Εφαρμογές Μηχανικής Μάθησης στο Μάρκετινγκ

ΚΩΝΣΤΑΝΤΙΝΟΣ ΔΙΑΜΑΝΤΑΡΑΣ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
Α.Τ.Ε.Ι. ΘΕΣΣΑΛΟΝΙΚΗΣ

Εφαρμογές Μηχανικής Μάθησης στην πράξη

- Συνεργασία με την εταιρία MSENSIS
- Πρόγραμμα ΠΑΒΕΤ 2013 [2014-15]
  - Αναγνώριση συναισθήματος σε κείμενο (Sentiment Analysis)
- Πρόγραμμα ICT4Growth [2014-15]
  - Πρόβλεψη αλλαγής τηλεπικοινωνιακού παρόχου (customer churning)
  - Συστήματα συστάσεων (Recommendation Systems)

Αναγνώριση συναισθήματος κειμένου

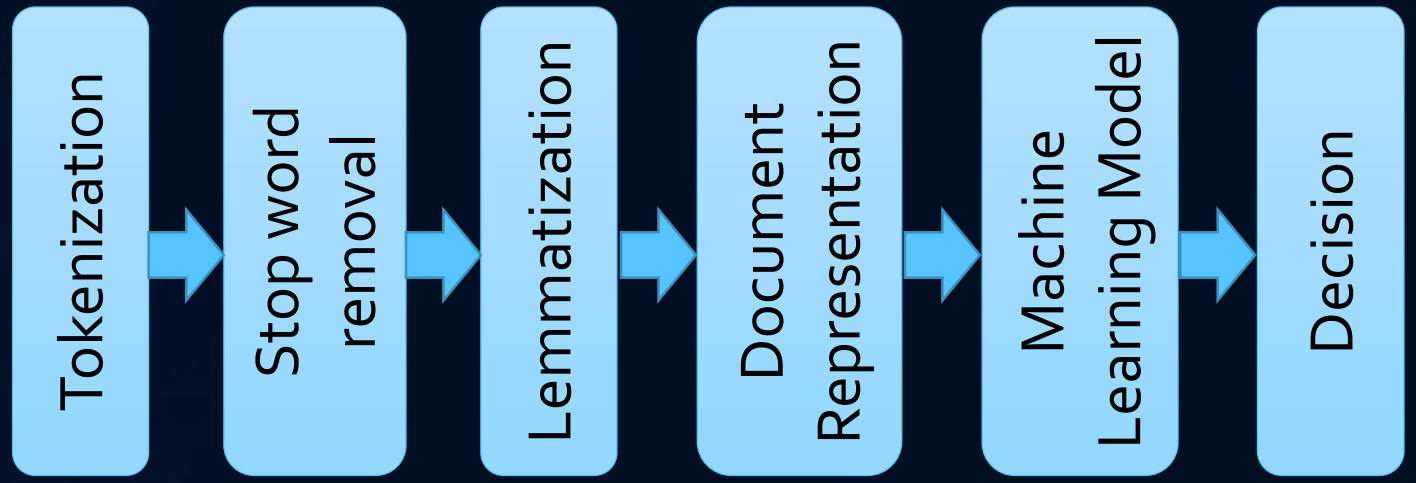
- Εξαγωγή της διάθεσης ή άποψης του συγγραφέα ενός ελεύθερου κειμένου (sentiment analysis ή opinion mining)
- Εξαγωγή συναισθήματος (χαρά, λύπη, έκπληξη, φόβος, αηδία, κλπ)
- Σημαντικό γιατί επιτρέπει την ανάλυση του impact ενός προϊόντος ή μιας υπηρεσίας σε πραγματικό χρόνο με βάση σχόλια στο διαδίκτυο (πχ.)
- Γλώσσες στόχοι: Ελληνικά, Αγγλικά
- Διαθέσιμα εργαλεία στα Αγγλικά
  - Natural Language Toolkit (http://www.nltk.org)
- Ελάχιστα ανοιχτά / ελεύθερα εργαλεία στα Ελληνικά.
- Αναπτύχθηκαν οι παρακάτω πόροι:
  - Ελληνικό λεξικό με όλες τις φόρμες κλίσης των λέξεων
  - Λογισμικό συντακτικής ανάλυσης (αναγνώριση κλίσης, λημματισμός, κλπ)

Αναγνώριση συναισθήματος κειμένου: Σώματα κειμένου (Corpora)

