MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Πανεπιστήμιο Κύπρου - Τεχνητή Νοημοσύνη

MAI612 - ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Διάλεξη 2: Προετοιμασία δεδομένων

Βασίλης Βασιλειάδης, PhD

Χειμερινό Εξάμηνο 2022/23

CYENS CENTRE OF EXCELLENCE

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Διάλεξη 2: Προετοιμασία δεδομένων

Μαθησιακά αποτελέσματα

Θα καταλάβετε:

1. τη διαδικασία συλλογής δεδομένων και προετοιμασίας συνόλων δεδομένων για την ανάπτυξη λύσεων μηχανικής μάθησης
2. τη σημασία της προετοιμασίας των δεδομένων
3. τα στάδια της προεπεξεργασίας δεδομένων και του μετασχηματισμού δεδομένων
4. τη σημασία της απεικόνισης των δεδομένων και της διερευνητικής ανάλυσης δεδομένων
5. γιατί χωρίζουμε ένα σύνολο δεδομένων σε σύνολο εκπαίδευσης και σύνολο δοκιμών

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Ταξινομητής εικόνων φρούτων με χρήση Μηχανικής Μάθησης (ΜΜ)

Είστε μηχανικός ΜΜ στην εταιρεία XYZ και σας ανατέθηκε να δημιουργήσετε μια εφαρμογή για κινητά που ταξινομεί διαφορετικές εικόνες φρούτων, στις αντίστοιχες κατηγορίες τους (μήλο, πορτοκαλί κ.λπ.).

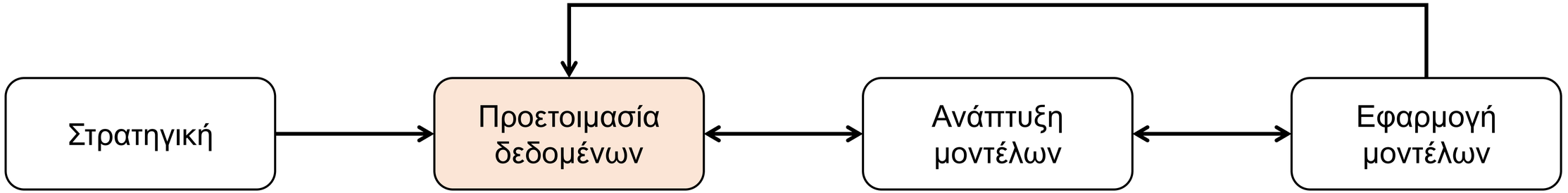
Θέλετε να χρησιμοποιήσετε τη ΜΜ

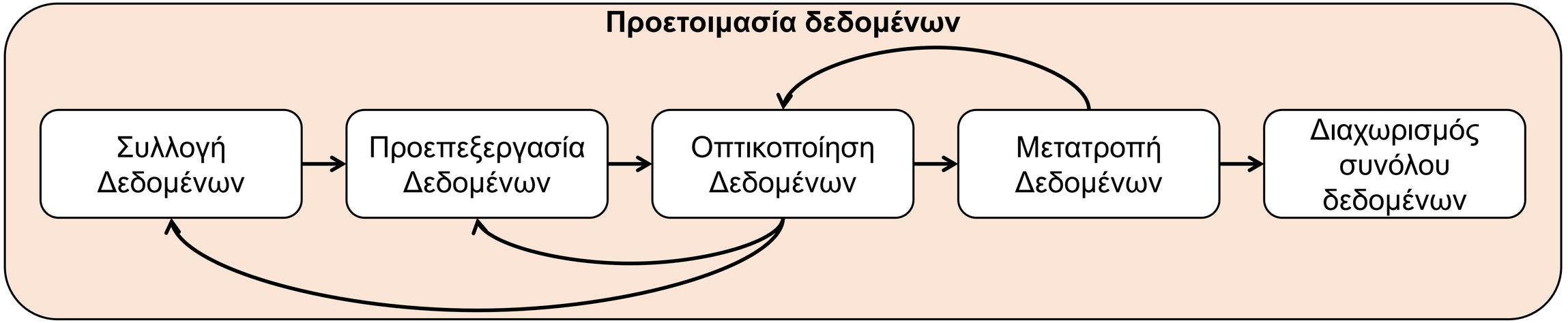
Ποιο είναι το πρώτο βήμα της διαδικασίας;
- Συλλογή δεδομένων

Πώς;
- Στο διαδίκτυο
- Τραβήξτε φωτογραφίες και συνδέστε τις με ετικέτες

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

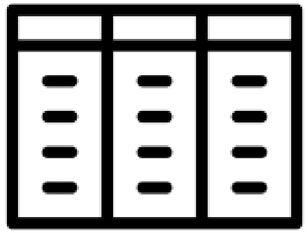
Προετοιμασία δεδομένων





This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Δεδομένα

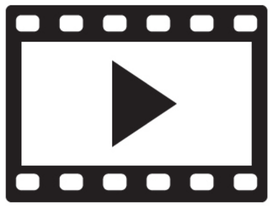


πίνακας

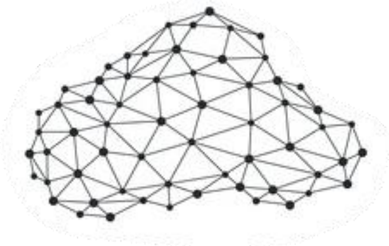
κείμενο

σήματα

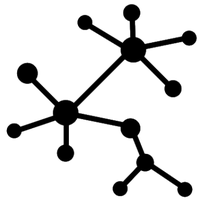
εικόνες



βίντεο



point clouds



γράφοι

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Συλλογή δεδομένων

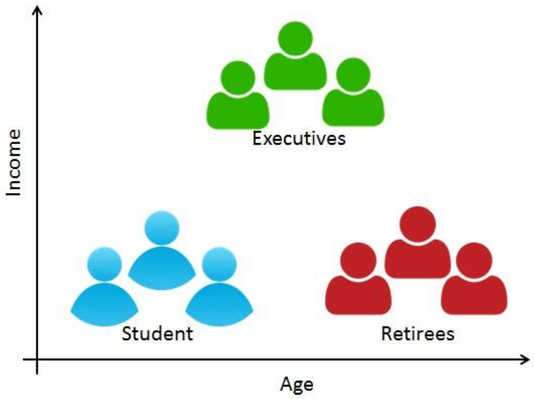
Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Συλλογή δεδομένων

Ενδέχεται να χρειαστεί να αποκτηθούν δεδομένα για τη συγκεκριμένη εργασία

- εικόνες φρούτων για ταξινομητή εικόνων φρούτων



- ερωτηματολόγιο με δημογραφικά στοιχεία για την κατάτμηση των πελατών

- ένας ρομποτικός βραχίονας μαθαίνει να συλλαμβάνει διαφορετικά αντικείμενα

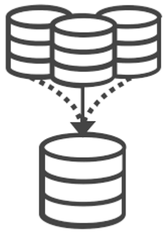
Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Συλλογή δεδομένων

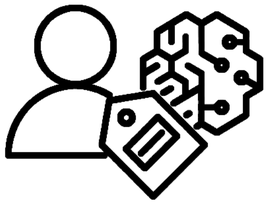
Τα δεδομένα υπάρχουν ήδη:

