $P(X) = 1$
2. $P(\emptyset) = 0$
3. Hvis $A \cap B = \emptyset$ så er $P(A \cup B) = P(A) + P(B)$

I det generelle tilfellet skal (3) også gjelde for tellbart uendelige sekvenser.
Vi vil stort sett se på det endelige tilfellet. Da vil vi bruke antallsmålet.

Hvis $A \cap B = \emptyset$, $A \cap C = \emptyset$ og $B \cap C = \emptyset$, så er $P(A \cup B \cup C) = P(A) + P(B) + P(C)$

1. Hva er sannsynligheten for at neste bil som kommer rundt hjørnet er rød?
2. Hva er sannsynligheten for å kaste kron?
3. Hva er sannsynligheten for å kaste to kron på rad?
4. Hva er sannsynligheten for å få 7 med to terninger?

Eks2: ½ = 50%

Eks3: $\frac{1}{2} * \frac{1}{2} = \frac{1}{4} = 25\%$

Eks4: Mulige utfall: 36, 7 i 6 av tilfellene, $\frac{6}{36} = \frac{1}{6}$

Flere begivenheter, uavhengighet

Tre størrelser på biler: små, mellomstore og store.
Hva er sannsynligheten for at neste bil er liten og rød?
Hva er sannsynligheten for at neste bil er liten hvis den er rød?
Hva er sannsynligheten for at neste bil er rød hvis den er liten?

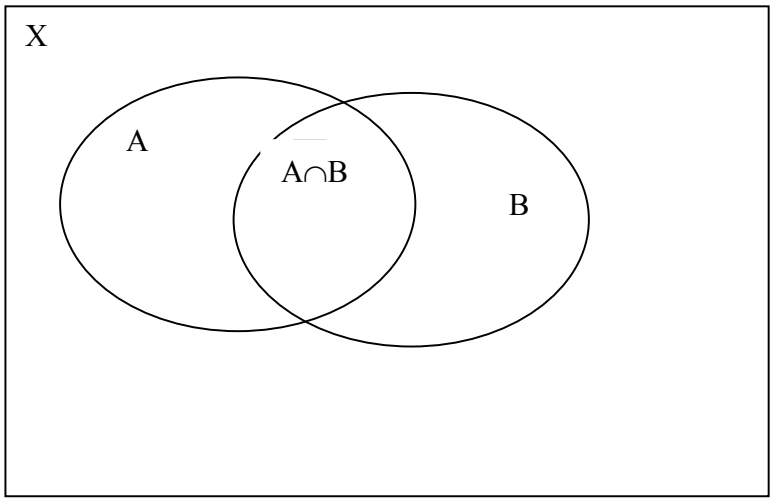
B = mengden av små biler (som kan komme rundt hjørnet)

Sannsynligheten for at bilen er liten og rød: $P(A \cap B) = \frac{|A \cap B|}{|X|}$

Hvis de to begivenhetene ikke har noe med hverandre å gjøre vil

$$P(A \cap B) = \frac{|A \cap B|}{|X|} = \frac{|A|}{|X|} * \frac{|B|}{|X|} = P(A) * P(B)$$

Dette er tilfelle hvis vi kaster to mynter, men neppe hvis vi ser på bilfarge.
Små biler er oftere rød, store biler er oftere sorte.



Mer generelt skriver vi P(A, B) for $P(A \cap B)$ ("the joint probability of A and B")

Betinget sannsynlighet

Sannsynligheten for at en liten bil er rød: $P(A \mid B) = \frac{|A \cap B|}{|B|}$

Merk at: $P(A \mid B) = \frac{|A \cap B| * |X|}{|B| * |X|} = \frac{|A \cap B|}{|X|} * \frac{|X|}{|B|} = \frac{P(A \cap B)}{P(B)}$

Bayes teorem

Vi ser at da er også $P(A \mid B)P(B) = P(A \cap B) = P(B \cap A) = P(B \mid A)P(A)$

Og $P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$

Dette er Bayes' teorem.
Bør læres!
Egentlig brukt for å oppdatere en kjent sannsynlighet P(A) i lys av ny viten om B.
Nyttig i NLP fordi det kan være lettere å beregne P(B|A) enn P(A|B)

