4 Markovin piilomallit

Edellisestä sääntöpohjaisesta menetelmästä poiketen useimmat nykyiset sanaluokkien tunnistajat perustuvat tilastollisiin menetelmiin. Tilastollisissa menetelmissä sanaluokkien tunnistaminen mielletään lingvistisen lähestymistavan sijaan koneoppimisongelmaksi, tai tarkemmin sarjanluokitteluongelmaksi. Sarjanluokitteluongelmassa tavoitteena on oppia funktio $f : \mathcal{X} \to \mathcal{Y}$, joka luokittelee kunkin syötteen x johonkin luokkaan y. Todennäköisyyslaskennan kautta funktio f voidaan määritellä muodossa

$$f(x) = \arg\max_{y \in \mathcal{Y}} P(y|x) \tag{4.1}$$

missä todennäköisyyttä $P(y|x)$ estimoidaan harjoitusaineiston perusteella.

Tilastolliset mallit voidaan vuorostaan jakaa diskriminatiivisiin ja generatiivisiin malleihin. Diskriminatiiviset mallit käsittelevät suoraan todennäköisyyttä $P(y|x)$, eli ne eivät ota kantaa syötteeseen x. Generatiiviset mallit puolestaan käsittelevät koko yhteisjakaumaa $P(x, y)$, josta todennäköisyys $P(y|x)$ johdetaan Bayesin säännön avulla. Diskriminatiiviset mallit ovat siten generatiivisia malleja rajoittuneempia, mutta sarjanluokitteluongelman kannalta mallin rajoitteilla ei ole väliä. Intuition mukaan generatiivinen malli on diskriminatiivista mallia tehottomampi, sillä yhteisjakauman mallintaminen on laskennallisesti vaativampaa kuin ehdollisen jakauman. Ng & Jordan (2002) kuitenkin osoittavat, ettei tämä välttämättä aina pidä paikkaansa: sanaluokkien tunnistamiseen onkin olemassa kumpaankin malliin perustuvia tehokkaita ratkaisuja. Tässä tutkielmassa esitellään generatiivinen Markovin piilomalli sekä diskriminatiivinen log-lineaarinen malli.

4.1 Markovin malli

Markovin malli (mm. Rabiner, 1989) kuvaa sellaista stokastista prosessia, joka toteuttaa ns. Markov-ominaisuuden: seuraava tila riippuu aina vain N:stä edeltävästä tilasta. Markov-ominaisuus on siis eräänlainen riippumattomuusoletus, joka yksinkertaistaa stokastisen prosessin tilan estimointia rajoittamalla tilasiirtymien historian määrää. N:nen asteen Markov-ominaisuuden toimiessa esimerkiksi todennäköisyys

$$P(x_k|x_1,\ldots,x_{k-1}) \tag{4.2}$$

voidaan laskea huomattavasti yksinkertaisemmin tarkastelemalla vain N:ää edellistä tilaa:

$$P(x_k|x_{k-N},\ldots,x_{k-1}) \tag{4.3}$$

Markov-ominaisuudesta puhuttaessa on yleensä kyse juuri ensimmäisen asteen Markov-ominaisuudesta, jolloin $N = 1$. Käytännössä tämä tarkoittaa sitä, että tarkastellaan jotakin kahta peräkkäistä tilaa — nykyistä sekä tulevaa. Markov-ominaisuutta voidaan kuitenkin laajentaa myös korkeampiin asteisiin, jolloin myös tarkasteltavien tilasarjojen pituudet kasvavat. Sanaluokkia tunnistaessa näitä tilasarjoja vastaavat n:n peräkkäisen sanaluokan sarjat, eli ns. n-grammit.

Yksinkertaisimmillaan Markovin mallia voidaan kuvata tilakoneena, joka koostuu havaittavia tapahtumia kuvaavista tiloista, sekä tilasiirtymämatriisista, josta ilmenevät todennäköisyydet siirtyä kustakin tilasta mihin tahansa muuhun tilaan. Kukin tila on siis itsenäinen ja muistiton.

4.1.1 Markovin piilomalli

Markovin mallin hyödyllisyyttä rajoittaa se, että malli ei pysty itsessään mallintamaan prosesseja, joiden tilat eivät ole suoraan havaittavissa. Useimmissa mielen-

kiintoisissa tapauksissa prosessin tilat eivät kuitenkaan suoraan vastaa havaintoja, eli prosessin tila — vaikkakin havainnosta riippuvainen — on piilotettu: esimerkiksi sanaluokkien tunnistamisongelmassa havainto (sana) on tiedossa, tila (sanaluokka) on riippuvainen havainnosta, mutta tila itsessään ei ole tiedossa. Tällöin Markovin malli tulee laajentaa Markovin piilomalliksi, jossa havainto on aina sitä vastaavan tilan todennäköisyysfunktio.