- μπορεί να χρειαστεί να ενσωματωθούν από πολλαπλές πηγές



π.χ. στον εντοπισμό απάτης: χρονοδιάγραμμα μεταξύ δύο διαδοχικών συναλλαγών και απόσταση μεταξύ των τόπων όπου πραγματοποιήθηκαν οι συναλλαγές

- μπορεί να χρειαστεί να επισημανθούν
  - όλες οι εποπτευόμενες ρυθμίσεις μάθησης
  - απαιτεί ανθρώπινη προσπάθεια (συχνά ειδικοί τομέα)
  - μπορούν να ανατεθούν σε εξωτερικούς συνεργάτες (υπηρεσίες επισήμανσης δεδομένων, π.χ., Amazon Mechanical Turk)
  - μεταφορά της μάθησης: χρησιμοποιήστε ένα ήδη εκπαιδευμένο μοντέλο με τις ετικέτες του

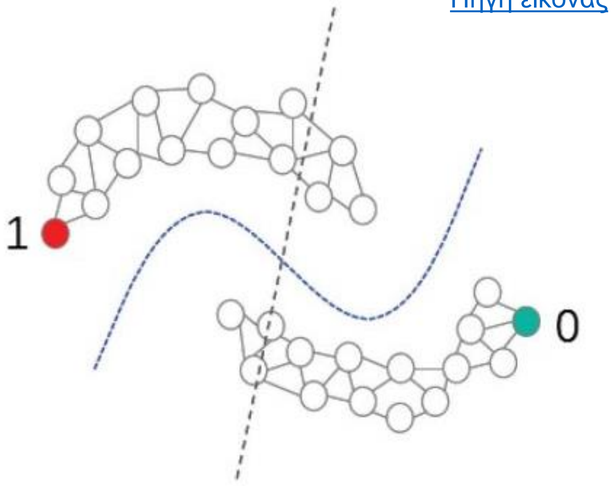


This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Συλλογή δεδομένων

Πιο προηγμένοι τύποι ΜΜ που σχετίζονται με τη συλλογή δεδομένων:

Πηγή εικόνας



Ημι-εποπτευόμενη μάθηση

- Μεγάλος όγκος μη εποπτευόμενων δεδομένων, μικρή ποσότητα δεδομένων με επισήμανση
- Μεταξύ μη εποπτευόμενης και εποπτευόμενης μάθησης

Ενεργός Μάθηση

- Χρησιμοποιείται όταν η επισήμανση είναι δαπανηρή
- Επιλέξτε έξυπνα το επόμενο σημείο δεδομένων για να επισημάνετε με τρόπο που ελαχιστοποιεί την προσπάθεια επισήμανσης

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Συλλογή δεδομένων

Άλλα θέματα:

Προστασία προσωπικών δεδομένων

- Ανάγκη να εργαστείτε με ανωνυμοποιημένα δεδομένα
- Το στάδιο της προεπεξεργασίας αφορά ευαίσθητα δεδομένα

Ανισορροπία κλάσεων

- Πρέπει να έχουν περίπου τον ίδιο αριθμό περιπτώσεων για όλες τις κλάσεις σε εργασίες ταξινόμησης

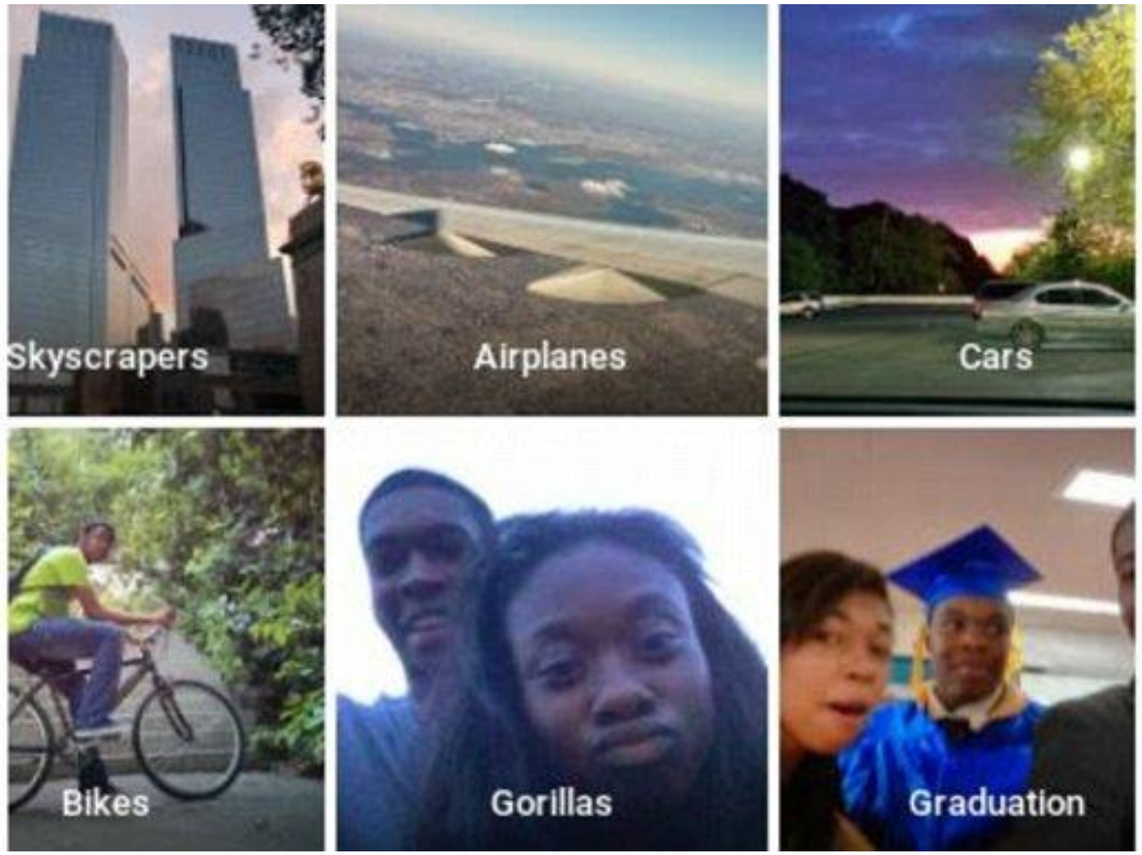
Προκατάληψη & Δικαιοσύνη

- Είναι πολύ σημαντικό να έχουμε ποικιλομορφία στα δεδομένα μας
- Δείτε τις επόμενες διαφάνειες

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Προκατάληψη & δικαιοσύνη: Το περιστατικό των «Gorillas»

