Szekvenciatanulás

A szekvenciatanulás (sequence labeling, stuctured prediction) problémája abban tér el a klasszikus osztályozási feladattól, hogy itt nem egyetlen szeparált egyed címkéjének előrejelzésére építünk statisztikai modellt, hanem egyidőben egyedek egy sorozatára. Természetesen ebben az esetben nem élhetünk az egyedek közötti függetlenség feltevésével sem.

Az információkinyerés problémáját gyakran fogalmazzák meg szekvenciatanulásként. Tekintsük példaként a tulajdonnév-felismerést. Ekkor a cél szavak egy sorozatának (általában egy mondatnak) felcímkézése tulajdonnév kategóriákkal. A kimenet is tehát címkék szekvenciája lesz:

> *Sólyom*$_{\mathrm{PER}}$ *László*$_{\mathrm{PER}}$ *Magyarország*$_{\mathrm{LOC}}$ *köztársasági*$_{\mathrm{NON}}$ *elnöke*$_{\mathrm{NON}}$ *az*$_{\mathrm{NON}}$ *MTV-nek*$_{\mathrm{ORG}}$ *elmondta*$_{\mathrm{NON}}$ ...

Alább bemutatjuk a három legismertebb szekvencia alapú gépi tanulási modellt.

Rejtett Markov-modell

A legkorábbi szekvencia alapú tanuló algoritmus a *rejtett Markov-modell* (hidden Markov model, HMM, Manning & Schütze, 1999) . A modell alaptevései, hogy a rendszer állapotai generálják a megfigyeléseket (pl. szavak) és hogy egy elsőrendű Markov-modellben a rendszer állapota a t időpontban (y_t) csak a megelőző állapottól függ, azaz

$$P(y_t|y_{t-1},y_{t-2},\dots,y_1) = P(y_t|y_{t-1})$$

valamint a t időpontbeli megfigyelés (x_t) csak az aktuális y_t állapottól függ. A modell az állapotsorozat és megfigyeléssorozat együttes valószínűségét becsli, célja a legvalószínűbb állapotsorozat megtalálása az adott megfigyeléssorozathoz (a gyakorlatban általában a rendszer állapotait egyértelműen megfeleltetik a cél címkéknek). A rejtett Markov-modell három típusú valószínűséget használ:

- Az átmeneti valószínűségek $P(y_t|y_{t-1})$ hivatottak jelezni az egymást követő állapotok közti átmenetek valószínűségeit, tehát például a $P(\mathrm{ORG}|\mathrm{LOC})$ feltételes valószínűség jelöli, hogy milyen eséllyel követ egy földrajzi kifejezést egy szervezetre vonatkozó kifejezés.
- A kezdő állapot eloszlása $P(y_1)$.
- Az emissziós eloszlások $P(x_t|y_t)$ azt írják le, hogy az egyes állapotok milyen valószínűséggel generálják az egyes megfigyeléseket.

Ezek felhasználásával a becsülendő együttes eloszlás:

$$P(\mathbf{x},\mathbf{y}) = P(y_1)P(x_1|y_1)\prod_{2}^{T} P(x_t|y_{t-1})P(x_t|y_t)$$

Ezeket az eloszlásokat egy tanító halmazból a *Baum–Welsh-eljárás* (egy Expectation-Maximization módszer) segítségével becsülhetjük. Ezen becslések ismeretében és adott szekvencia esetén a legvalószínűbb utat (állapotszekvenciát) a *Viterbi algoritmussal* találhatjuk meg. Ez az algoritmus dinamikus programozáson alapul, lényegében a t-ik legvalószínűbb állapotot az alapján választjuk meg, hogy a $(t-1)$.ik állapotokba milyen eséllyel kerülhet a rendszer. A Baum–Welsh- és Viterbi-algoritmusok részletesebb leírása megtalálható például Manning & Schütze, 1999 és Rabiner, 1990 munkákban.

A rejtett Markov-modellekkel kapcsolatban két problémát kell tisztán látnunk:

- Az emissziós eloszlásoknál szükséges $P(x)$ becslése. Természetesen maga az x megfigyelés jellemzők egy halmaza (például a szó nagybetűvel kezdődik-e). A rejtett Markov-modell él a naív Bayes feltevéssel, azaz felteszi az egyes jellemzők függetlenségét. Nyilvánvaló, hogy a legtöbb esetben a jellemzők nem függetlenek (példánknál maradva: egy szó kezdőbetűje és a megelőző szó nem független események) ami különösen megbízhatatlanná teszi ezeket a modelleket.

