Επεξεργασία Φυσικής Γλώσσας & Μηχανική Μάθηση

Μάθηση Βασισμένη στη Μνήμη στην ΕΦΓ

Κάτια Κερμανίδου
kerman@ionio.gr

Μάθηση Βασισμένη στη Μνήμη (Memory-based Learning)

- Λέγεται και
  - Instance-based learning
  - Lazy learning
- Τα δεδομένα εκπαίδευσης διατηρούνται αυτούσια...
  - ...σε αντίθεση με τις άλλες μεθόδους μηχανικής μάθησης οι οποίες κωδικοποιούν τα παραδείγματα εκπαίδευσης σε μια συμπαγή περιγραφή.
- Όταν ένα τέτοιο σύστημα κληθεί να αποφασίσει για την κατηγορία μιας νέας περίπτωσης, εξετάζει εκείνη τη στιγμή τη σχέση της με τα ήδη αποθηκευμένα παραδείγματα.
- Κάνει την παραδοχή ότι τα διάφορα παραδείγματα μπορεί να αναπαρασταθούν ως σημεία σε κάποιον n-διάστατο Ευκλείδειο χώρο Rn όπου n ο αριθμός των χαρακτηριστικών (ανεξάρτητων μεταβλητών).
- Κάθε νέα περίπτωση τοποθετείται στο χώρο αυτό ως νέο σημείο και η τιμή του η προσδιορίζεται με βάση το χαρακτηρισμό των k γειτονικών σημείων.

Ο αλγόριθμος των Πλησιέστερων Γειτόνων
The Nearest Neighbors Algorithm

- Η πιο απλή υλοποίηση του αλγορίθμου:
- Τα παραδείγματα εκπαίδευσης σαρώνονται γραμμικά, και υπολογίζεται η απόσταση του καθενός από το παράδειγμα ελέγχου.
- 1-NN: Όποιο παράδειγμα εκπαίδευσης έχει την μικρότερη απόσταση από το παράδειγμα ελέγχου, είναι ο πιο κοντινός γείτονας.
  - Το παράδειγμα ελέγχου παίρνει σαν τιμή στην κλάση ταξινόμησης ό,τι τιμή έχει στην κλάση ταξινόμησης ο πιο κοντινός γείτονας
- k-NN: Τα k παραδείγματα εκπαίδευσης που έχουν την μικρότερη απόσταση από το παράδειγμα ελέγχου, είναι οι k πιο κοντινοί γείτονες.
  - Το παράδειγμα ελέγχου παίρνει σαν τιμή στην κλάση ταξινόμησης ό,τι τιμή έχει στην κλάση ταξινόμησης η πλειοψηφία των k πιο κοντινών γειτόνων **(majority voting)**

Πλεονεκτήματα-Μειονεκτήματα

- Ο ταξινομητής μπορεί να επικαιροποιηθεί (επεκταθεί), δηλ. να εκπαιδευτεί σε περισσότερα δεδομένα, με την απλή προσθήκη καινούριων παραδειγμάτων εκπαίδευσης.
- Σε πολλές περιπτώσεις είναι πολύ ακριβής η ταξινόμηση
  - Εξαιρέσεις
- Δεν πραγματοποιείται εκπαίδευση (δεν επάγεται κάποιο γενικευμένο μοντέλο)
- Αφιερώνει πολύ χρόνο για τον υπολογισμό των αποστάσεων
  - Η χρονική πολυπλοκότητα είναι ανάλογη του γινομένου του αριθμού των χαρακτηριστικών επί τον αριθμό των παραδειγμάτων εκπαίδευσης
- Θεωρεί ότι όλα τα χαρακτηριστικά είναι ισοβαρή (έχουν την ίδια σημασία για την μάθηση της έννοιας)
  - Αυτό μπορεί να αντιμετωπιστεί με απόδοση βαρών στα χαρακτηριστικά (weighting)

Παραλλαγή του k-NN (Memory-based shallow parsing, Daelemans, Buchholz, Veenstra, 1999)

- IB1-IG
  - Κάθε χαρακτηριστικό έχει διαφορετικό βάρος – το Κέρδος Πληροφορίας (Information Gain)
- Αναγνώριση υποκειμένου-αντικειμένου
  - Κάθε ουσιαστικό αποτελεί ένα παράδειγμα μάθησης (πιθανό υποκείμενο ή αντικείμενο)
  - Έξοδος: υποκείμενο, αντικείμενο, άλλο
  - Χαρακτηριστικά
    - Απόσταση (σε αριθμό φράσεων) ανάμεσα στο ρήμα και το ουσιαστικό
    - Αριθμός άλλων ρηματικών φράσεων που παρεμβάλλονται ανάμεσα στο ρήμα και το ουσιαστικό
    - Αριθμός κομμάτων που υπάρχουν ανάμεσα στο ρήμα και το ουσιαστικό
    - Το ρήμα καθεαυτό και το μέρος του λόγου του
    - Οι δυο λέξεις που προηγούνται του ουσιαστικού και τα μέρη του λόγου τους
    - Η λέξη καθεαυτή του ουσιαστικού
    - Η μια λέξη που έπεται του ουσιαστικού και το μέρος του λόγου της

Memory-based shallow parsing, Daelemans, Buchholz, Veenstra, 1999

[**NP** *My*/PRP$ *sisters*/NNS **NP**] [**VP** *have*/VBP *not*/RB *seen*/VBN **VP**] [**NP** *the*/DT *old*/JJ *man*/NN **NP**] *lately*/RB ./.

Feature	1	2	3	4	5	6	7	8	9	10	11	12	13	Class
Weight	39	40	4	3	2	10	12	18	29	18	31	13	24	
Inst.1	-1	0	0	seen	VBN	-	-	-	-	sisters	PRP$	seen	VBN	S
Inst.2	1	0	0	seen	VBN	sisters	PRP$	seen	VBN	man	NN	lately	RB	O
Inst.3	2	0	0	seen	VNB	seen	VBN	man	NN	lately	RB	.	.	-

Memory-based shallow parsing, Daelemans, Buchholz, Veenstra, 1999

	Together				Subjects			Objects		
# relations	51629				32755			18874		
Method	acc.	prec.	rec.	$F_{\beta=1}$	prec.	rec.	$F_{\beta=1}$	prec.	rec.	$F_{\beta=1}$
Random baseline		3.9	3.9	3.9	4.5	4.5	4.5	2.7	2.5	2.6
Heuristic baseline		65.9	66.5	66.2	69.3	61.6	65.2	61.6	75.1	67.7
IGTree	96.9	79.5	73.2	76.2	80.9	71.4	75.8	77.2	76.4	76.8
IB1-IG	96.6	74.4	**76.9**	75.6	76.2	**76.9**	76.5	71.5	**76.7**	74.0
IGTree & IB1-IG unanimous	97.4	**89.8**	68.6	**77.8**	**89.7**	67.6	**77.1**	**89.8**	70.4	**79.0**

Table 3: Results of the 10-fold cross validation experiment on the subject–verb/object–verb relations data. We trained one classifier to detect subjects as well as objects. Its performance can be found in the column *Together*. For expository reasons, we also mention how well this classifier performs when computing precision and recall for subjects and objects separately.

Random baseline: δίνοντας τυχαία τιμή, βάσει της κατανομής των τιμών της εξόδου

Heuristic baseline: Θεωρώντας όλα τα ουσιαστικά που προηγούνται του ρήματος υποκείμενα, και όλα αυτά που έπονται αντικείμενα.