ΔΙΑΧΕΙΡΙΣΗ ΠΕΡΙΕΧΟΜΕΝΟΥ ΠΑΓΚΟΣΜΙΟΥ ΙΣΤΟΥ ΚΑΙ ΓΛΩΣΣΙΚΑ ΕΡΓΑΛΕΙΑ

Data Mining - Classification

Data Mining

- Ανακάλυψη προτύπων σε μεγάλο όγκο δεδομένων.
- Σαν πεδίο περιλαμβάνει κλάσεις εργασιών:
  - **Anomaly Detection**: Αναγνώριση δεδομένων που αποκλίνουν από τα συνήθη.
  - **Association Rule Learning**: Εκμάθηση κανόνων συσχέτισης μεταξύ μεταβλητών.
  - **Clustering**: Ομαδοποίηση δεδομένων
  - **Classification**: Ανάθεση κατηγοριών σε δεδομένα
  - **Regression**: Εύρεση συναρτήσεων που προσεγγίζουν τα δεδομένα με το ελάχιστο λάθος
  - **Summarization**: Συμπυκνωμένη αναπαράσταση των δεδομένων

Machine Learning

- Εποπτευόμενες Μέθοδοι (Supervised Methods)
  - Στόχος η εκμάθηση μιας λειτουργίας
  - Επιτυγχάνεται με χρήση δεδομένων εκπαίδευσης
  - Εφαρμόζεται η λειτουργία σε άγνωστα δεδομένα
  - πχ Classification
- Μη-Εποπτευόμενες Μέθοδοι (Unsupervised Methods)
  - Στόχος η ομαδοποίηση ενός συνόλου δεδομένων
  - Δεν υπάρχει επισημείωση στα δεδομένα
  - Επιτυγχάνεται με βάση ορισμένα από πριν χαρακτηριστικά
  - πχ Clustering

Classification

- Στο πεδίο του NLP, η σημαντικότερη εφαρμογή είναι η κατηγοριοποίηση (classification).
- Το πρόβλημα του classification ορίζεται ως εξής:
  - Έστω ένα σύνολο από ενδεχόμενα
  - Έστω ένα σύνολο από κλάσεις με τις αντίστοιχες ετικέτες τους
  - Ένας κατηγοριοποιητής (classifier) είναι μια συνάρτηση η οποία:
    - Δέχεται ως είσοδο ένα ενδεχόμενο και μια κλάση
    - Χρησιμοποιεί χαρακτηριστικά (features) του ενδεχομένου
    - Επιστρέφει ως έξοδο αν ανήκει το ενδεχόμενο στην κλάση ή όχι

Classifying Classification

- Με βάση τον αριθμό των κλάσεων
  - Δυαδική κατηγοριοποίηση (binary)
  - Κατηγοριοποίηση πολλαπλών κλάσεων (multiclass)
- Με βάση τον αριθμό των κλάσεων που μπορούν να ανατεθούν σε ένα ενδεχόμενο
  - Μοναδικής ετικέτας (single label)
  - Πολλαπλών ετικετών (multi label)
- Με βάση τον τύπο της ανάθεσης σε κλάση
  - Hard Classification: Ανήκει ή όχι στην κλάση
  - Soft Classification: Πιθανότητα να ανήκει στην κλάση
- Με βάση την οργάνωση των κατηγοριών
  - Flat Classification: Λίστα κατηγοριών
  - Hierarchical Classification: Δενδρική δομή κατηγοριών