- A tanuló algoritmus tulajdonképpen a megfigyeléssorozat valószínűségében maximalizál, holott a cél a legvalószínűbb állapot-(címke) sorozat megtalálása.

Maximum entrópia Markov-modell

A maximum entrópia Markov-modell (maximum entropy Markov model, MEMM, McCallum et al, 2000) nem tételez fel függetlenséget az egyes — megfigyeléseket leíró — jellemzők között. Alapja egy osztályozó eljárás a *maximum entrópia* (más néven logisztikus regresszió) módszer ami a osztálycímkék eloszlását, mint függvényt közelíti. Alapfeltevése, hogy a feltételes valószínűség logaritmusa lineáris modellel leírható:

$$P(y|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\exp\{\sum_{j=1}^{K}\lambda_{y,j}x_j\},$$

ahol K darab x_j jellemző van, azok súlya a $\lambda_{y,j}$ paraméterek, valamint $Z(\mathbf{x})$ egy normalizációs faktor:

$$Z(\mathbf{x}) = \sum_y \exp\{\sum_{j=1}^{K} \lambda_{y,j} x_j\}$$

A maximum entrópia Markov-modell nem használ külön átmeneti és emissziós valószínűségeket, hanem egyetlen feltételes valószínűségbe vonja ezeket össze: mekkora az egyes címkék valószínűsége a megelőző címke és a megfigyelt jellemzővektor függvényében. Ezt az eloszlást becsli az exponenciális modell segítségével:

$$P(y_t|y_{t-1},\mathbf{x}) = \frac{1}{Z(y_{t-1},\mathbf{x})} \exp\{\sum_{j=1}^{K} \lambda_j f_j(x, y_t, y_{t-1})\},$$

ahol f_j egy jelölési egyszerűsítés az $f(y_t, y_{t-1})$ és $f(y_t, x_t)$ jellemzőfüggvények megadására. A tanulás itt tehát az egyes jellemzők súlyának ($\lambda_{y,j}$) a megtanulása. A becsült $P(y_t|y_{t-1},\mathbf{x})$ felhasználásával azután a Viterbi-algoritmushoz hasonló módon kiszámíthatjuk a legvalószínűbb címkesorozatot.

A MEMM tanítása csak némileg komplexebb, mint a HMM-é, viszont empirikusan bizonyított, hogy a legtöbb probléma esetén — a jellemzők közötti összefüggések kiaknázásának köszönhetően — sokkal hatékonyabb.

A MEMM problémája az, hogy a tanítás folyamán megelőző címkeként az etalon (gold standard) címkéket használja, a predikciós fázisban azonban az előző címke nem biztos, hogy helyes, így ez a „jellemző” igen zajossá válhat. Ez a probléma a gyökere az irodalomban *label bias probléma*ként (Lafferty et al, 2001) emlegetett jelenségnek.

Feltételes valószínűségi mezők

A HMM és a MEMM is lokális eloszlásbecsléseket hajtanak végre, majd a Viterbi-algoritmus segítségével kiválasztják a legvalószínűbb utat. Ezzel szemben a label bias problémára megoldást kínáló *feltételes valószínűségi mező* (conditional random fields, CRF, Lafferty et al, 2001) valóban az egész struktúra előrejelzését végzi el. A CRF nem $P(y_t|x_t)$ jellegű lokális valószínűségeket becsül, hanem az egész szekvencia feltételes valószínűségét:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\{\sum_t \sum_{j=1}^{K} \lambda_j f_j(x, y_t, y_{t-1})\}$$