- Στα Ελληνικά
  - πηγή www.skroutz.gr : αξιολογήσεις κινητών τηλεφώνων (2800 προτάσεις)
  - πηγή www.kotsovolos.gr [Agathangelou e.a. 2014] : αξιολογήσεις ηλεκτρικών συσκευών (1976 προτάσεις training, 3329 προτάσεις testing)
- Στα Αγγλικά
  - Movie Reviews, [Pang & Lee 2005] : Αξιολογήσεις ταινιών από το www.rottentomatos.com (10662 προτάσεις)

Αναγνώριση συναισθήματος κειμένου

- Μεθοδολογία ανεξαρτήτως γλώσσας



Αναγνώριση συναισθήματος: Προεπεξεργασία

- Tokenization
  - Διαχωρισμός λέξεων
- Stop word removal
  - Αφαίρεση λέξεων χωρίς σημασιολογικό περιεχόμενο (άρθρα, σύνδεσμοι, κλπ)
- Lemmatization
  - Μετατροπή λέξεων στην αρχική μορφή λήμματος λεξικού
  - Πχ. running → run, διάλεξε → διαλέγω, οθόνες → οθόνη, κλπ

Αναγνώριση συναισθήματος: διανυσματική αναπαράσταση κειμένων

- Αναπαράσταση κειμένου = άθροισμα των αναπαραστάσεων των λέξεων
- Αναπαράσταση λέξεων:
  - Bag of Words : Διάνυσμα διάστασης L = πλήθος λημματισμένων λέξεων.
    - $b[i] = 0,$ αν η λέξη i δεν εμφανίζεται στο κείμενο,
    - $b[i] = \text{tf}-\text{idf} = \log\left(\frac{N}{n_i}\right) f_i,$ αν η λέξη εμφανίζεται στο κείμενο
  - Λεξικό συναισθημάτων
    - Αγγλικά: *NRC Word-Emotion Association Lexicon*, ή *EmoLex* [Mohammad and Turney 2013] (14182 λέξεις)
    - Ελληνικά: *expanded Greek Sentiment Lexicon* (4658 λέξεις) επέκταση του [Tsakalidis e.a. 2014]

Αναγνώριση συναισθήματος: διανυσματική αναπαράσταση λέξεων

- Word2Vec [Mikolov 2013]
  - Νευρωνικό Μοντέλο.
  - Αναπαράσταση λέξεων με διανύσματα διάστασης W που επιλέγει ο χρήστης (τυπικά $W = 300$)
  - Δημιουργεί διανύσματα που διατηρούν το σημασιολογικό περιεχόμενο των λέξεων. Πχ οι λέξεις «Ρώμη» και «Ιταλία» έχουν κοντινή αναπαράσταση στο διανυσματικό χώρο
  - Απαιτεί εκπαίδευση σε πολύ μεγάλο corpus (πχ. όλη την Wikipedia)
  - Ταχύτατη μέθοδος (λίγα λεπτά για όλη την Ελληνική Βικιπαίδεια)

Διαχείριση αρνητικών φράσεων

- Αναζήτηση αρνητικών όρων στα Αγγλικά (e.g. no, never, don’t) και στα Ελληνικά (π.χ. μην, δεν, όχι).
- Αναπαράσταση άρνησης:
  - REVERSE: Αναπαράσταση της αρνητικής φράσης με αρνητικό πρόσημο

    $$rep(neg\ x) = -rep(x)$$

    - Πχ. $rep(\text{καλός}) = 1,\ \ rep(\text{όχι καλός}) = -1$
  - DOUBLE: Διπλασιασμός του μεγέθους του διανύσματος αναπαράστασης. Για κάθε λέξη δημιουργούνται δύο θέσεις στο διάνυσμα, μια για την άρνηση και μια για την κατάφαση. Μια από τις δύο θέσεις έχει τιμή 0.

Μοντέλο Μηχανικής Μάθησης

- Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)
- Ισχυρά μοντέλα με χαμηλό πλήθος παραμέτρων
- Επιλογή μη γραμμικής συνάρτησης μετασχηματισμού πυρήνα: Γραμμικός (linear), Gaussian (RBF)
- Επιλογή καλύτερων παραμέτρων με cross-validation

Αποτελέσματα (μόνο BoW+Sentiment)

- Χωρίς Word2Vec
- *Διάνυσμα αναπαράστασης : BoW-DOUBLE και Average emotion-DOUBLE*

Dataset	Accuracy
Movie Reviews (Αγγλικά)	63,84%
Mobile-Sen (Ελληνικά)	73,36%
Mobile-Par (Ελληνικά)	84,59%

Αποτελέσματα (μόνο Word2Vec)