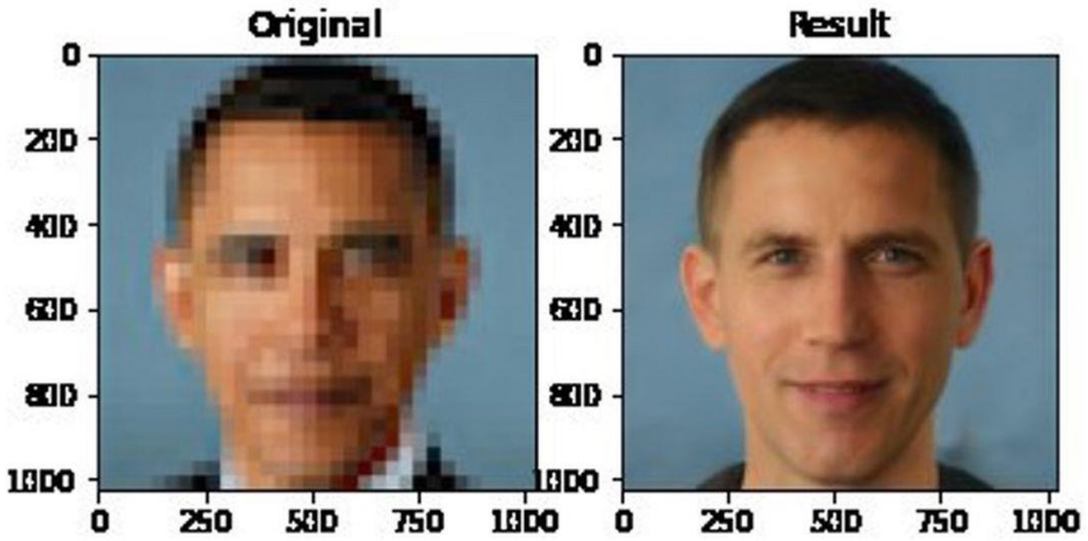


Μία από αυτές τις προβλέψεις είναι πολύ λάθος!

Γιατί νομίζετε ότι συνέβη αυτό;

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Προκατάληψη & δικαιοσύνη: Ο Ομπάμα ανακατασκευάστηκε ως λευκός άνδρας



Το σύστημα αναδειγματοληψίας προσώπου κάνει όλους να φαίνονται λευκοί, επειδή εκπαιδεύτηκε σε ένα σύνολο δεδομένων που περιέχει κυρίως εικόνες λευκών ανθρώπων.

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Προκατάληψη & δικαιοσύνη: Ποσοστό σφάλματος ανά φύλο και φυλή

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Όλα τα μοντέλα έχουν περισσότερο πρόβλημα να ταξινομήσουν σωστά πιο έγχρωμες γυναίκες από λευκούς άντρες.

- Οι ταξινομητές δεν εκπαιδεύτηκαν με επαρκή ποικιλία
- Σημασία της ύπαρξης ισορροπημένων συνόλων δεδομένων

Το μοντέλο σας είναι τόσο καλό όσο τα δεδομένα σας!

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Δεδομένα πίνακα

Μεταβλητή/Στοιχείο/Χαρακτηριστικό

Παράδειγμα/περίπτωση

Στήλη 1	Στήλη 2	Στήλη 3	Στήλη 4

Μεταβλητές εισόδου: στήλες που παρέχονται στο μοντέλο για την πρόβλεψη

Μεταβλητή εξόδου: στήλη που προβλέπεται από το μοντέλο

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Μεταβλητοί τύποι δεδομένων

Ποιοτικός/Κατηγορικός

Φτιαγμένη από λέξεις

1. Ονομαστική ονομασία (ονομαζόμενες κατηγορίες)
   - π.χ. φύλο

2. Ordinal (Θέματα Τάξης)
   - π.χ. διάθεση

Ποσοτικός/Αριθμητικός

Φτιαγμένη από αριθμούς

1. Διακριτό (integers)
   - π.χ. μέγεθος παπουτσιών

2. Συνεχής (οποιοσδήποτε αριθμός)
   - π.χ. βάρος

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Στοιχεία μεταβλητοί τύποι κουίζ

Κατηγορηματικό (ονομαστικό, προφορικό) ή αριθμητικό (διακριτό, συνεχές);

1. Εθνικότητα	Ονομαστικό
2. Ηλικία	Συνεχής ή διακριτή
3. Χρώμα	Ονομαστικό
4. Θερμοκρασία	Συνεχής
5. Αριθμός παιδιών	Διακριτή
6. Βαθμολογία ικανοποίησης (δυστυχής/ουδέτερη/ευτυχής)	Ordinal
7. Ημερομηνία	String: Ανάγκη μετασχηματισμού χαρακτηριστικών

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Προεπεξεργασία δεδομένων

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Προεπεξεργασία δεδομένων

Τι είναι η προεπεξεργασία δεδομένων;

- Διαδικασία εισαγωγής **ακατέργαστων** δεδομένων σε μορφή κατάλληλη για μοντελοποίηση
- Τυπικά περιλαμβάνει: Καθαρισμός δεδομένων, κωδικοποίηση δεδομένων

Γιατί είναι σημαντικό;

Οι αλγόριθμοι ΜΜ αναμένουν αριθμούς (π.χ. ορισμένοι δεν μπορούν να ασχοληθούν με κατηγορηματικά χαρακτηριστικά)

- Οι αλγόριθμοι ΜΜ έχουν απαιτήσεις (π.χ. δεν μπορούν να αντιμετωπίσουν τις ελλείπουσες τιμές)

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Προεπεξεργασία δεδομένων

Καθαρισμός δεδομένων

- Διορθώστε σφάλματα: τυπογραφικά λάθη, εσφαλμένη κεφαλαιοποίηση, ασυνέπειες κ.λπ. (π.χ. «Α/Α» έναντι «Άνευ αντικειμένου», «Άνδρας» έναντι «αρσενικού»,...)
- Ελλείποντα δεδομένα (αφαίρεση, αντικατάσταση με 0 ή παρεμβολή)
- Αφαίρεση διπλών γραμμών/στήλων
- Αφαίρεση άσχετων δεδομένων (π.χ. κατά την ανάλυση δεδομένων για πελάτες χιλιετηρίδων, αφαίρεση παλαιότερων γενεών)
- Αφαίρεση ακραίων τιμών (μπορεί να οφείλεται σε σφάλμα στη συλλογή δεδομένων)

Κωδικοποίηση δεδομένων

- Κατηγορηματικές μεταβλητές
- Ανωνυμοποίηση δεδομένων

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Κωδικοποίηση δεδομένων

Κωδικοποιήστε τις **κατηγορηματικές** μεταβλητές χρησιμοποιώντας μια αριθμητική αναπαράσταση. Τυπικά:

- Nominal variables: κωδικοποίηση«one-hot»: όσες νέες δυαδικές μεταβλητές έχουν ονομαστεί κατηγορίες

Φύλο = {αρσενικό, θηλυκό} -> GenderMale = {0,1}, GenderFemale = {0,1}
Χρώμα = {κόκκινο, πράσινο, μπλε} -> ColorRed = {0,1}, ColorGreen = {0,1}, ColorBlue = {0,1}