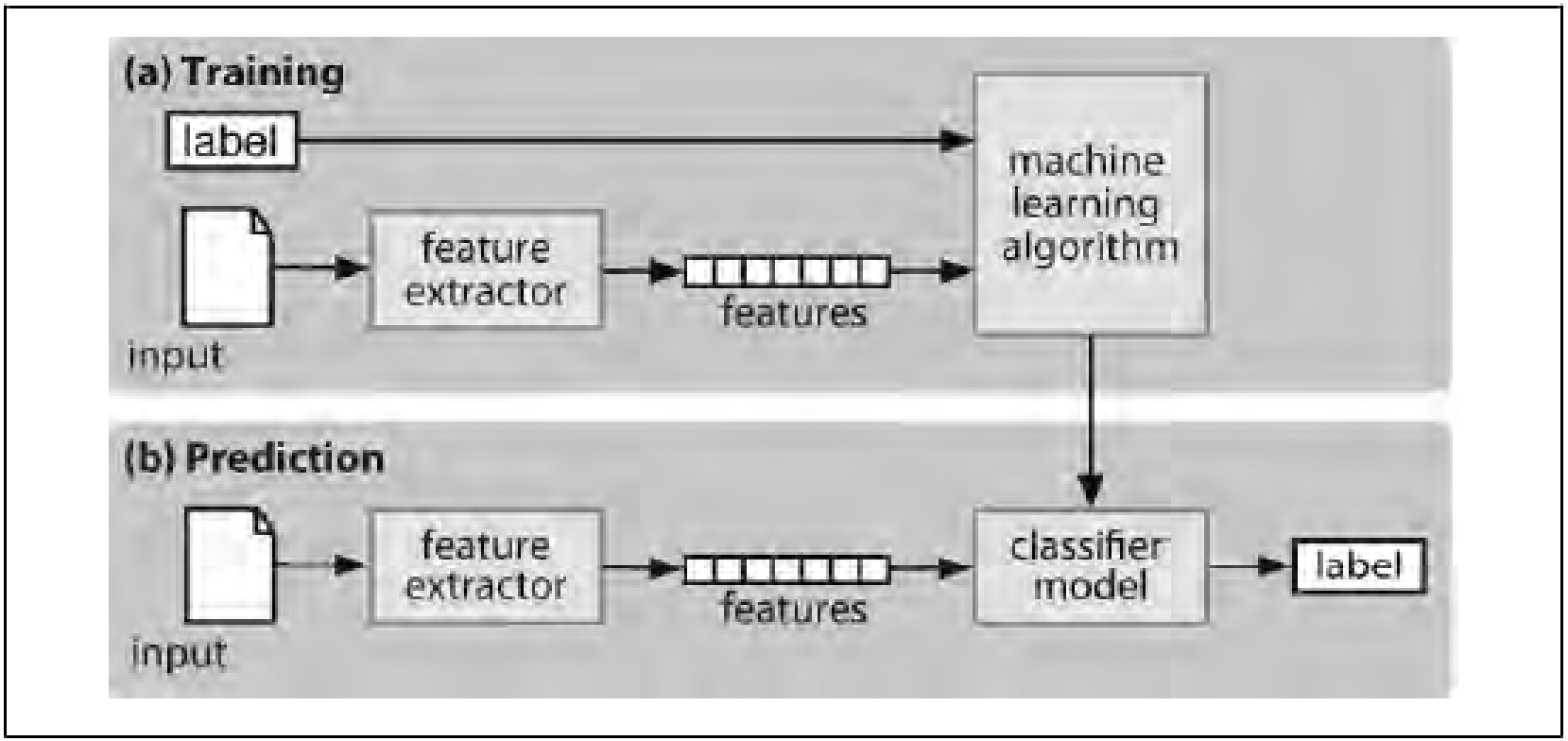
Text Classification

- Η κατηγοριοποίηση κειμένου ορίζεται ως εξής:
  - Έστω ένα σύνολο κειμένων
  - Έστω ένα σύνολο κλάσεων και οι αντίστοιχες ετικέτες τους.
  - Ένας κατηγοριοποιητής είναι μια συνάρτηση που αποφασίζει με βάση τα χαρακτηριστικά ενός κειμένου αν ανήκει σε μια συγκεκριμένη κλάση.
- Υπο-προβλήματα κατηγοριοποίησης κειμένου:
  - Κατηγοριοποίηση θέματος (subject)
    - Οι κλάσεις αντιπροσωπεύουν θέματα
  - Κατηγοριοποίηση είδους (genre)
    - Οι κλάσεις αντιπροσωπεύουν λειτουργίες ενός κειμένου, πχ νέα, επιστολή, πεζό κλπ
  - Κατηγοριοποίηση αισθήματος (sentiment)
    - Οι κλάσεις αντιπροσωπεύουν το αίσθημα που εκφράζει το κείμενο, πχ θετική, αρνητική ή ουδέτερη άποψη.
  - Γενικά υπάρχουν τόσα υποπροβλήματα όσα σύνολα κλάσεων μπορούν να δημιουργηθούν