- Με Word2Vec, εκπαιδευμένο ως εξής:
  - Για τα Αγγλικά με Google-EN (3M λέξεις) πηγή: http://code.google.com/p/word2vec/
  - Για τα Ελληνικά με Wiki-EL (Ελληνική Βικιπαίδεια)

Dataset	Καλύτερη επίδοση
Movie Reviews (Αγγλικά)	77,56%
Mobile-Sen (Ελληνικά)	81,96%
Mobile-Par (Ελληνικά)	85,43%

Αποτελέσματα (Υβριδικό μοντέλο)

- Word2Vec + BoW + Sentiment
  - Αγγλικά : Οριακά καλύτερο σε σχέση με την περίπτωση χωρίς W2V
  - Ελληνικά: Καλύτερο σε σχέση με την περίπτωση χωρίς W2V.
  - Mobile-PAR : Σημαντικά καλύτερο σε σύγκριση με [Agathangelou 2014] όπου Accuracy = 78,05%

Dataset	Accuracy
Movie Reviews (Αγγλικά)	77,84%
Mobile-SEN (Ελληνικά)	83,86%
Mobile-PAR (Ελληνικά)	86,21%

Δημοσιεύσεις

- M. Giatsoglou, M. Vozalis, K. I. Diamantaras, A. Vakali, G. Sarigiannidis, K. Ch. Chatzisavvas, "Sentiment Analysis Leveraging Emotions and Word Embeddings", Expert Systems with Applications, vol. 69, pp. 214-224, Elsevier, March 2017
- P. Stalidis, M. Giatsoglou, K. I. Diamantaras, G. Sarigiannidis, K. Ch. Chatzisavvas, "Machine Learning Sentiment Prediction based on Hybrid Document Representation", arXiv:1511.09107, Nov. 2015

Πρόβλεψη αλλαγής παρόχου

- Churn rate = πιθανότητα ο πελάτης να προβεί σε αλλαγή παρόχου μιας υπηρεσίας, πχ. τηλεπικοινωνιακής, τραπεζικής, ασφαλιστικής, κλπ.
- Σημαντικό για την αγορά υπηρεσιών
- Κόστος Customer Acquisition = 20 × Κόστος Customer Retention
- Η επιτυχής πρόβλεψη του churn rate ιδιαίτερα χρήσιμη για Customer loyalty management

Churn prediction: Σύγκριση μοντέλων μηχανικής μάθησης

- Μεθοδολογία:
- Εκπαίδευση διαφορετικών μοντέλων
  - Νευρωνικά Δίκτυα 2 στρωμάτων
  - Μοντέλα διανυσμάτων υποστήριξης (SVM)
  - Naive Bayes
  - Δέντρα αποφάσεων (Decision trees)
  - Logistic regression
- Δοκιμή όλων των μοντέλων με boosting και χωρίς boosting
- Εκτίμηση της επίδοσης πρόβλεψης με χρήση cross-validation

Churn prediction: Σύγκριση μοντέλων μηχανικής μάθησης (2)

- Boosting (Adaboost.M1): Αύξηση επίδοσης με συνδυασμό ταξινομητών
- Δείγματα: x_i , Στόχοι: y_i , $i = 1, \cdots, N$. Δίνουμε βάρη στα δείγματα. Αρχικά όλα τα βάρη είναι ίσα: $w_1(i) = \frac{1}{N}$
- Επαναληπτικά, στο βήμα $t = 1, \cdots, T$ εκπαιδεύουμε έναν ταξινομητή $h_t(x)$ ώστε να ελαχιστοποιήσει το κόστος

$$E_t = \sum_{i:\, h_t(x_i) \neq y_i} w_t(i)$$

- Ενημερώνουμε τα βάρη ώστε να αυξηθούν μόνο για τα πρότυπα που έγινε λάθος ταξινόμηση: $w_{t+1}(i) = w_t(i)\beta_t / Z_t$

$$\beta_t = (1 - E_t)/E_t$$

- Απόφαση = συνδυασμός πολλών ταξινομητών $h_t, t = 1, \cdots, T$

$$h_{total}(x) = \arg\max_y \sum_{t:\, h_t(x)=y} \log(\beta_t)$$

Churn Dataset

- Churn Dataset δημόσια διαθέσιμο στο πακέτο C50 της R https://cran.r-project.org/web/packages/C50/C50.pdf

Χαρακτηριστικό	Τύπος	Χαρακτηριστικό	Τύπος
account_length	Num	total_intl_minutes	Num
total_eve_charge	Num	total_day_minutes	Num
area_code	Num	total_intl_calls	Num
total_night_minutes	Num	total_day_calls	Num
international_plan	Yes/no	total_intl_charge	Num
total_night_calls	Num	total_day_charge	Num
voice_mail_plan	Yes/no	number_customer_service_calls	Num
total_night_charge	Num	total_eve_minutes	Num
number_vmail_messages	Num	total_eve_calls	Num