- Τακτικές μεταβλητές: τακτικός κωδικοποίηση: μετατροπή σε ακέραιες τιμές

Ικανοποίηση = {sad,neutral,happy} -> {0,1,2}

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Κωδικοποίηση στοιχείων: One-hot παράδειγμα κωδικοποίησης

```
In [1]: # example of a one-hot encoding
   ...: from numpy import asarray
   ...: from sklearn.preprocessing import OneHotEncoder
   ...:
   ...: # define data
   ...: data = asarray([['red'], ['green'], ['blue']])
   ...: print(data)
   ...:
   ...: # define one hot encoding
   ...: encoder = OneHotEncoder(sparse=False)
   ...:
   ...: # transform data
   ...: onehot = encoder.fit_transform(data)
   ...: print(onehot)
[['red']
 ['green']
 ['blue']]
[[0. 0. 1.]
 [0. 1. 0.]
 [1. 0. 0.]]
```

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Κωδικοποίηση στοιχείων: Ordinal παράδειγμα κωδικοποίησης

```
In [1]: # example of an ordinal encoding
   ...: from numpy import asarray
   ...: from sklearn.preprocessing import OrdinalEncoder
   ...:
   ...: # define data
   ...: satisfaction_rating = asarray([['sad'], ['neutral'], ['happy']])
   ...: print(satisfaction_rating)
   ...:
   ...: # define ordinal encoding
   ...: encoder = OrdinalEncoder()
   ...:
   ...: # transform to numbers
   ...: result = encoder.fit_transform(satisfaction_rating)
   ...: print(result)
[['sad']
 ['neutral']
 ['happy']]
[[2.]
 [1.]
 [0.]]
```


This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Οπτικοποίηση δεδομένων

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Οπτικοποίηση δεδομένων

Ανάλυση διερευνητικών δεδομένων

- υπάρχουν εμφανή μοτίβα (μπορεί να δώσει συμβουλές σχετικά με τη μέθοδο για να επιλέξετε);
- υπάρχουν προφανή προβλήματα (θόρυβος ετικέτας ή υπερβολικές τιμές);
- περιλαμβάνει βασικά στατιστικά στοιχεία (π.χ. μέσος όρος, διακύμανση, συντελεστής συσχέτισης) και γραφήματα

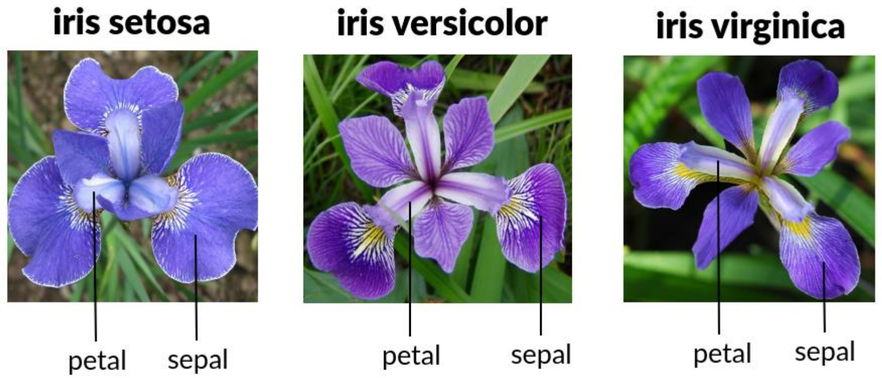
- Δεδομένα πίνακα με μικρό αριθμό χαρακτηριστικών: χρησιμοποιήστε **pair plot**

- Υψηλής διάστασης στοιχεία: χρήση της **μείωσης της διάστασης**
  - π.χ., ανάλυση βασικών στοιχείων
  - θα μελετήσει σε μεταγενέστερες διαλέξεις

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

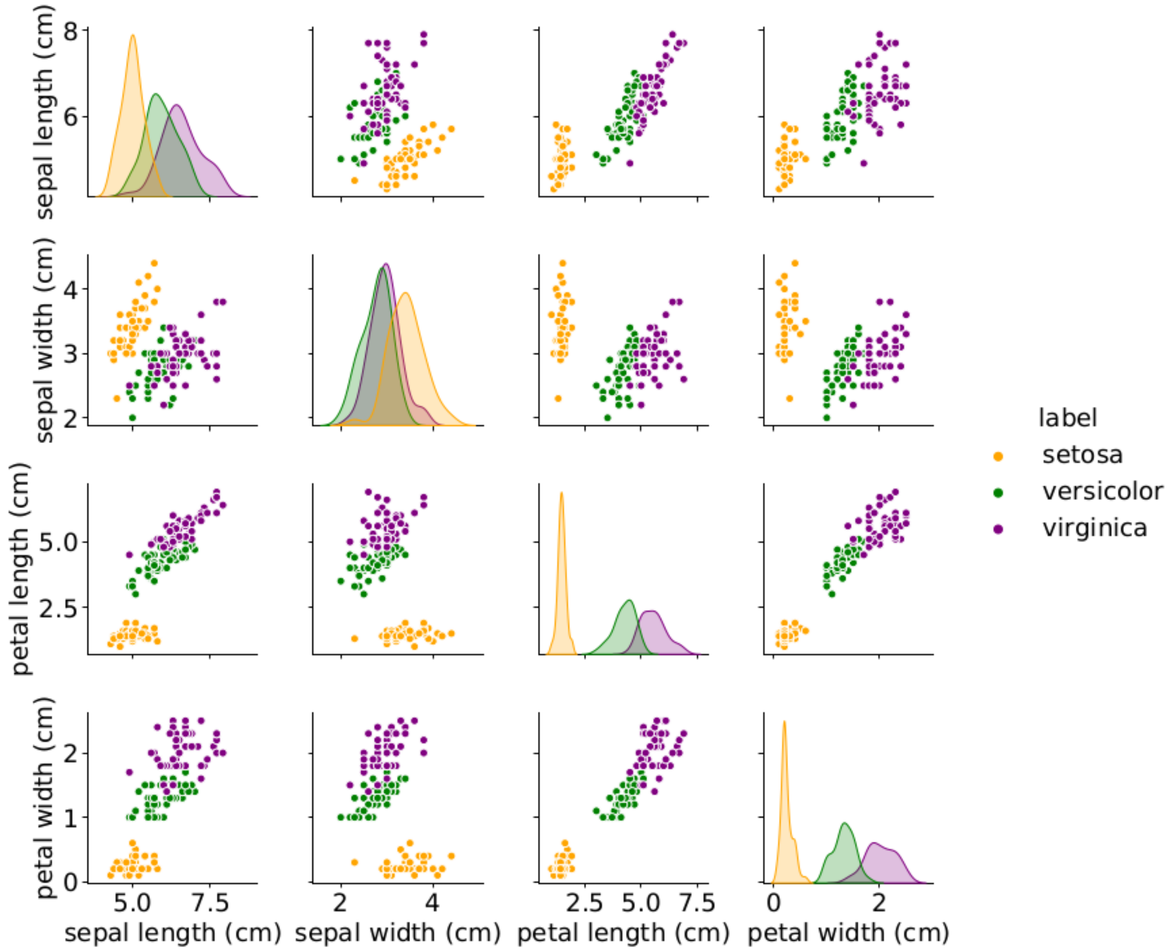
Σύνολο δεδομένων Iris

- 3 κλάσεις των 50 περιπτώσεων η κάθε μία
- 4 χαρακτηριστικά (μήκος/πλάτος πέτου/σεπάλ σε cm)



Ποια τάξη είναι γραμμικά διαχωρίσιμη από τις άλλες;

