Επεξεργασία Φυσικής Γλώσσας & Μηχανική Μάθηση

Πιθανοτικά Μοντέλα στην ΕΦΓ

Κάτια Κερμανίδου
kerman@ionio.gr

Naïve Bayes

Πιθανοτική Ταξινόμηση

- Στόχος της πιθανοτικής ταξινόμησης είναι η εκτίμηση της πιθανότητας P(c|d), δηλ. της πιθανότητας στο κείμενο d να αποδίδεται η έξοδος c.

$$P(c|\vec{d}_j) = \frac{P(c)P(\vec{d}_j|c)}{P(\vec{d}_j)}$$

- Μεγάλο υπολογιστικό κόστος, μια και το d απαρτίζεται από n τυχαίες μεταβλητές (όσος ο αριθμός των χαρακτηριστικών αναπαράστασης των κειμένων)
- Παραδοχή ανεξαρτησίας (Conditional Independence Assumption)
  - Τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους δεδομένης της τιμής της εξόδου

$$P(\vec{d}|c) = \prod_{k=1}^{|T|} P(t_k|c)$$

Σπάνια/Ανύπαρκτα φαινόμενα (Sparse data)

- Τι γίνεται αν στα παραδείγματα εκπαίδευσης δεν εμφανίζεται παράδειγμα με τιμή c_j στην έξοδο, και τιμή a_i στο χαρακτηριστικό i:
- Τότε $\hat{P}(a_i|c_j) = 0$, $\hat{P}(c_j)\prod_i \hat{P}(a_i|c_j) = 0$
- Εξομάλυνση

$$\hat{P}(a_i|c_j) \leftarrow \frac{n_c + mp}{n + m}$$

- n is number of training examples for which $C = c_j$,
- n_c number of examples for which $C = c_j$ and $A = a_i$
- p is prior estimate for $\hat{P}(a_i|c_j)$
- m is weight given to prior (i.e. number of "virtual" examples)

Συνεχή Χαρακτηριστικά

- $P(x|C) = \frac{1}{\sigma_c\sqrt{2\pi}} e^{-\frac{(x-\mu_c)^2}{2\sigma_c^2}}$

- όπου, σ_c η τυπική απόκλιση των τιμών του αριθμητικού χαρακτηριστικού για τιμή c της εξόδου
- μ_c η μέση τιμή των τιμών του αριθμητικού χαρακτηριστικού για τιμή c της εξόδου
- x η τιμή του χαρακτηριστικού στο παράδειγμα ταξινόμησης

Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques, Shweta Rana ; Archana Singh

- https://ieeexplore.ieee.org/abstract/document/7877399
- 1000 θετικά και 1000 αρνητικά σχόλια χρηστών για ταινίες http://www.cs.comell.edu/individuals/pabol-movie-review-data/
- Tokenization
  - Διάσπαση των κειμένων σε λέξεις (tokens)
  - Επιλογή tokens από 4 έως 25 χαρακτήρες
- Porter Stemming
  - connect, connects, connecting, connected, connection -> connect
- Αφαίρεση λειτουργικών λέξεων
  - 'a', 'the' κλπ

Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques, Shweta Rana ; Archana Singh

- Αποτελέσματα με τον Naïve Bayes

Action Movies			
	true actionpos	true actionneg	class precision
Pred.actionpos	9	1	90.00%
Pred.actionneg	11	19	63.33%
class recall	45.00%	95.00%	

Adventure Movies			
	true adventurepos	true adventureneg	class precision
Pred.adventurepos	14	3	66.67%
Pred.adventureneg	6	13	68.42%
class recall	70.00%	65.00%	

Drama Movies			
	true dramapos	true dramaneg	class precision
Pred.dramapos	16	4	80.00%
Pred.dramaneg	4	16	80.00%
class recall	80.00%	80.00%	