Figyeljük meg, hogy a normalizációs faktor itt már csak a megfigyelés függvénye, így nem jelentkezhet a label bias probléma. A rendszer tanítása (λ_j becslése) általában a logaritmikus feltételes valószínűség maximalizálásával történik:

$$\max_{\lambda} \ell(\lambda) = \max \sum_{i=1}^{N} p(y^{(i)}|x^{(i)}),$$

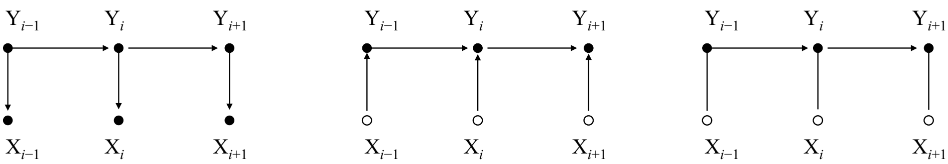
ahol a tanító halmaz N darab $x^{(i)}$ megfigyelés-szekvenciából és $y^{(i)}$ állapotszekvenciából áll. Mivel a $\ell(\lambda)$ függvény $g(x) = log(\sum_i \exp x_i))$ alakú így szigorúan konkáv is. A CRF optimalizálást általában valamilyen kvázi-Newton-módszerrel (például BFGS (Byrd et al, 1995)) szokták elvégezni.

A CRF egyetlen nagy hátránya a tanítás időigénye, ami égyzetesen függ a osztálycímkék számától és lineárisan a tanító példák számától valamint az átlagos szekvenciahossztól. Azokra a problémákra ahol ez az idő elfogadható nagyságrendű napjaink legsikeresebb rendszerei CRF-et használnak.

A fent bemutatott CRF modell (linear-chain CRF) a legegyszerűbb feltételes valószínűségi mezőn alapul ahol y_t csak y_{t-1} és y_{t+1}-től függ. Elméletileg a modell tetszőleges valószínűségi mező felett értelmezhető. Ekkor lehetővé válik például távoli egyedek közötti (pl. mondatokon átívelő) összefüggések leírása is. Ezeknek az *általános* CRF-eknek (Sutton & McCallum, 2007) az effektív tanítása még nyitott kutatási kérdés, gyakorlati problémák megoldására (egy-két speciális esettől eltekintve) nem alkalmazhatóak.

A három modell összehasonlítása

Összefoglalásként tekintsük át a három bemutatott modellt azok gráfos ábrázolásán keresztül (1. ábra).



1. ábra. A HMM-, a MEMM- és a lánc-CRF-modellek gráfos reprezentációja (balról jobbra). Az üres pont azt jelöli, hogy a változót nem a modell generálja (Forrás: Lafferty et al, 2001.)

A HMM alapfeltevése, hogy a megfigyeléseket az állapotok generálják. Legnagyobb hátránya, hogy a megfigyeléseket leíró jellemzőket egymástól függetlennek tételezi fel. A MEMM nem igényli a $P(\mathbf{x})$ kiszámítását, így a jellemzők közötti összefüggések modellezését kikerüli. A modellben az egyes címkék eloszlását egy exponenciális modellel írhatjuk le, ahol tulajdonképpen a jellemzők sokdimenziós terében egy függvény-approximációt hajtunk végre a tanítás folyamán.

A MEMM már adottnak tekinti a megfigyeléssorozatot ellentétben a HMM-el) és célja az optimális állapotszekvencia megtalálása (középső ábra) azonban magával hozza a label bias problémáját. Ennek kiküszöbölésére a CRF modell a lokális valószínűség-becslések helyett az egész állapotszekvencia feltételes valószínűségében optimalizál. Ez megteremti a kapcsolatot a t időpontbeli címke és a későbbi ($t <$) megfigyelések közt is (nem irányított a modell).

Forrás:

R. H. Byrd, P. Lu, J. Nocedal, and C. Y. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(6):1190–1208, 1995.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML-01, 18th Int. Conf. on Machine Learning*, pages 282–289, Williamstown, USA, 2001.

Ch. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, 1999.

A. McCallum, D. Freitag, and F. C. N. Pereira. Maximum entropy Markov models for information extraction and segmentation. In *Proc. of ICML-00, 17th Int. Conf. on Machine Learning*, pages 591–598, Stanford, USA, 2000.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In A. Waibel and K.-F. Lee, editors, *Readings in Speech Recognition*, pages 267–296. Kaufmann, San Mateo, CA, 1990.

Ch. Sutton and A. McCallum. An introduction to conditional random fields for relational learning. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press, 2007. To appear.