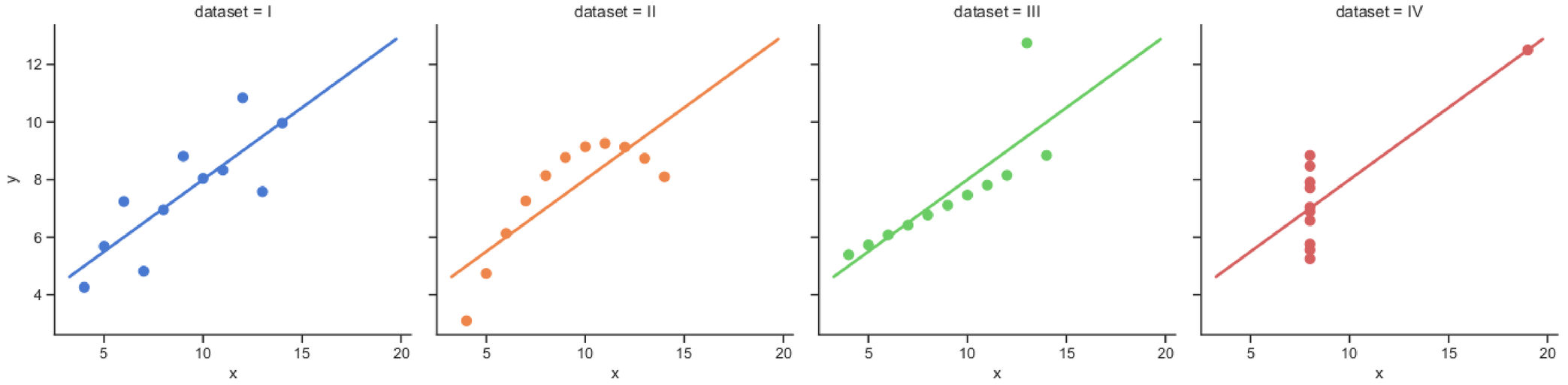
Πηγή: Μέρφι, Κ. (2022). Πιθανοτική μηχανική μάθηση — μια εισαγωγή. MIT Press



This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Σημασία της απεικόνισης των δεδομένων σε σχέση με τις συνοπτικές στατιστικές

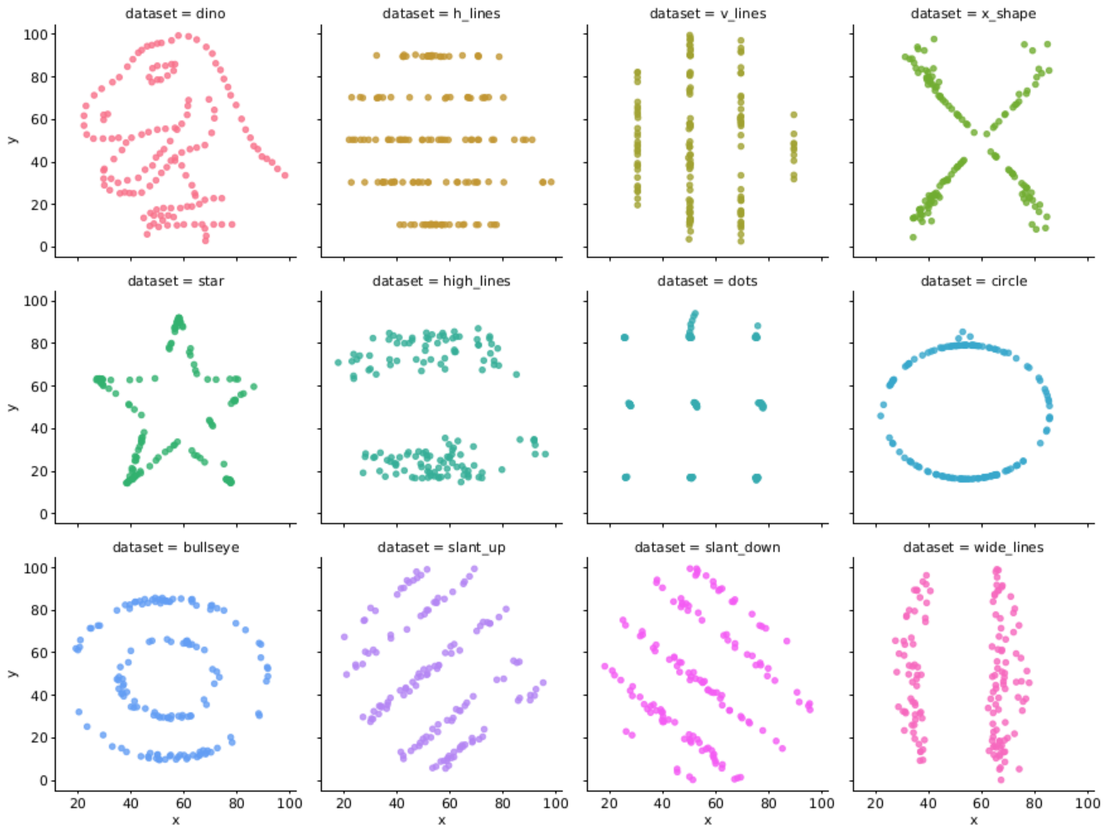
Πηγή: Murphy, K. (2022). Probabilistic Machine Learning - An Introduction. MIT Press



Το κουαρτέτο του Anscombe: Όλα αυτά τα σύνολα δεδομένων έχουν τον ίδιο μέσο όρο, συντελεστή διακύμανσης και συσχέτισης.

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Σημασία της απεικόνισης των δεδομένων σε σχέση με τις συνοπτικές στατιστικές



Πηγή: Murphy, K. (2022). Probabilistic Machine Learning - An Introduction. MIT Press

Ντουζίνα Datasaurus: Όλα αυτά τα σύνολα δεδομένων έχουν τον ίδιο μέσο όρο, συντελεστή διακύμανσης και συσχέτισης.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Μετασχηματισμός δεδομένων

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Μετασχηματισμός δεδομένων

Τι είναι ο μετασχηματισμός των δεδομένων;

- Διαδικασία μετασχηματισμού δεδομένων σε μορφή **καταλληλότερη** για μοντελοποίηση
- Τυπικά περιλαμβάνει: Κλιμάκωση Χαρακτηριστικών, Επιλογή Χαρακτηριστικών, Εξαγωγή Χαρακτηριστικών και Κατασκευή Χαρακτηριστικών (ή Μηχανική)

Μερικές φορές: Αύξηση δεδομένων, Δειγματοληψία δεδομένων

Γιατί είναι σημαντικό;

- Οι αλγόριθμοι ΜΜ έχουν απαιτήσεις (π.χ., η πολυπλοκότητα ορισμένων αλγορίθμων κλιμάκων με αριθμό σημείων δεδομένων)
- Η απόδοση του μοντέλου εξαρτάται από τα δεδομένα (π.χ., διαφορετικές περιοχές χαρακτηριστικών επηρεάζουν πολλούς αλγορίθμους)

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Χαρακτηριστικό γνώρισμα Κλιμάκωση

Ελάχιστη-μέγιστη κανονικοποίηση (ή γραμμική κλιμάκωση):