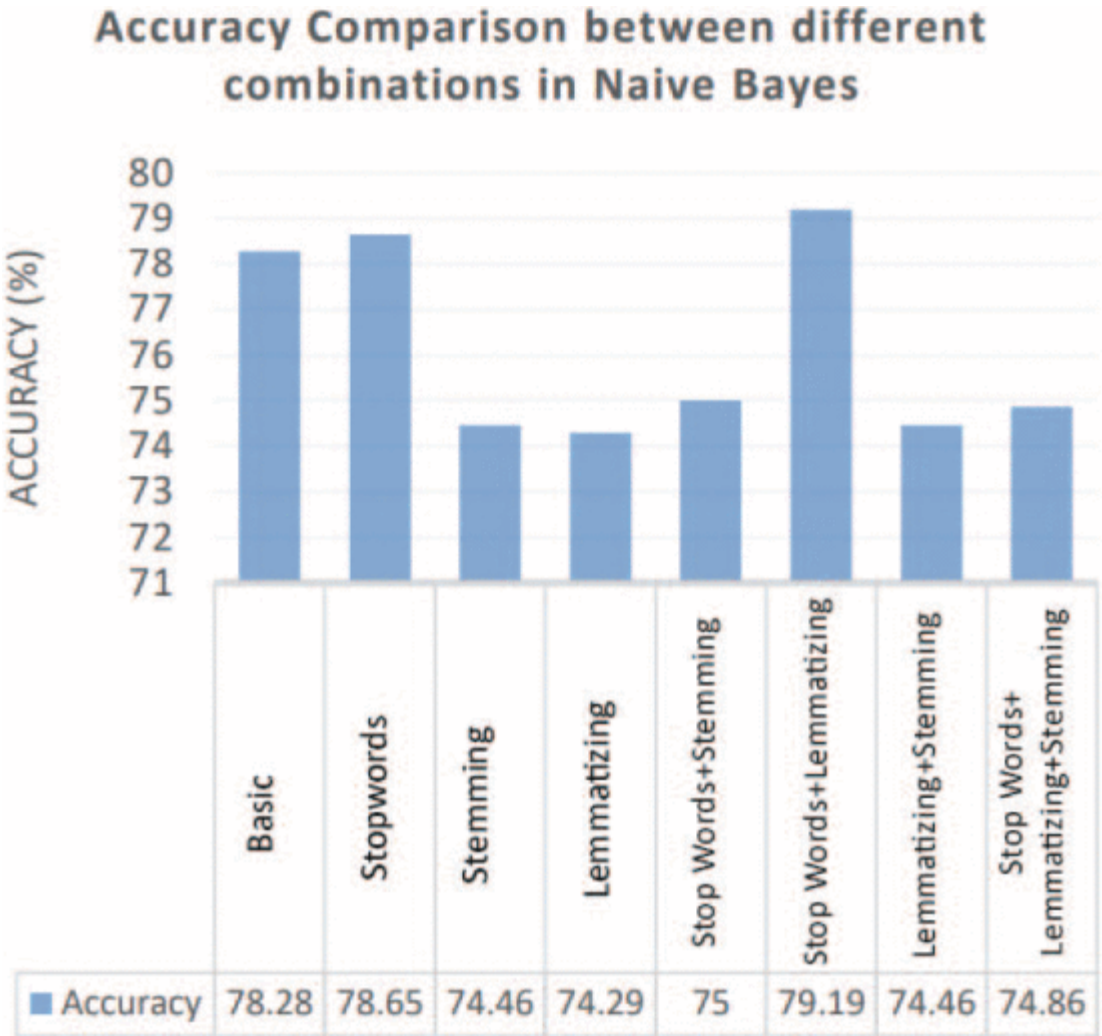
Romantic movies			
	true romancepos	true romanceneg	class precision
Pred.romaancepos	15	4	78.95%
Pred.romanceneg	5	16	76.19%
class recall	75.00%	80.00%	

A comparative approach to email classification using Naive Bayes classifier and hidden Markov model, Sebastian Romy Gomes et al.