Supervised Classification Process

- Παρέχεται ένα σύνολο κλάσεων και ενδεχόμενα για κάθε κλάση
- Αναπαρίσταται κάθε ενδεχόμενο ως ένα διάνυσμα χαρακτηριστικών με μια τιμή για κάθε χαρακτηριστικό
- Εκπαίδευση
  - Επιλέγεται μέρος των ενδεχομένων
  - Ο κατηγοριοποιητής «μαθαίνει» τη συνάρτηση ανάθεσης σε κλάση με βάση τα χαρακτηριστικά.
- Αξιολόγηση
  - Για τα υπόλοιπα ενδεχόμενα θεωρείται άγνωστη η κλάση
  - Εφαρμόζεται η συνάρτηση ανάθεσης
  - Υπολογίζεται το ποσοστό επιτυχίας

Supervised Classification Process



Supervised Classification Example

- Με βάση δυο λίστες που περιλαμβάνει το NLTK για αντρικά και γυναικεία ονόματα, θα εκπαιδεύσουμε και θα αξιολογήσουμε ένα κατηγοριοποιητή.

```
>>> names=nltk.corpus.names
>>> names.fileids()
['female.txt', 'male.txt']
>>> male_names=names.words('male.txt')
>>> female_names=names.words('female.txt')
>>> len(male_names)
2943
>>> len(female_names)
5001
>>> male_names[:10]
['Aamir', 'Aaron', 'Abbey', 'Abbie', 'Abbot', 'Abbott', 'Abby', 'Abdel', 'Abdul'
, 'Abdulkarim']
>>> female_names[:10]
['Abagael', 'Abagail', 'Abbe', 'Abbey', 'Abbi', 'Abbie', 'Abby', 'Abigael', 'Abi
gail', 'Abigale']
```

Supervised Classification Example

- Ορισμός του προβλήματος
  - Το σύνολο ενδεχομένων αποτελείται από 2943+5001 ενδεχόμενα.
  - Το σύνολο των κλάσεων περιλαμβάνει 2 κλάσεις: αντρικό και γυναικείο.
  - Για κάθε ενδεχόμενο δίνεται το όνομα σαν λεκτικό και η κλάση/κλασεις στην οποία ανήκει.
  - Ζητάται η δημιουργία της συνάρτησης του κατηγοριοποιητή.

Supervised Classification Example

- Επιλογή Χαρακτηριστικών
  - Χρήσιμα για την κατηγοριοποίηση, με μεγάλη ικανότητα διαχωρισμού.
  - Επιλογή αναπαράστασης/μετρήσιμου μεγέθους
  - Συνήθης μέθοδος (kitchen sink): περιλαμβάνουμε τα πάντα και στην πορεία με δοκιμές επιλέγουμε τα χρήσιμα (feature selection).
- Για τα ονόματα ποιο χαρακτηριστικό μπορεί να διαχωρίσει αντρικά από γυναικεία?

```
>>> def gender_features(name):
...     return {'last_letter':name[-1]}
...
>>> def gender_features2(name):
...     features={}
...     features['first_letter']=name[0]
...     features['length']=len(name)
...     return features
...
>>> gender_features('Mary')['last_letter']
'y'
>>> gender_features2('Mary')['first_letter']
'M'
>>> gender_features2('Mary')['length']
4
```

Supervised Classification Example

- Διαχωρισμός του συνόλου εκπαίδευσης από το σύνολο

```
>>> from nltk.corpus import names
>>> import random
>>> names=([(name,'male') for name in names.words('male.txt')]+
...     [(name,'female') for name in names.words('female.txt')])
>>> random.shuffle(names)
>>>
>>> feature_set=[(gender_features(n),g) for (n,g) in names]
>>> train_set=feature_set[500:]
>>> test_set=feature_set[:500]
>>>
>>> feature_set2=[(gender_features2(n),g) for (n,g) in names]
>>> train_set2=feature_set2[500:]
>>> test_set2=feature_set2[:500]
>>>
```