- χρησιμοποιημένος όταν το χαρακτηριστικό γνώρισμα κατανέμεται ομοιόμορφα σε μια σταθερή περιοχή
- συνήθως κλιμακώνεται στο εύρος [0,1]
- μπορεί να επηρεαστεί σε μεγάλο βαθμό από τις ακραίες τιμές

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Τυποποίηση (ή εξομάλυνση z-score):

- χρησιμοποιείται όταν δεν γνωρίζουμε τη σειρά χαρακτηριστικών
- οι κλιμακούμενες κατανομές χαρακτηριστικών έχουν μέσο = 0 και sd=1
- λιγότερο επηρεαζόμενες από ακραίες τιμές

$$x_{scaled} = \frac{x - mean}{sd}$$

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επιλογή χαρακτηριστικών

Διαδικασία που επιλέγει ένα υποσύνολο χαρακτηριστικών M από το αρχικό σύνολο χαρακτηριστικών N (M < N), έτσι ώστε ο χώρος χαρακτηριστικών να μειώνεται βέλτιστα σύμφωνα με ένα συγκεκριμένο κριτήριο.

Ρόλος:

1. μειώστε τη διάσταση του χώρου χαρακτηριστικών
2. επιταχύνετε έναν αλγόριθμο εκμάθησης
3. βελτίωση της προγνωστικής ακρίβειας ενός μοντέλου
4. βελτίωση της κατανόησης των αποτελεσμάτων

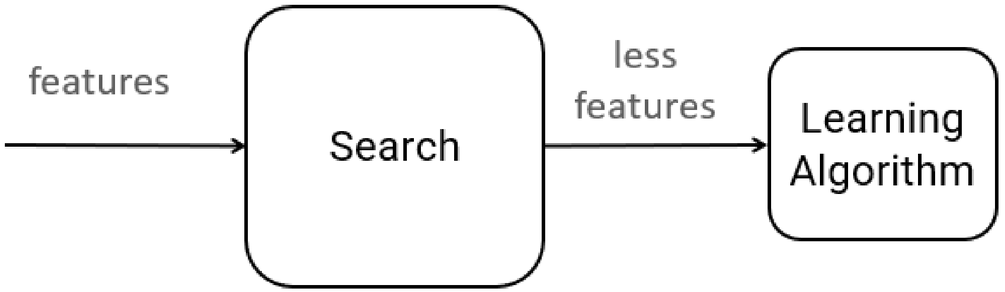
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

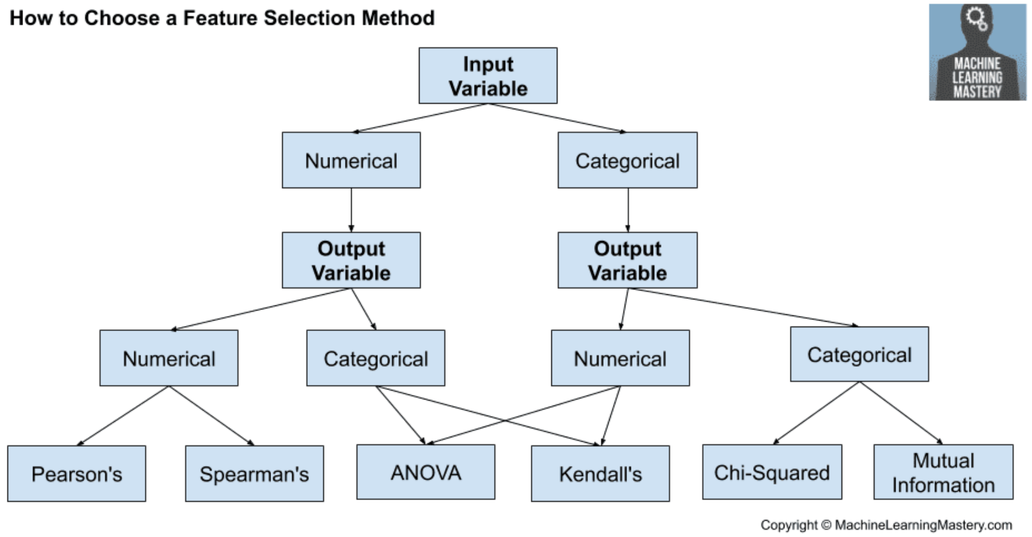
Επιλογή χαρακτηριστικών

Μέθοδοι φίλτρου

Επιλέξτε υποσύνολα των χαρακτηριστικών με βάση τη **σχέση τους με το στόχο**

Συνήθως με τη χρήση στατιστικών τεχνικών



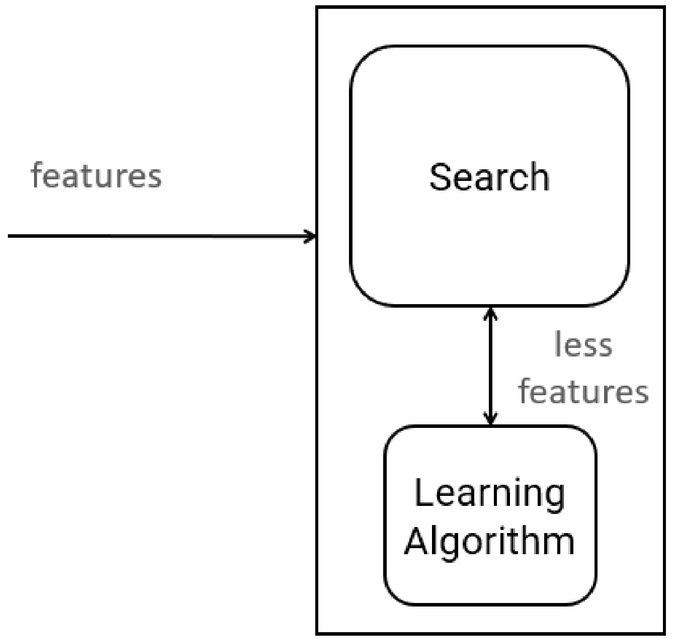


Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επιλογή χαρακτηριστικών

Μέθοδοι περιτυλίγματος



Αναζήτηση για **καλά αποδοτικά** υποσύνολα χαρακτηριστικών

Παραδείγματα:

- Αναδρομική Εξάλειψη Χαρακτηριστικών
- Ευρετικοί αλγόριθμοι αναζήτησης (π.χ. γενετικοί αλγόριθμοι, βελτιστοποίηση σμήνους σωματιδίων)

Υπολογιστικό κόστος;

Για κάθε υποσύνολο, ένα νέο μοντέλο εκπαιδεύεται και δοκιμάζεται για να επιτευχθεί η ακρίβεια

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επιλογή χαρακτηριστικών

Εγγενείς/έμμεσες/ενσωματωμένες μέθοδοι

Αυτόματη επιλογή χαρακτηριστικών κατά τη διάρκεια της εκπαίδευσης

Παραδείγματα:

- Μοντέλα που βασίζονται σε κανόνες
- Δενδρικά μοντέλα (θα μελετηθούν σε μεταγενέστερη διάλεξη)
- L1 τακτοποίηση (θα μελετήσει σε μια μεταγενέστερη διάλεξη)

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Εξαγωγή χαρακτηριστικών (μείωση διάστασης)

Διαδικασία που εξάγει ένα σύνολο νέων χαρακτηριστικών M από τα αρχικά χαρακτηριστικά N (M < N) μέσω κάποιας λειτουργικής χαρτογράφησης.