- https://ieeexplore.ieee.org/abstract/document/8255404
- 1500 σημαντικά και 4000 spam email (Enron email dataset)
- Tokenization
- Διαγραφή λειτουργικών λέξεων
- Χαρακτηριστικά: οι 2650 πιο συχνές λέξεις των σημαντικών email και οι 2650 πιο συχνές λέξεις των spam

A comparative approach to email classification using Naive Bayes classifier and hidden Markov model,

Sebastian Romy Gomes et al.



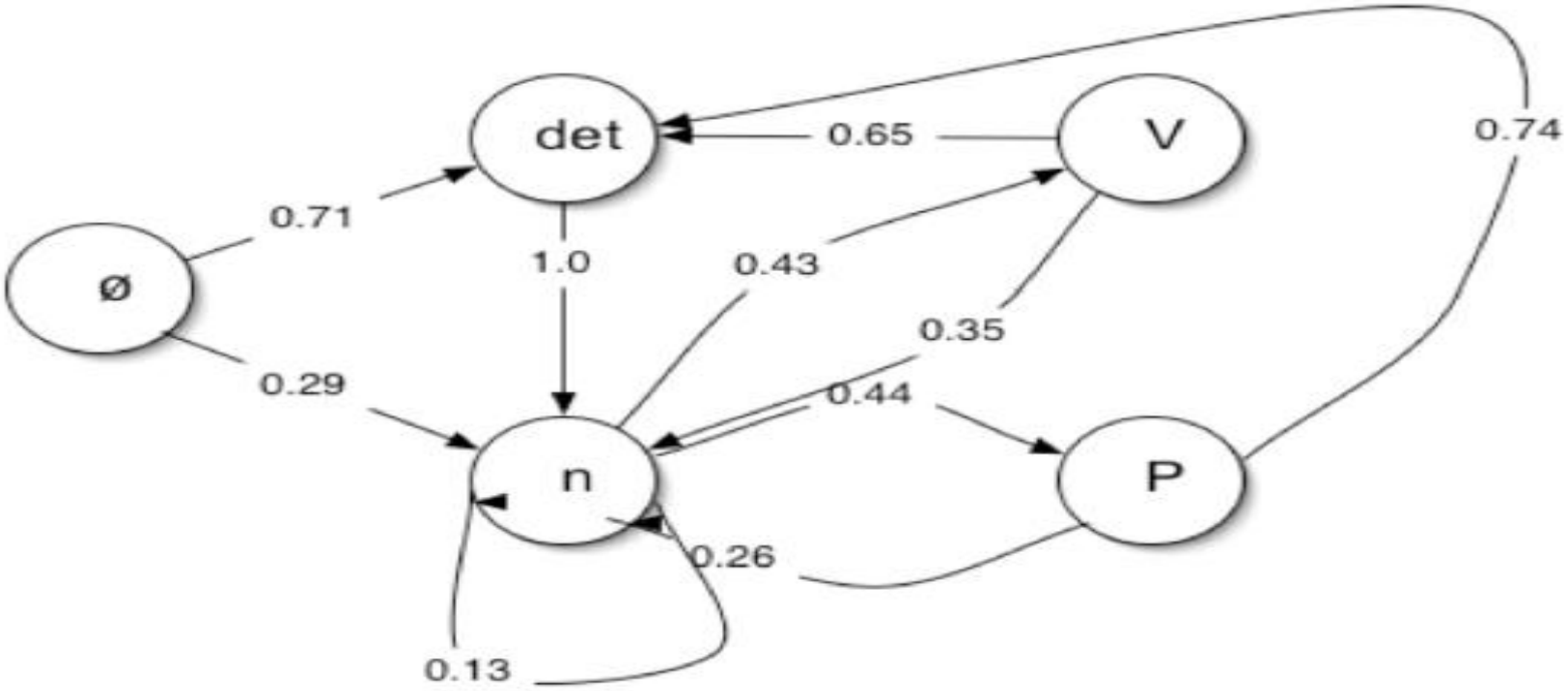
Κρυμμένα Μοντέλα Markov (Hidden Markov models)

