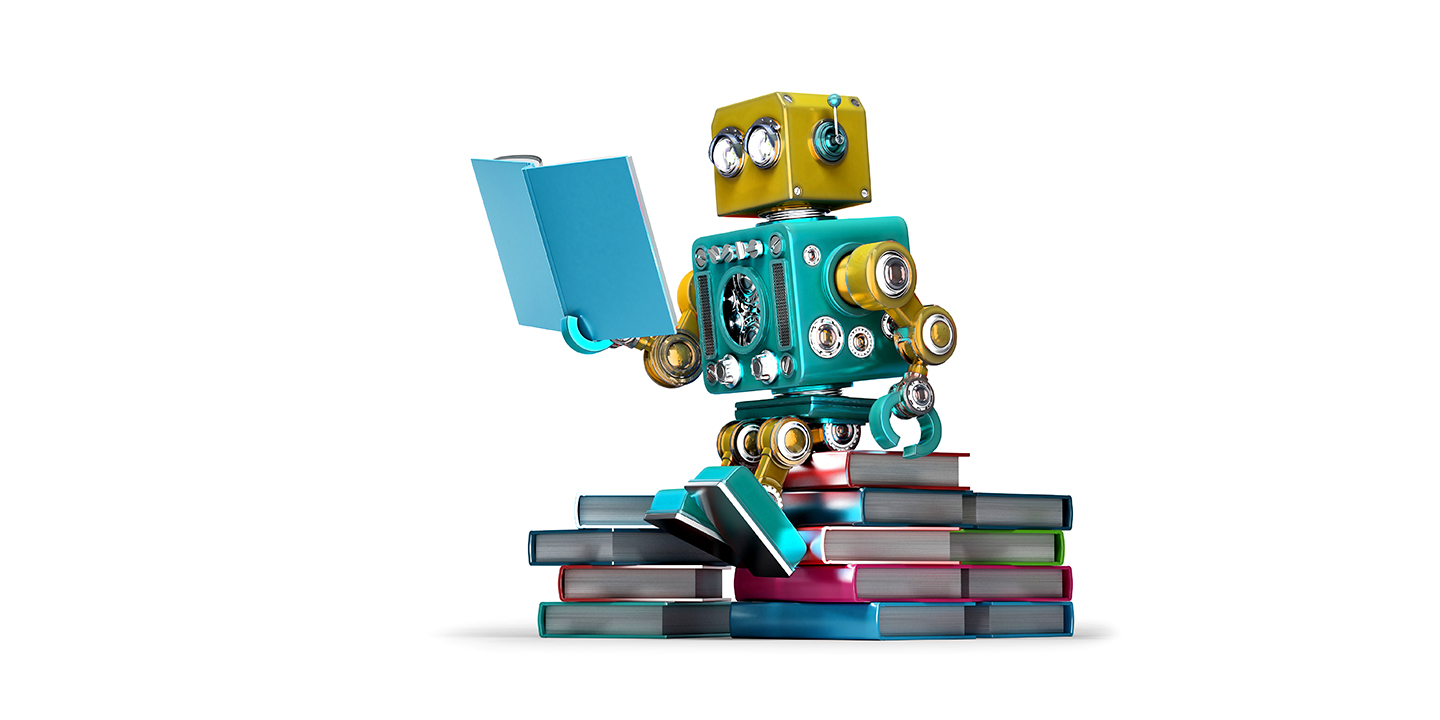
**Μηχανική Μάθηση Εργασία 2**



**Μπόζας Αριστείδης**

**ΑΜ:740**

**Δερμεντζόγλου Ιωάννης**

**ΑΜ:743**

**Περιεχόμενα**

[1. Μέρος Α 4](#_Toc511454016)

[2. Μέρος Β 4](#_Toc511454017)

[3. Μέρος Γ 5](#_Toc511454018)

[Παράρτημα κώδικα 1 5](#_Toc511454019)

[Παράρτημα κώδικα 2 5](#_Toc511454020)

[Παράρτημα κώδικα 3 5](#_Toc511454021)

Κατάλογων Πινάκων

Κατάλογων Εικόνων

# Μέρος Α

Στο πρώτο μέρος της εργασίας εξετάστηκε το αντικείμενο της μάθησης από δεδομένα πολλαπλών ετικετών. Το dataset που χρησιμοποιείται είναι το σύνολο δεδομένων [DeliciousMIL](https://github.com/hsoleimani/MLTM/tree/master/Data) το οποίο περιέχει δεδομένα πολλαπλών ετικετών.

Το Dataset καθαρίστηκε από τα που είχε και έμειναν μόνο η αναπαράσταση κάθε εγγράφου αφορά στις λέξεις που περιλαμβάνει (bag of words) και δημιουργήσαμε διανύσματα. Αλλά επειδή είχαν διαφορετικό μήκος τα διανύσματα για να γίνουν όλο το ίδιο μήκος, πήραμε το μεγαλύτερο διάνυσμα και βάση αυτό προσθέσαμε μηδενικά στα υπόλοιπα μηδενικά.