Στόχος: αναζητήστε ένα ελάχιστο σύνολο νέων χαρακτηριστικών μέσω κάποιου μετασχηματισμού σύμφωνα με κάποιο μέτρο απόδοσης.

Προσέγγιση:

- Ανάλυση κύριων συστατικών (και άλλες μέθοδοι μείωσης της διάστασης)
- Νευρωνικά δίκτυα
- Θα μελετηθούν σε μεταγενέστερες διαλέξεις

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Κατασκευή χαρακτηριστικών (μηχανική χαρακτηριστικών)

Διαδικασία που ανακαλύπτει ελλείπουσες πληροφορίες σχετικά με τις σχέσεις μεταξύ των χαρακτηριστικών και αυξάνει το χώρο των δυνατοτήτων με τη συναγωγή ή τη δημιουργία πρόσθετων χαρακτηριστικών.

Μπορεί να γίνει με τη χρήση:

- Αυτοματοποιημένες μέθοδοι - παραδείγματα:
  - Αριθμητικά χαρακτηριστικά: πολυωνυμική επέκταση
  - Ονομαστικά χαρακτηριστικά γνωρίσματα: σύνδεση, αποσύνδεση, άρνηση
- Γνώση τομέα (π.χ., SurfaceArea = Ύψος x Πλάτος)

Στόχος: αυξήστε την εκφραστική δύναμη των αρχικών χαρακτηριστικών

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα: Πολυωνυμικά χαρακτηριστικά

```
In [1]: import numpy as np
   ...: from sklearn.preprocessing import PolynomialFeatures
   ...: X = np.arange(6).reshape(3, 2)
   ...: print(X)
   ...:
   ...: poly = PolynomialFeatures(2)
   ...: poly.fit_transform(X)
[[0 1]
 [2 3]
 [4 5]]
Out[1]:
array([[ 1.,  0.,  1.,  0.,  0.,  1.],
       [ 1.,  2.,  3.,  4.,  6.,  9.],
       [ 1.,  4.,  5., 16., 20., 25.]])
```

Πολυώνυμο 2ου βαθμού:
Είσοδος: [α, β]
Έξοδος: [1, a, b, a^2, ab, b^2].

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Κουίζ

- Η επιλογή χαρακτηριστικών συνήθως μειώνει τον αριθμό των χαρακτηριστικών

ΣΩΣΤΟ

- Η εξαγωγή χαρακτηριστικών συνήθως αυξάνει τον αριθμό των χαρακτηριστικών

ΛΑΘΟΣ

- Χαρακτηριστικό κατασκευής αυξάνει συνήθως τον αριθμό των χαρακτηριστικών

ΛΑΘΟΣ

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Αύξηση δεδομένων (ή υπερδειγματοληψία δεδομένων)

Τεχνική που εισάγει πρόσθετα σημεία δεδομένων (περιστάσεις)

Χρησιμοποιείται όταν έχουμε:
- ένα **μικρό σύνολο δεδομένων**
- **σύνολο δεδομένων με ανισορροπία** (π.χ. 100 δείγματα κατηγορίας 0, 10 δείγματα της κατηγορίας 1)

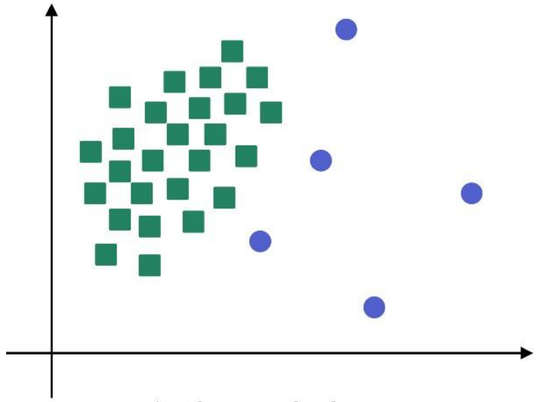
Προσεγγίσεις:
- απλή τυχαία υπερδειγματοληψία (επαναλαμβανόμενα δεδομένα)
- προσθέστε μικρό θόρυβο Gaussian (για αριθμητικά χαρακτηριστικά)
- Smote [1] και παραλλαγές (για αριθμητικά χαρακτηριστικά)

Δείτε imbalanced-learn library

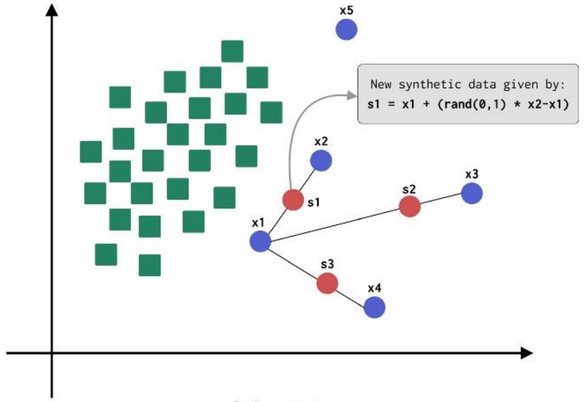
[1] Nitesh et al. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357, 2002.

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

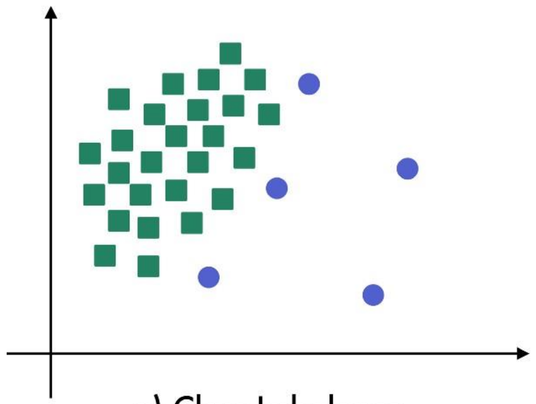
Αύξηση δεδομένων (ή υπερδειγματοληψία δεδομένων)