HMM στην Αναγνώριση μερών του λόγου (POS tagging) - Μετάβαση

- Έστω οι παρακάτω μετρήσεις (πιθανότητες μετάβασης – transition probabilities) σε ένα επισημειωμένο ως προς το μέρος του λόγου σώμα κειμένων 300 προτάσεων

POS	freq	bigram	freq	Prob	estimate
Ø	300	Ø, DT	213	P [DT| Ø]	0.71
		Ø, NN	87	P [NN| Ø]	0.29
DT	558	DT, NN	558	P [NN| DT]	1.00
NN	833	NN, VB	358	P [VB| NN]	0.43
		NN, NN	108	P [NN| NN]	0.13
		NN, IN	366	P [IN| NN]	0.44
VB	300	VB, NN	75	P [NN| VB]	0.35
		VB, IN	194	P [IN| VB]	0.65
IN	307	IN, DT	226	P [DT| IN]	0.74
		IN, NN	81	P [NN| IN]	0.26

HMM στην Αναγνώριση μερών του λόγου (POS tagging) - Μετάβαση



Prob ["Ø DT NN VB IN DT NN"] =

0.71 · 1.0 · 0.43 · 0.65 · 0.74 · 1.0 = 0.14685

HMM στην Αναγνώριση μερών του λόγου (POS tagging) – Εκπομπή

- P(flower|NN) = count(flower,NN) / count(NN) = 53/833

	NN	VB	DT	IN	Total
flies	21	23	0	0	44
fruit	49	5	1	0	55
like	10	30	0	21	61
a	1	0	201	0	202
the	1	0	300	2	303
flower	53	15	0	0	68
flowers	42	16	0	0	58
birds	64	1	0	0	65
others	592	210	56	284	1142
total	833	300	558	307	1998

HMM στην Αναγνώριση μερών του λόγου (POS tagging): Απόδοση πιο πιθανής ακολουθίας μερών του λόγου σε ακολουθία λέξεων

- « *Flies*/NN *like*/VB *a*/DT *flower*/NN »?
- Prob [$w_{1,n} \mid t_{1,n}$] · Prob [$t_{1,n}$]
- Prob [$t_{1,n}$] = 0.29 · 0.43 · 0.65 · 1 = 4.68E-06
- Prob [$w_{1,n} \mid t_{1,n}$] = 0.0252 · 0.1 · 0.3602 · 0.0636 = 5.778E-05
- Prob [$w_{1,n} \mid t_{1,n}$] · Prob [$t_{1,n}$] = 4.68E-06 ← η πιο πιθανή ακολουθία

Με κανονικοποίηση: 0.556

Η δεύτερη πιο πιθανή ακολουθία: NN-IN-DT-NN (0.443)

HMM στην αναγνώριση συναισθήματος

- Mangi Kang, Jaelim Ahn, Kichun Lee. Opinion mining using ensemble text hidden Markov models for text classification, Expert Systems with Applications, Elsevier. 2018.
- Κατασκευή του tern-sentence matrix
- Εφαρμογή LSI για την μείωση της διαστατικότητας
  - Οι λέξεις αναπαρίστανται σε χώρο χαμηλότερης διαστατικότητας
- Εφαρμογή k-means clustering για την ομαδοποίηση των λέξεων βάσει της ομοιότητας συνημιτόνου (50-250 clusters)
- Μετάβαση: από μια ομάδα λέξεων σε μια άλλη
- Εκπομπή: μια συγκεκριμένη λέξη να ανήκει στο συγκεκριμένο cluster
- Με την παραπάνω διαδικασία σχηματίζονται δυο HMMs, ένα για το θετικό συναίσθημα (από τα θετικά παραδείγματα), και ένα για το αρνητικό (από τα αρνητικά παραδείγματα).
- Εάν η πιθανότητα που αποδίδει στην καινούρια ακολουθία λέξεων το θετικό HMM είναι μεγαλύτερη από ότι το αρνητικό, τότε στην ακολουθία αποδίδεται θετικό συναίσθημα
- Εφαρμογή σε πέντε datasets

Μοντέλα Γλώσσας - Language Models

Language Model

- Ένα μοντέλο γλώσσας εκτιμά την πιθανότητα εμφάνισης μιας ακολουθίας λέξεων σε μια γλώσσα
- Όσο πιο μεγάλη αυτή η πιθανότητα, τόσο πιο δόκιμη η συγκεκριμένη ακολουθία λέξεων στην γλώσσα
- P(I like bungee jumping off high bridges) = ?