Για την επίλυση και το πρόβλημα της ταξινόμησης πολλαπλών ετικετών υπάρχουν τρεις μέθοδοι:

1. **Πρόβλημα μετασχηματισμού (Problem Transformation)**
2. **Προσαρμοσμένος Αλγόριθμος (Adapted Algorithm)**
3. **Ensemble approaches**

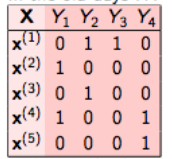
Στη μέθοδο του προβλήματος μετασχηματισμού που θα ασχοληθούμε σε αυτή την εργασία προσπαθούμε να μετατρέψουμε το πρόβλημα πολλαπλών ετικετών σε πρόβλημα μία ετικέτας. Αυτό μπορεί να επιτευχθεί με τις τρεις παρακάτω μεθόδους:

1. **Binary Relevance**
2. **Classifier Chains**
3. **Label Powerset**

## 1.1 Binary Relevance

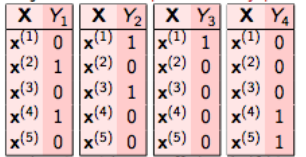
Αυτή είναι η απλούστερη τεχνική, η οποία βασικά αντιμετωπίζει κάθε ετικέτα ως ξεχωριστό πρόβλημα ταξινόμησης μια ενιαίας κατηγορίας.

Για παράδειγμα, έχουμε ένα σύνολο δεδομένων όπου το **Χ** είναι ανεξάρτητο χαρακτηριστικό και η **Υ** είναι μεταβλητή στόχος.



Εικόνα 1:Πρόβλημα πριν τη μέθοδο Binary Relevance

Με τη μέθοδο **Binary Relevance** ,το πρόβλημα χωρίζεται σε 4 διαφορετικά προβλήματα ταξινόμησης μιας μόνο κλάσης όπως φαίνεται στην εικόνα 2

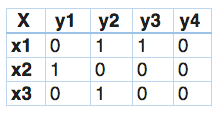


Εικόνα 2: Πρόβλημα μετά τη μέθοδο Binary Relevance

## 1.2 Classifier Chains

Στη μέθοδο αυτή, ο πρώτος ταξινομητής εκπαιδεύεται μόνο στα δεδομένα εισόδου και στη συνέχεια κάθε επόμενος ταξινομητής εκπαιδεύεται στο χώρο εισόδου και όλοι οι προηγούμενοι ταξινομητές στην αλυσίδα.

Για παράδειγμα, στο παρακάτω σύνολο δεδομένων έχουμε το Χ ως είσοδο και το Υ ως τις ετικέτες



Εικόνα 3: Πρόβλημα πριν τη μέθοδο Classifier Chains

Με τη μέθοδο αυτή, το πρόβλημα θα μετατραπείς σε τέσσερα διαφορετικά προβλήματα μόνο μίας ετικέτας, όπως φαίνεται παρακάτω.

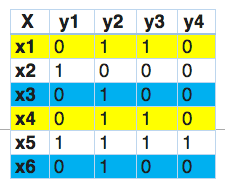


Εικόνα 4: Πρόβλημα μετά τη μέθοδο Classifier Chains

## 1.3 Label Powerset

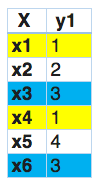
Στη μέθοδο αυτή ,μετατρέπουμε το πρόβλημα σε ένα πρόβλημα πολλαπλών τάξεων με έναν ταξινομητή πολλαπλών τάξεων που εκπαιδεύεται σε όλους τους μοναδικούς συνδυασμούς ετικετών που βρίσκονται στα δεδομένα εκπαίδευσης

Για παράδειγμα,



Εικόνα 5: Πρόβλημα πριν τη μέθοδο Label Powerset

Στο παραπάνω πρόβλημα φαίνεται ότι τα χ1 και χ3 έχουν τις ίδιες ετικέτες, ομοίως, τα χ3 και χ4 έχουν το ίδιο σύνολο ετικετών. Έτσι, η μέθοδος μετατρέπει το πρόβλημα σε ένα ενιαίο πρόβλημα πολλαπλών κατηγοριών όπως φαίνεται παρακάτω:



Εικόνα 6: Πρόβλημα μετά τη μέθοδο Label Powerset

## 1.4 Αποτελέσματα

Οι μετρικές που χρησιμοποιήσαμε είναι:

* Subset Accuracy:
* Hamming loss:
* F1 macro:
* F1 micro:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithms** | **Example Based metrics** | | **Label Based metrics** | |
| **Accuracy** | **Hamming loss** | **F1 macro** | **F1 micro** |
| **Binary Relevance - SVC** | 0.058750 | 0.144175 | 0.024628 | 0.027548 |
| **Classifier Chains - SVC** |  |  |  |  |
| **Label Power set - SVC** |  |  |  |  |
| **Binary Relevance - Naive Bayes** | 0.000000 | 0.458436 | 0.241542 | 0.227292 |
| **Classifier Chains - Naive Bayes** | 0.000000 | 0.458411 | 0.241646 | 0.227397 |
| **Label Power set - Naive Bayes** | 0.001506 | 0.252410 | 0.209972 | 0.165988 |
| **Binary Relevance - DesicionTree** | 0.005272 | 0.237836 | 0.218657 | 0.172143 |
| **Classifier Chains - DesicionTree** | 0.010043 | 0.239355 | 0.215705 | 0.168589 |
| **Label Power set - DesicionTree** | 0.013558 | 0.223550 | 0.219837 | 0.171287 |

Πίνακας 2: κόστους

## 1.5 Συμπεράσματα

# Μέρος Β

# Μέρος Γ