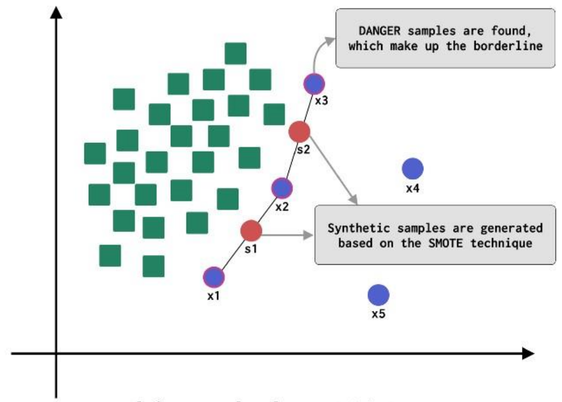
a) Class Imbalance



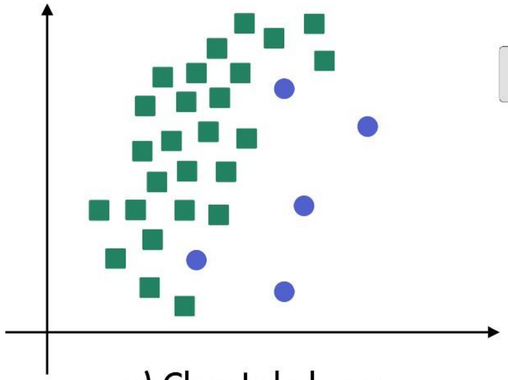
b) SMOTE



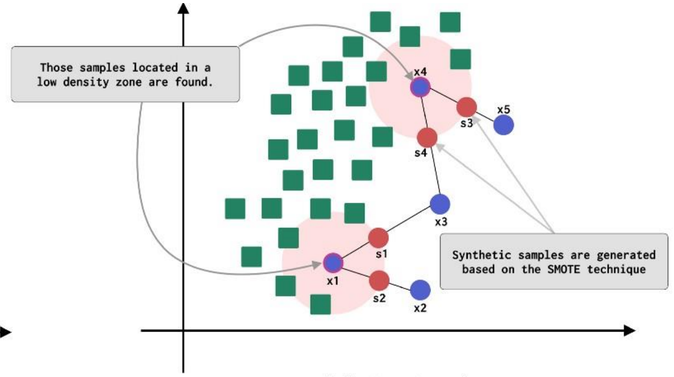
a) Class Imbalance



b) Borderline-SMOTE



a) Class Imbalance



b) ADASYN

Πηγή εικόνων

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Δειγματοληψία δεδομένων (ή μη δειγματοληψία δεδομένων)

Τεχνική που αφαιρεί τα σημεία δεδομένων (instances)

Χρησιμοποιείται όταν έχουμε:

- ένα **τεράστιο σύνολο δεδομένων** (που συνήθως δεν μπορεί να χωρέσει στη μνήμη)
- **σύνολο δεδομένων με ανισορροπία** (π.χ. 1000 δείγματα κατηγορίας 0, 100 δείγματα της κατηγορίας 1)

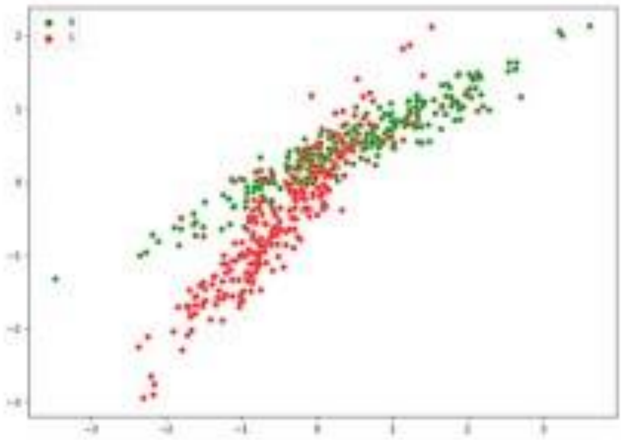
Προσεγγίσεις:

- απλή τυχαία δειγματοληψία
- δειγματοληψία συστάδων

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

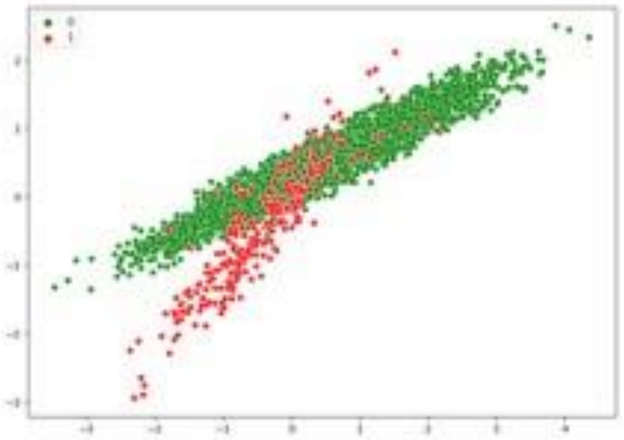
Υποδειγματικότητα και υπερδειγματοληψία δεδομένων

Under-sampling



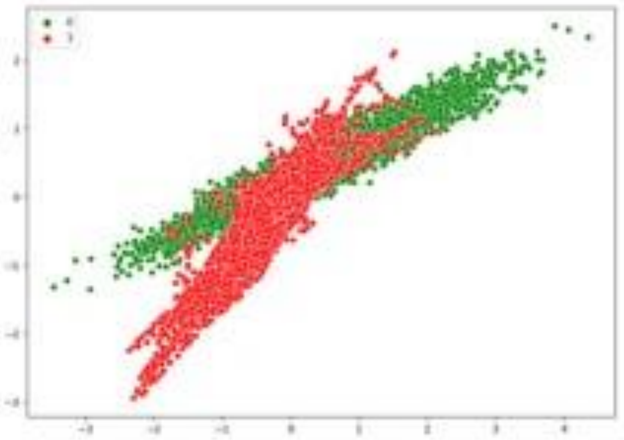
Samples of class 1 = 267
Samples of class 0 = 267

Over-sampling



Samples of class 1 = 4733
Samples of class 0 = 4733

SMOTE



Samples of class 1 = 4733
Samples of class 0 = 4733

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Διαχωρισμός συνόλου δεδομένων

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Διαχωρισμός συνόλου δεδομένων

70 %

Χρησιμοποιείται για την προσαρμογή του μοντέλου: βελτίωση των επιδόσεών του

30 %

Χρησιμοποιείται για τον έλεγχο των επιδόσεων γενίκευσης του μοντέλου

Γιατί το χρειαζόμαστε αυτό;

Το σύνολο δεδομένων είναι ένα δείγμα από την υποκείμενη κατανομή δεδομένων.

Αν χρησιμοποιήσουμε ολόκληρο το σύνολο, διατρέχουμε τον κίνδυνο «υπερπροσαρμογής» του μοντέλου στις αποχρώσεις του δείγματος.

Περισσότερα για αυτό σε μεταγενέστερες διαλέξεις...

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Συνήθες πρόβλημα: Διαρροή δεδομένων

Κατά την προετοιμασία των δεδομένων διαρρέεται η γνώση του συνόλου διακρατούμενων (δοκιμών)

Χαρακτηριστικό παράδειγμα: η κλιμάκωση χαρακτηριστικών εξετάζει ολόκληρο το σύνολο δεδομένων αντί μόνο για το σύνολο εκπαίδευσης

Μπορεί να οδηγήσει σε λανθασμένες επιδόσεις του μοντέλου σε νέα δεδομένα

Πώς να το αποφύγετε: τοποθετήστε τη διαδικασία μετατροπής δεδομένων στο σύνολο εκπαίδευσης και αξιολογήστε μόνο το σύνολο δοκιμών

Το σύνολο δοκιμών δεν χρησιμοποιείται μόνο για την αξιολόγηση του εκπαιδευμένου μοντέλου, αλλά και για τη διαδικασία μετασχηματισμού δεδομένων.

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επόμενη Διάλεξη

Μέρος 2: Εποπτευόμενη Μάθηση

- Παλινδρόμηση
- Ταξινόμηση

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Σας ευχαριστούμε

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423