Language model: Μονόγραμμο

- Εκπαίδευση: Από μεγάλο σώμα κειμένων εξαγωγή πιθανοτήτων μονογράμμου:

Unigram probabilities

$$p(w_1) = \frac{count(w_1)}{total\ words\ observed}$$

- P(I like bungee jumping off high bridges)

=P(I)*P(like)*P(bungee)*P(jumping)*P(off)*P(high)*P(bridges)*P(<s>)

Language model: Δίγραμμο

- Εκπαίδευση: Από μεγάλο σώμα κειμένων εξαγωγή πιθανοτήτων διγράμμου:

Bigram probabilities

$$p(w_2|w_1) = \frac{count(w_1 w_2)}{count(w_1)}$$

- P(I like bungee jumping off high bridges)

=P(I|<s>)*P(like|I)*P(bungee|like)*P(jumping|bungee)*P(off|jumping)*P(high|off)*P(bridges|high)*P(<s>|bridges)

Language model: Τρίγραμμο

- Εκπαίδευση: Από μεγάλο σώμα κειμένων εξαγωγή πιθανοτήτων τριγράμμου:

Trigram probabilities

$$p(w_3|w_1w_2) = \frac{count(w_1w_2w_3)}{count(w_1w_2)}$$

- P(I like bungee jumping off high bridges)

=P(I|<s><s>) * P(like|<s> I) * P(bungee|I like) * P(jumping|like bungee) * P(off|bungee jumping) * P(high|jumping off) * P(bridges|off high) * P(<s>|high bridges)

Αποτελέσματα του διγράμμου

- Μοντέλο bigrams: κάθε λέξη εξαρτάται μόνο από την προηγούμενή της
  - $P(w_1,w_2,\ldots,w_n) = \Pi\ P(w_i \mid w_{i-1})$
  - Πολλές φορές δεν αρκεί αυτό το μοντέλο:
    π.χ. "I hire the men who is good pilots"

Όσο μεγαλύτερο το N-γραμμο

- Όσο μεγαλύτερες ακολουθίες λέξεων χρησιμοποιώ για τον υπολογισμό των πιθανοτήτων, τόσο πιο απίθανο είναι να συναντήσω αυτές τις ακολουθίες στα δεδομένα
- Πρόβλημα σπάνιων δεδομένων (sparse data)
- Λύση: backing off (smoothing – εξομάλυνση)
- Χρησιμοποιώ συνδυασμό unigrams + bigrams + trigrams με αντίστοιχο βάρος στο καθένα

$$
\begin{aligned}
&.8 * p(w_3|w_1 w_2) + \\
&.15 * p(w_3|w_2) + \\
&.049 * p(w_3) + \\
&.001
\end{aligned}
$$

Αξιολόγηση μοντέλου

- Έστω οι προτάσεις ελέγχου S1, S2, …Sn
- Υπολογίζω το γινόμενο των πιθανοτήτων που μου παράγει για αυτές τις προτάσεις το μοντέλο

$$\prod_{i=1}^{n} P(S_i)$$

$$\log \prod_{i=1}^{n} P(S_i) = \sum_{i=1}^{n} \log P(S_i)$$

$$\text{Perplexity} = 2^{-x} \quad \text{where} \quad x = \frac{1}{W} \sum_{i=1}^{n} \log P(S_i)$$

- Και W είναι ο συνολικός αριθμός λέξεων στις προτάσεις ελέγχου.
- Όσο μικρότερο το perplexity τόσο καλύτερο το μοντέλο