T-(538|725)-MALV, Málvinnsla
Mörkun - með tölfræði

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]Tölvunarfræðideild, Háskólinn í Reykjavík

September 2007

Outline

Outline

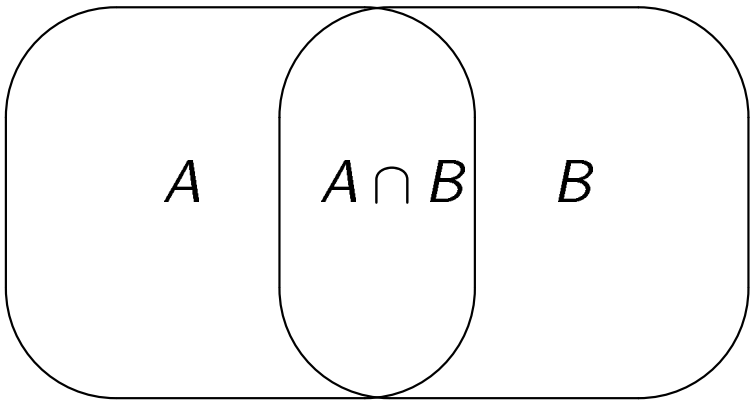
Skilyrtar líkur

Líkur á einum atburði að gefnum öðrum atburði.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

Bayes regla:

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A) \tag{2}$$



Líklegasta röð af mörkum

- Markaruna: $T = t_1, t_2, \ldots, t_n$
- Orðaruna: $W = w_1, w_2, \ldots, w_n$
- Viljum finna þá markarunu $\hat{\mathrm{T}}$ sem hámarkar: $P(T|W)$
  - $\hat{\mathrm{T}} = max_{t_1, t_2, \ldots t_n}\ P(t_1, t_2, \ldots t_n | w_1, w_2, \ldots w_n)$
- Skv. Bayes: $\mathrm{P(W)P(T|W) = P(T)P(W|T)}$
- $$\hat{\mathrm{T}} = \arg\max_T \quad \frac{P(T)P(W|T)}{P(W)} \tag{3}$$
- Fyrir tiltekna orðarunu er P(W) fasti, $\Rightarrow$

$$\hat{\mathrm{T}} = \arg\max_T \quad P(T)P(W|T) \tag{4}$$

Þrístæðu nálgun fyrir markarununa

$$P(T) = P(t_1, t_2, \ldots, t_n) \approx P(t_1)P(t_2|t_1)\prod_{i=3}^{n} P(t_i|t_{i-2}, t_{i-1}) \quad (5)$$

Þessar líkur eru nálgaðar með **sennileikalíkunum** (MLE):

$$P_{MLE}(t_i|t_{i-2}, t_{i-1}) = \frac{C(t_{i-2}, t_{i-1}, t_i)}{C(t_{i-2}, t_{i-1})} \quad (6)$$

Þjálfunarmálheildir eru ekki nógu stórar $\Rightarrow$ tíðnitölur vantar. Þá oft notuð línuleg brúun (e. linear interpolation)

$$P_{LinearInter}(t_i|t_{i-2}, t_{i-1}) = \lambda_1 P(t_i|t_{i-2}, t_{i-1}) + \lambda_2 P(t_i|t_{i-1}) + \lambda_3 P(t_i) \quad (7)$$

þar sem $0 \leq \lambda_i \leq 1$ og $\sum_{i=1}^{3} \lambda_i = 1$

Nálgun fyrir P(W|T)

$$P(W|T) = P(w_1, w_2, \ldots, w_n|t_1, t_2, \ldots, t_n) \approx \prod_{i=1}^{n} P(w_i|t_i) \quad (8)$$

Sérhvert $P(w_i|t_i)$ er síðan nálgað með sennileikalíkum:

$$P_{MLE}(w_i|t_i) = \frac{C(w_i, t_i)}{C(t_i)} \quad (9)$$

Endanleg formúla

Þessi formúla:

$$\hat{\mathrm{T}} = \arg\max_T \quad P(T)P(W|T) \tag{10}$$

. . . var einfölduð og útkoman varð markaröðin $\hat{\mathrm{T}}$ sem hámarkar:

$$P(t_1)P(t_2|t_1)\prod_{i=3}^{n} P(t_i|t_{i-2},t_{i-1})\prod_{i=1}^{n} P(w_i|t_i) \tag{11}$$

Þetta er **þrístæðulíkan**.

Outline

Markov líkan

- Notað til að lýsa röð af slembibreytum sem eru að einhverju leyti háðar.
- T.d. þegar gildi slembibreytna er háð gildi á fyrri slembibreytum í röðinni.

Markov keðja (e. Markov chain)

- Röð slembibreytna $\{X_1, X_2, \ldots, X_T\}$
- Fá gildi sitt úr endanlegum stöðumengi $\{q_1, q_2, \ldots, q_n\}$
- Skilgreinir slembiumskipti (e. random transition) úr einni stöðu í aðra.