Αποτελέσματα

- Κριτήρια επίδοσης:

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp+fn}, \qquad \text{F-measure} = \frac{2\times\text{precison}\times\text{recall}}{\text{precision}+\text{recall}}$$

$$\text{precision} = \frac{tp}{tp+fp}, \qquad \text{recall} = \frac{tp}{tp+fn}$$

Ταξινομητής	Χωρίς Boosting Accuracy (%)	Με Boosting Accuracy (%)	Χωρίς Boosting F-measure (%)	Με Boosting F-measure (%)
Naïve Bayes	86.94	-	53.31	-
Logistic Regression	87.94	-	14.46	-
BPN	94.06	95.05	77.48	80.97
SVM-RBF	93.18	96.05	73.16	84.22
SVM-Polynomial	93.04	**96.85**	73.11	**84.57**
DT-C50	94.15	95.09	77.04	83.87

Δημοσιεύσεις

- T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," Simulation Modelling Practice and Theory, vol. 55, pp. 1–9, Jun. 2015.

Συστήματα συστάσεων (Recommendation Systems)

- Συστήματα στοχευμένων, προσωποποιημένων προτάσεων προς πελάτες / χρήστες.
- Πλεονεκτήματα:
  - Αύξηση αξίας προϊόντων
  - Αύξηση εμπλοκής του πελάτη με την υπηρεσία
- Παραλλαγές με βάση τους τύπους των δεδομένων:
  - Χρήση ιστορικού αξιολόγησης προϊόντων από τον χρήστη (ratings)
  - Χρήση περιεχομένου προϊόντων, πχ. περίληψη ενός βιβλίου, ηθοποιοί μιας ταινίας, κλπ. (μέθοδοι content-based)
  - Χρήση δεδομένων κοινωνικού χαρακτήρα (social recommenders)
  - Υβριδικές μέθοδοι, συνδυάζοντας τα παραπάνω

Προτάσεις με βάση ιστορικό αξιολόγησης

- Πιο απλή και συνηθισμένη περίπτωση. Διαθέσιμα δεδομένα:

$$\text{rating matrix } \mathbf{R} = [r_{ui}],\ u\text{=user},\ i\text{=item}$$

- Μέθοδοι:
  - Collaborative Filtering / User-based: Εκτιμάμε την αρέσκεια του χρήστη u προς το αντικείμενο i παίρνοντας ζυγισμένο μέσο όρο της αρέσκειάς του χρήστη προς παρόμοια αντικείμενα (s_{ij}=similarity between items i, j)

$$\hat{r}_{ui} = \sum_j s_{ij} r_{uj} \,/\, \sum_j s_{ij}$$

  - Collaborative Filtering / Item-based: Εκτιμάμε την αρέσκεια του χρήστη u προς το αντικείμενο i παίρνοντας ζυγισμένο μέσο όρο της αρέσκειάς παρόμοιων χρηστών προς αυτό το αντικείμενο (s_{uv}=similarity between users u, v)

$$\hat{r}_{ui} = \sum_j s_{uv} r_{vi} \,/\, \sum_v s_{uv}$$

Προτάσεις με βάση ιστορικό αξιολόγησης

- Μέθοδοι (συνέχεια):
  - Hybrid item-based: παραλλαγή της μεθόδου item-based όπου χρησιμοποιούνται τα ratings των πιο όμοιων αντικειμένων υπό την προϋπόθεση ότι έχουν αξιολογηθεί από ένα ελάχιστον πλήθος όμοιων χρηστών.
  - Παραγοντοποίηση του πίνακα $\mathbf{R}$ (matrix factorization) : $\mathbf{R} \approx \widehat{\mathbf{R}} = \mathbf{PQ}$

$$\hat{r}_{ui} = \mathbf{p}_u^T \mathbf{q}_i$$

  - Παραγοντοποίηση με υβριδικό SVD. Έστω ότι $0 \leq r_{ij} \leq 5$
    - Αρχικά παραγοντοποιούμε $\mathbf{R} \approx \widehat{\mathbf{R}} = \mathbf{PQ}$. Αν κάπου $\hat{r}_{ij} > 5$ ή $\hat{r}_{ij} < 0$ κάνουμε κλιπ τις τιμές μεταξύ 0 και 5 και επαναλαμβάνουμε την παραγοντοποίηση για το νέο πίνακα $\widehat{\mathbf{R}}$.