Supervised Classification Example

- Δημιουργία του classifier και αξιολόγηση με βάση το test set.
- Στο παράδειγμα φαίνεται ότι ο ένας classifier έχει καλύτερη απόδοση από τον δεύτερο. Γιατί?

```
>>> classifier=nltk.NaiveBayesClassifier.train(train_set)
>>> classifier.classify(gender_features('Vivi'))
'female'
>>> print nltk.classify.accuracy(classifier,test_set)
0.746
>>>
>>> classifier2=nltk.NaiveBayesClassifier.train(train_set2)
>>> print nltk.classify.accuracy(classifier2,test_set2)
0.642
```

Supervised Classification Example

- Βελτίωση με επιλογή καλύτερου χαρακτηριστικού

```
>>> def gender_features3(name):
...     return {'suffix':name[-2:]}
...
>>> feature_set3=[(gender_features3(n),g) for (n,g) in names]
>>> train_set3=feature_set3[500:]
>>> test_set3=feature_set3[:500]
>>> classifier3=nltk.NaiveBayesClassifier.train(train_set3)
>>> print nltk.classify.accuracy(classifier3,test_set3)
0.766
>>>
>>> def gender_features4(name):
...     return {'suffix':name[-2:],'last_letter':name[-1]}
...
>>> feature_set4=[(gender_features4(n),g) for (n,g) in names]
>>> train_set4=feature_set4[500:]
>>> test_set4=feature_set4[:500]
>>> classifier4=nltk.NaiveBayesClassifier.train(train_set4)
>>> print nltk.classify.accuracy(classifier4,test_set4)
0.776
```

Text Classification

- Στην πράξη, η πιο συχνή μορφή του προβλήματος κατηγοριοποίησης κειμένου ορίζεται ως εξής:
  - Το σύνολο εκπαίδευσης αποτελείται από κείμενα και κάθε κείμενο είναι ένα ενδεχόμενο.
  - Το σύνολο των κλάσεων περιλαμβάνει κλάσεις οι οποίες αντιπροσωπεύουν θέματα κειμένου (πχ τέχνη, πολιτική, ιστορία κλπ).
  - Κάθε ενδεχόμενο αντιπροσωπεύεται από ένα σύνολο χαρακτηριστικών όρων και μια τιμή για κάθε όρο (διάνυσμα χαρακτηριστικών όρων με βάρη).
  - Για κάθε ενδεχόμενο δίνεται η κλάση/κλασεις στην οποία ανήκει.
  - Η συνάρτηση του κατηγοριοποιητή δημιουργείται με βάση τα διανύσματα των κειμένων.

Web Pages Classification

- Σε σχέση με την παραδοσιακή κατηγοριοποίηση κειμένου διαφέρει:
  - Σε κλάσεις κατηγοριοποίησης
    - Ορίζεται για παράδειγμα το functional classification (σε personal page, registration page κλπ)
  - Σε features που χρησιμοποιούνται
    - Η σελίδα είναι ημι-δομημένο κείμενο σε HTML που μπορεί να ληφθεί υπόψη
    - Μπορούν να χρησιμοποιηθούν και τα links μεταξύ σελίδων
      - Πληροφορία που λαμβάνεται από γειτονικές σελίδες
  - Σε training set
    - Είναι εξαιρετικά δύσκολο να δημιουργηθεί αντιπροσωπευτικό σώμα εκπαίδευσης για το web!

Supervised Classification Algorithms

- Decision Trees
- Nearest Neighbors (kNN, weighted kNN)
- Relevance Feedback (Rocchio)
- Naive Bayes
- Support Vector Machines
- Ensemble

Δέντρα Απόφασης

- Αλγόριθμος Επαγωγής (ID3)
  - Κατασκευάζει ένα δέντρο απόφασης από ένα σύνολο δειγμάτων εκπαίδευσης
- Έστω $TS=\{T_1,T_2,T_3,\ldots,T_n\}$ το σύνολο των δειγμάτων εκπαίδευσης.
- Κάθε δείγμα εκπαίδευσης Τ
  - Περιγράφεται από ένα σύνολο χαρακτηριστικών