Markov líkan

Markov skilyrðið

- **Limited history**. Núverandi staða (gildi) er eingöngu háð föstum fjölda af undanfarandi stöðum.
  - T.d. "First order model":

$$P(X_t = q_j | X_1, X_2, \ldots, X_{t-1}) = P(X_t = q_j | X_{t-1})$$

  - T.d. "Second order model":

$$P(X_t = q_j | X_1, X_2, \ldots, X_{t-1}) = P(X_t = q_j | X_{t-2}, X_{t-1})$$

- **Independent of time; stationary**.
  - T.d. "First order model":

$$P(X_t = q_j | X_{t-1} = q_i) = a_{ij}$$

Markov líkan og stöðuvélar

- Líta má á Markov líkan sem stöðuvél með færslulíkum (a_{ij}) á hverjum legg (sjá Fig. 7.3. bls 168).
- Það eru engin "long distance dependencies" og næsta staða er eingöngu háð núverandi stöðu.
- Í "Visible Markov model" er vitað í hvaða stöður vélin fer í gegnum.

$P(X_1, X_2, \ldots, X_T) = P(X_1) * P(X_2|X_1) * P(X_3|X_2, X_1), \ldots,$
$P(X_T|X_1, X_2, \ldots, X_{T-1})) =$
$P(X_1) * P(X_2|X_1) * P(X_3|X_2), \ldots, P(X_T|X_{T-1})$
(limited history; first order)

Hidden Markov Model (HMM)

- Í tilviki mörkunar getum við litið svo á að stöður í líkaninu standi fyrir mörk.
- Röð staða sem líkanið fer í gegnum er ekki þekkt (hidden).
- En hvað með orðin sjálf?
  - Sérhver staða "sendir frá sér" (e. emits) orð með tilteknum frálagslíkum.
- Markov líkanið er með undirliggjandi falin mörk (stöður) sem búa til orð með tilteknum líkum.
- P(T): færslulíkur (e. transition probabilities)
- P(W|T): frálagslíkur (e. emission probabilities)

Hidden Markov Model (HMM)

- Lítum sem sagt á röð marka í texta sem Markov líkan.
- Við gerum ráð fyrir að mark tiltekins orðs sé eingöngu háð undanfarandi tveimur/þremur mörkum (limited horizon)
- Og það breytist ekki eftir því sem tíminn líður (e. stationary)
  - T.d. ef líkurnar á því að sögn komi á eftir fornafni í upphafi setningar eru 0,2 þá breytast ekki þessar líkur þegar afgangurinn af setningunni er markaður (eða nýjar setningar).
- Markmiðið er að finna líklegustu röðina af mörkum fyrir röð af orðum.
- Þ.e. að finna líklegustu röðina af stöðum sem líkanið fer í gegnum fyrir röð af orðum.

Outline

Hvernig finnum við líklegustu markarununa?

$$P(t_1)P(t_2|t_1)\prod_{i=3}^{n} P(t_i|t_{i-2},t_{i-1})\prod_{i=1}^{n} P(w_i|t_i)$$

Reikna út líkur fyrir alla möguleika á $t_1, t_2, \ldots, t_n$?

- Óskilvirkt, því möguleikarnir geta verið mjög margir.
- Tímaflækjan er (að hámarki) n^T, þar sem T er fjöldi marka í markamenginu.

Því er notuð kvik bestun (e. dynamic programming) – **Viterbi** algrímið.

Viterbi algrímið

- Í stað þess að þræða allar mögulegar leiðir í Markov líkaninu til að reikna út líklegustu markarununa þá ...
- Reiknar Viterbi líklegustu "hlutleiðir" (e. subpaths) fyrir sérhverja stöðu í stöðuvélinni um leið og ferðast er um í henni.
- Þeim leiðum sem eru ekki líklegastar er hafnað.
- Sjá Fig. 7.2 bls. 166.

Outline

Þrístæðumarkari

- Byggir á HMM.
- Stöður standa fyrir pör af mörkum.
- Færslulíkur: Líkur á tilteknum marki að gefnum tveimur undanfarandi mörkum.
  - e. transition/contextual probabilities
- Frálagslíkur: Líkur á tilteknum orði að gefnu tilteknu marki.
  - e. emission/lexical probabilities
- Líkur eru metnar út frá mörkuðu þjálfunarsafni (sennileikalíkur)
- Markarunan $t_1, t_2, \ldots, t_n$ er fundin sem hámarkar:

$$P(t_1)P(t_2|t_1)\prod_{i=3}^{n} P(t_i|t_{i-2}, t_{i-1})\prod_{i=1}^{n} P(w_i|t_i)$$

Þrístæðumarkari - Óþekkt orð

- Hvernig eru óþekkt orð meðhöndluð, þ.e. orð sem ekki koma fyrir í þjálfunarsafninu?
- Ein leið er að framkvæma greiningu á endingum.
- Líkindadreifing fyrir tiltekna endingu er búin til með því að nota orð í þjálfunarsafninu sem hafa sameiginlega endingu af einhverri hámarkslengd (t.d. 10)
- $P(w_i|t_i)$ verður þá $P(e_i|t_i)$, þar sem e_i stendur fyrir endingu á orði i af tiltekinni lengd.

Þrístæðumarkari - TnT

- TnT - Trigrams 'n Tags
- Brants 2000.
- Skilvirkur og markvirkur tölfræðimarkari sem byggir á HMM líkani.
- Útfærður á þann máta sem við höfum fjallað um.