Προτάσεις με βάση το περιεχόμενο

- Πχ. Περιγραφή βιβλίων. Αναπαράσταση περιεχομένου με το μοντέλο Bag of Words (BoW)
- Μέθοδος Rocchio
  - Αναπαράσταση του περιεχομένου με διανύσματα TF-IDF
  - Δημιουργούνται δύο διανύσματα προφίλ για κάθε χρήστη: ένα θετικό με τα αντικείμενα που του αρέσουν και ένα αρνητικό. Ταξινόμηση αντικειμένου με βάση το πιο κοντινό διάνυσμα
- Clustering με τη μέθοδο ArtMAP
  - Παραλλαγή του Rocchio αλλά γίνεται clustering των αντικειμένων με βάση το διάνυσμα TF-IDF. Έτσι δημιουργούνται περισσότερα προφίλ του χρήστη, ένα για κάθε cluster.

Datasets

- Amazon books I: 116 χρήστες, 110 βιβλία. Υποσύνολο του Amazon Ratings library https://snap.stanford.edu/data/web-Amazon.html/
- Amazon books II: 304 χρήστες, 110 βιβλία . Υποσύνολο του Amazon Ratings library
- Jester: Συλλογή από 100 ανέκδοτα αξιολογημένα από 24983 χρήστες

Αποτελέσματα

- Κριτήριο: Mean Absolute Error (MAE) $= \langle |\hat{r}_{ij} - r_{ij}| \rangle$

Μέθοδος	Amazon (I)	Amazon (II)	Jester
User-based CF	1.42	1.45	1.09
Item-based CF (Pearson similarity)	1.19	1.22	1.04
Item-based CF (Adjusted cosine sim.)	1.78	1.27	1.02
Rocchio	0.02	0.04	0.58
ArtMAP	0.04	0.06	0.49
Hybrid item-based	-	-	0.88

Προτάσεις με βάση social information

- Μαζί με τις αξιολογήσεις διαθέτουμε πληροφορίες για τη σχέση φιλίας / εμπιστοσύνης μεταξύ χρηστών
- Συγκριτική δοκιμή 5 αλγορίθμων
  - **SocialMF**: [Breese 1998].
  - **SoRec**: [Deshpande 2004].
  - **SoReg**: [Goldberg 1992].
  - **TrustMF**: [Kitts 2000].
  - **TrustSVD**: [Linden 2003].

Αποτελέσματα – Social Recommenders

- Dataset: Epinion.com http://www.trustlet.org/wiki/Downloaded_Epinions_dataset
  - 49290 users, 139738 items, 664824 ratings και 487181 trust statements

Αλγόριθμος	RMSE	indicative training run time (min)
SocialMF	1.1	1.32
SoRec	1.1	2.08
SoReg	1.47	0.08
TrustMF	1.22	8.14
TrustSVD	1.1	1.29

Βιβλιογραφία

- Agathangelou, P., Katakis, I., Kokkoras, F. and Ntonas, K., (2014). "Mining domain-specific dictionaries of opinion words". In Int. Conf. on Web Information Systems Engineering (pp. 47-62). Springer International Publishing.
- Pang, B. and Lee, L., (2005). «Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales». In Proc. of the 43rd annual meeting on association for computational linguistics (pp. 115-124). Association for Computational Linguistics.
- Tsakalidis, A., Papadopoulos, S. and Kompatsiaris, I., (2014). "An ensemble model for cross-domain polarity classification on twitter". In Int. Conf. on Web Information Systems Engineering (pp. 168-177). Springer International Publishing.
- Breese J. S., Heckerman D., Kadie C., (1998). "Empirical Analysis of Predictive Algorithms for Collaborative Filtering." In Uncertainty in AI. Proc. 14th Conf., pp. 43–52.
- Deshpande M., Karypis G., (2004). "Item-based top-N recommendation algorithms." ACM Transations on Information Systems, 22(1), 143–177.
- Goldberg D., Nichols D., Oki B. M., Terry D., (1992). "Using collaborative filtering to weave an information tapestry." Communications of the ACM, 35(12), 61–70.
- Kitts B., Freed D., Vrieze M., (2000). "Cross-sell: a fast promotion-tunable customer-item recommendation method based on conditionally independent probabilities." In KDD '00: Proc. 6th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining, pp. 437–446. ACM.
- Linden G., Smith B., York J., (2003). "Amazon.com Recommendations: Item-to-Item Collaborative Filtering." IEEE Internet Computing, 7(1), 76–80.

Ευχαριστώ!