Multiplikasjonsregelen

$$P(A \cap B) = P(A)P(B \mid A)$$

$$P(A \cap B \cap C) = P(A)P(B \mid A)P(C \mid A \cap B)$$

$$P(A \cap B \cap C \cap D) = P(A)P(B \mid A)P(C \mid A \cap B)P(D \mid A \cap B \cap C \cap D)$$

$$P(A1 \cap A2 \cap ... \cap An) = P(A1)P(A2 \mid A1)...P(An \mid A1 \cap A2... \cap An-1)$$

Et eksempel

Anta vi har en test for en bestemt type kreft.
Den er svært nyttig. Den finner 50% av alle svulster av denne typen.
Den har en liten feilkilde. Hvis en person ikke har kreft vil den i 99% si at vedkommende ikke har det, men i 1% av tilfellene vil den feilaktig si at de har det.
Hva er sannsynligheten for at du har kreft hvis testen er positiv?

La oss si at vi screener alle norske kvinner i en årgang: 40 000.
Anta at 100 av dem har denne krefttypen.

	Positiv test: B	Negativ test: X-B
Har kreft: A	50	50
Har ikke kreft: X-A	399	39501

P(B|A) = ½ = 50%

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{0.5 * (100/40000)}{(449/40000)} = \frac{0.5 * 100}{449} = ca0.1$$

Her vil P(A) være sannsynligheten for å ha sykdommen før en utfører testen, det som på engelsk kalles ”prior probability”.
Her er den 100/40000, dvs 1/400, eller 0,0025.

Sannynligheten etter å ha utført testen kalles ”the posterior probability”.
Selv om den er liten, har den økt betraktelig.

Litt statistikk

En random variabel er en funksjon fra et utvalgsrom til de reelle tallene.

Eksempel:
Utvalgsrom: Kast to terninger
Randomvariabel: X, Summen av de to terningene
Da kan vi i neste omgang spørre om sannsynligheten for ulike verdier:

$P(X=2) = 1/36$
$P(X=7) = 1/6$

Argmax-notasjon

Vi er ofte interessert i å finne den A som gir størst sannsynlighet gitt B.
Det skriver vi slik

$$\arg\max_A P(A \mid B)$$

Merk at $\arg\max_A P(A \mid B) = \arg\max_A \frac{P(B \mid A)P(A)}{P(B)} = \arg\max_A P(B \mid A)P(A)$

Litt matematisk notasjon

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + a_3 + \ldots + a_n$$

$$\sum_{i=1}^{7} i = 1 + 2 + 3 + 4 + 5 + 6 + 7 = 28$$

$$\sum_{i=2}^{5} i^2 = 4 + 9 + 16 + 25 = 54$$

$$\prod_{i=1}^{n} a_i = a_1 * a_2 * a_3 * \ldots * a_n$$

$$\prod_{i=1}^{7} i = 1 * 2 * 3 * 4 * 5 * 6 * 7 = 7! = 5040$$

Ordmeningsentydighetsgjøring (WSD)

regulært uttrykk: **"([(word="doserer" %c)]) ;"**

valg:

Antall treff: **4**

Resultatsider: **1**

AV03Un0202.936 Professor Emeritus Dag Østerberg **doserer** i klassisk byteori , og Knut Halvorsen fra Oslo Teknopol i synergieffektene mellom byer og forskningsmiljøer .

SK01NeJo01.2859 Her sitter jeg på en bar , tenkte Harry , og hører på en transvestitt som **doserer** om australsk politikk .

SK01OlPa01.3709 Han **doserer** teen etter tempoet i fortellingen hennes , ser ikke ned på den for å forvisse seg om at koppen er på rett kjøl , ser ufravendt på henne , henger ved hennes lepper - teen kommer av seg selv .

SK04RaAn02.250 Som **doserer** morfinen og holder en knivskarp balanse mellom bevissthet og smerteterskel .

Hvorfor greier vi å entydiggjøre disse?

Hva ser vi på?

Ord i omgivelsene.
Den rollen de spiller.

Vi ser på et vindu rundt ordet og hva som befinner seg der.