    $\boldsymbol{FS} = \{\boldsymbol{F_1}, \boldsymbol{F_2}, \boldsymbol{F_3}, \ldots \boldsymbol{F_{|FS|}}\}$
  - Ανήκει σε μια κατηγορία **c**

    $\boldsymbol{C} = \{\boldsymbol{c_1}, \boldsymbol{c_2}, \boldsymbol{c_3}, \ldots \boldsymbol{c_{|C|}}\}$

$$\boldsymbol{T} = \left(\boldsymbol{v}_{F_1}, \boldsymbol{v}_{F_2}, \boldsymbol{v}_{F_3}, \ldots \boldsymbol{v}_{F_{|FS|}}, C \right)$$

Επαγωγή Δέντρων Απόφασης – ID3

- Γενική Ιδέα του ID3:
  - Για όλες τις αχρησιμοποίητα χαρακτηριστικά υπολόγισε την εντροπία σε σχέση με τα δείγματα.
  - Διάλεξε το χαρακτηριστικό που παρουσιάζει την ελάχιστη εντροπία (ή μέγιστο κέρδος πληροφορίας)
  - Φτιάξε κόμβο γι' αυτό το χαρακτηριστικό
- Ο αλγόριθμος βασίζεται στις έννοιες:
  - εντροπία πληροφορίας (information entropy)
  - κέρδος πληροφορίας (information gain)

Εντροπία Πληροφορίας (Information Entropy)

- Έστω S ένα σύνολο δεδομένων
  - Εντροπία Πληροφορίας
    - Χαρακτηρίζει το βαθμό αβεβαιότητας

$$E(S) = \sum_{i=1}^{n} p_i * \log_2 \frac{1}{p_i}$$

όπου $p_1, p_2, \ldots p_i$ οι πιθανότητες του κάθε ενδεχομένου που περιλαμβάνεται στο σύνολο

  - πχ. έστω δοχείο με N μπάλες: N*p λευκές και N(1-p) μαύρες
    - Αν όλες μαύρες ή όλες άσπρες => Εντροπία=0
    - Αν p=50% => Εντροπία=1 (μέγιστη)

Πληροφοριακό Κέρδος (Information Gain)

- Έστω χαρακτηριστικό A σε ένα S σύνολο δεδομένων
  - Κέρδος Πληροφορίας
    - Χαρακτηρίζει το πόση πληροφορία «φέρει» ένα χαρακτηριστικό

$$G(S,A) = E(S) - \sum_{i=1}^{m} f_s(A_i) * E(S_{A_i})$$

όπου

- **E(...)** η συνάρτηση εντροπίας
- **m** το πλήθος των τιμών που παίρνει το A στο S
- **f(Ai)** το ποσοστό των αντικειμένων στο S που παίρνουν την τιμή A_i
- $\mathbf{S_{Ai}}$ το υποσύνολο του S όπου η τιμή του A είναι A_i

Αλγόριθμος ID3

Κατασκεύασε_Δέντρο_1(Κόμβος N, Σύνολο_Δειγμάτων_Εκπαίδευσης TS, Σύνολο_Χαρακτηριστικών FS **)**

{

1. Βρες την πιο συχνή κατηγορία c στο TS και εκχώρησέ την στον κόμβο N;
2. Αν (η c είναι η μόνη κατηγορία στο TS) ή ($FS == \varnothing$) Τότε
3. Επίστρεψε; /* ο N είναι φύλλο */
4. Βρες το χαρακτηριστικό F από το FS με το μεγαλύτερο gain ratio και εκχώρησέ το στον κόμβο N;
5. Για κάθε τιμή v_i του F :

{

6. Συγκέντρωσε στο υποσύνολο TS_i όλα τα δείγματα του TS που έχουν την τιμή v_i για το χαρακτηριστικό F;
7. Αν $TS_i \neq \varnothing$ Τότε

{

8. Δημιούργησε ένα νέο κόμβο N' κάτω από την τιμή v_i του κόμβου N;
9. $FS' = FS - \{F\}$;
10. **Κατασκεύασε_Δέντρο_1(** N', TS_i, FS' **)**;