Hva er relevant?

- plassering?
- ordklasse, POS-tag?
- hvilke som helst ord, eller bare noen spesielle?

Kollokasjonstrekk

- lite vindu
- rekkfølge
- pos-tag

[*wi*−2,POS*i*−2,*wi*−1,POS*i*−1,*wi*+1, POS*i*+1,*wi*+2,POS*i*+2]

[guitar, NN, and, CC, player, NN, stand, VB]

"Bag-of words"

- større vindu
- ignorer plassering innenfor vinduet
- bruker noen termer som er overrepresentert i kontekster der det flertydige ordet forekommer

[*fishing, big, sound, player, fly, rod, pound, double, runs, playing, guitar, band*]

[0,0,0,1,0,0,0,0,0,0,1,0]

Kombinasjoner av de to teknikkene:

- en lang vektor, eller
- to klassifikatorer som kombineres

Naive Bayes-klassifikator

- w er et ord
- $\vec{f}$ er vektoren for en bestemt kontekst
- S er alle mulige meninger

Vi skal velge den mest sannsynlige kandidaten fra S gitt $\vec{f}$.

$$\hat{s} = \underset{s \in S}{\operatorname{argmax}} P(s|\vec{f})$$

Bayes formel:

$$\hat{s} = \underset{s \in S}{\operatorname{argmax}} \frac{P(\vec{f}|s)P(s)}{P(\vec{f})}$$

Alt for lite data: Vi har kanskje aldri sett $\vec{f}$ før.

Naiviteten

(Vi antar noe vi ikke vet om er riktig)

Vi antar at trekken er uavhengig av hverandre.
(Galt fordi *fiske*-trekk vil opptre sammen og *penge*-trekk vil opptre sammen.)

$$P(\vec{f}|s) \approx \prod_{j=1}^{n} P(f_j|s)$$

som gir

$$\hat{s} = \underset{s \in S}{\operatorname{argmax}} P(s) \prod_{j=1}^{n} P(f_j|s)$$

Vi får 20 størrelser å holde rede på i stedet for 2^{20}.

Trening

$$P(s_i) = \frac{\text{count}(s_i, w_j)}{\text{count}(w_j)}$$

og

$$P(f_j|s) = \frac{\text{count}(f_j, s)}{\text{count}(s)}$$

Dekoding

Regn ut

$$P(s)\prod_{j-1}^{n}P(f_j|s)$$

for hver s,
velg den som gir størst verdi.

Fordi dette blir å multiplisere sammen små tall er det fare for unøyaktigheter (”underflow”).
Derfor bruker vi logaritmer i stedet:

$$\log P(s) + \sum_{j=1}^{n} \log P(f_j|s)$$

Husk: log(a * b) = log(a) + log(b)

Glatting

Problem hvis en størrelse er null.
Erstatt

$$P(s_i) = \frac{count\,(s_i, w_j)}{count(w_j)}$$

med

$$P(s_i) = \frac{count\,(s_i, w_j) + 1}{count(w_j) + N}$$

der N er antall si-er
og

$$P(f_j|s) = \frac{count\,(f_j, s)}{count(s)}$$

med

$$P(f_j|s) = \frac{count\,(f_j, s) + 1}{count(s) + 2}$$

(Det er ikke entydig hva vi skal justere nevneren med. Vi har valgt 2 for to usette forekomster av s, en med fj og en uten fj. Andre fi-er spiller ikke inn her.)

Evaluering

Bunnlinje (base line):

- naivt 1/n, der *n* er antall betydninger
- mer vanlig: frekvensen til den hyppigste betydningen,

$$P(s_i) = \frac{\text{count}(s_i, w_j)}{\text{count}(w_j)}$$

- denne kan bli mer enn 0,5, jfr *banken*

Øvre grense:

- enighet mellom mennesker som blir satt til å gjøre samme oppgave

Pseudo-ord for evaluering:

banan-dør

Forskjell i vanskelighetsgrad

- Lettere: *bass* som fisk eller relatert til musikk (homonymer)
- Vanskeligere: de ulike musikkbetydningene (polysemer)
- pseudord har mye felles med homonymer