}

}

}

Δημιουργία Δέντρου Απόφασης - Παράδειγμα

- Θέλουμε δέντρο απόφασης για διάρκεια άθλησης στην ύπαιθρο ανάλογα με τον καιρό.
- Χαρακτηριστικά (FS):
  - ουρανός = {καθαρός, συννεφιά, βροχή}
  - θερμοκρασία = {υψηλή,μέτρια,χαμηλή}
  - υγρασία = {υψηλή, κανονική}
  - άνεμος = {δυνατός, αδύναμος}
- Κατηγορίες (C):
  - διάρκεια άθλησης = {μικρή,κανονική,καμία}

ΔΕΙΓΜΑΤΑ(TS)

Ουρανός	Θερμοκρασία	Υγρασία	Άνεμος	Διάρκεια
καθαρός	υψηλή	υψηλή	αδύναμος	μικρή
καθαρός	υψηλή	υψηλή	δυνατός	μικρή
συννεφιά	υψηλή	υψηλή	αδύναμος	κανονική
βροχή	μέτρια	υψηλή	αδύναμος	καμία
βροχή	χαμηλή	κανονική	αδύναμος	καμία
βροχή	χαμηλή	κανονική	δυνατός	καμία
συννεφιά	χαμηλή	κανονική	δυνατός	κανονική
καθαρός	μέτρια	υψηλή	αδύναμος	μικρή
καθαρός	χαμηλή	κανονική	αδύναμος	κανονική
βροχή	μέτρια	κανονική	δυνατός	καμία
καθαρός	μέτρια	κανονική	δυνατός	κανονική
συννεφιά	μέτρια	υψηλή	αδύναμος	κανονική
συννεφιά	υψηλή	κανονική	αδύναμος	κανονική
βροχή	μέτρια	υψηλή	δυνατός	καμία

Δημιουργία Δέντρου Απόφασης - Παράδειγμα

- Παραδείγματα υπολογισμών τιμών:
  - Εντροπία του συνόλου δεδομένων:

$$E(S) = \sum_{i=1}^{n} p_i * \log_2 \frac{1}{p_i} = \frac{3}{14}\log_2 \frac{14}{3} + \frac{6}{14}\log_2 \frac{14}{6} + \frac{5}{14}\log_2 \frac{14}{5}$$

όπου

- 3/14: η πιθανότητα να είναι η διάρκεια μικρή
- 6/14: η πιθανότητα να είναι κανονική
- 5/14: η πιθανότητα να είναι μηδενική

Δημιουργία Δέντρου Απόφασης - Παράδειγμα

- Παραδείγματα υπολογισμών τιμών:
  - Εντροπία του χαρακτηριστικού «Ουρανός»

$$E(Ουρανός) = \frac{5}{14}E(καθαρός) + \frac{4}{14}E(συννεφιά) + \frac{5}{14}E(βροχή)$$

όπου:

- 5/14: το ποσοστό των τιμών «καθαρό» στο S
- 4/14: το ποσοστό των τιμών «συννεφιά» στο S
- 5/14: το ποσοστό των τιμών «βροχή» στο S

Δημιουργία Δέντρου Απόφασης - Παράδειγμα

- Παραδείγματα υπολογισμών τιμών:
  - Εντροπία της τιμής «καθαρός» του χαρακτηριστικού «Ουρανός»

    όπου: $$E(\kappa\alpha\theta\alpha\rho\acute{o}\varsigma) = \frac{3}{5}\log_2\frac{5}{3} + \frac{2}{5}\log_2\frac{5}{2} + 0$$
    - 3/5: πιθανότητα όταν η τιμή είναι «καθαρός» η διάρκεια να είναι μικρή
    - 2/5: πιθανότητα η διάρκεια να είναι κανονική
    - 0: η πιθανότητα η διάρκεια να είναι μηδενική
  - Κέρδος πληροφορίας χαρακτηριστικού «Ουρανός»

$$Gain(S, Ουρανός) = E(S) - E(Ουρανός)$$

Recommended Reading

- "Natural Language Processing with Python"
  - Chapter 6: Learning to Classify Text