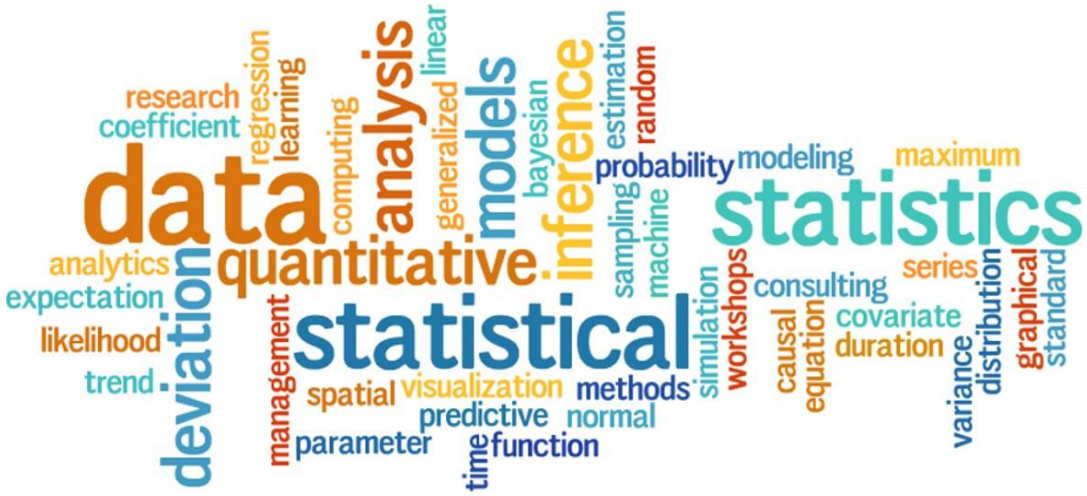
Ατομικό Project μαθήματος

«Ανάλυση Δεδομένων και Στατιστική»

Χειμερινό εξάμηνο 2023-2024

Γεράσιμος Ραζής - Ιωάννης Αναγνωστόπουλος

Ανάλυση συνόλων δεδομένων του αποθετηρίου Kaggle μέσω λογισμικού WEKA



Περιγραφή

Το αποθετήριο *Kaggle*[1], πέραν της δυνατότητας απόθεσης και τεκμηρίωσης των δεδομένων των χρηστών του, αποτελεί μία εξαιρετική υπηρεσία για ανάπτυξη, επαναχρησιμοποίηση, διόρθωση, αλλά και επαλήθευση ερευνητικών δεδομένων και προβλημάτων.

Αφού επιλέξετε από το αποθετήριο Kaggle δύο σύνολα δεδομένων τύπου CSV από δύο διαφορετικές θεματικές περιοχές των ενδιαφερόντων σας (π.χ. ιατρική, βιολογία, αξιολογήσεις προϊόντων, οικονομία, ασφάλεια, κ.α.), να απαντήσετε τα παρακάτω ερωτήματα μέσω της ανάλυσή τους με χρήση του λογισμικού WEKA (ένα σύνολο δεδομένων για κατηγοριοποίηση και ένα για συσταδοποίηση).

[1] https://en.wikipedia.org/wiki/Kaggle

<u>Α. Κατηγοριοποίηση (Α' σύνολο δεδομένων)</u>

1. Περιγράψτε εν συντομία πώς δημιουργήσατε το αρχείο arff για το επιλεγμένο σύνολο δεδομένων.

2. Διαβάστε το αρχείο και μελετήστε τα ανεπεξέργαστα δεδομένα (μπορείτε να δείτε τα ανεπεξέργαστα δεδομένα πιο εύκολα μέσω του κουμπιού «Edit…» ενώ βρίσκεστε στη λειτουργία προεπεξεργασίας (preprocess) του WEKA). Η κατανόηση του τομέα είναι ένα βασικό βήμα στην εξόρυξη και ανάλυση δεδομένων. Καταγράψτε τρία θέματα ή υποθέσεις που ιδανικά θα θέλατε να διερευνήσετε.

   a. Για παράδειγμα, εσείς θα μπορούσε να αλλάξετε τη μεταβλητή μιας κλάσης και να προσπαθήσετε να προβλέψετε αυτήν τη μεταβλητή από τις υπόλοιπες μεταβλητές.

   b. Μπορείτε επίσης να εξετάσετε άλλα πράγματα, όπως στρατηγικές για προεπεξεργασία των δεδομένων για πιθανή βελτίωση της προγνωστικής απόδοσης. Αλλά σε τέτοια περίπτωση θα πρέπει να δηλώσετε ότι θα εξετάσετε εάν μια τέτοια στρατηγική λειτουργεί ή όχι.

Εν ολίγοις, καταλήξτε σε τρία ζητήματα/στρατηγικές που θα μπορούσατε να βελτιώσετε/εφαρμόσετε.

3. Αφού διαβάσετε τα δεδομένα μέσω του αρχείου arff, ενώ βρίσκεστε στην καρτέλα προεπεξεργασίας, κοιτάξτε στο πώς κάθε χαρακτηριστικό συσχετίζεται με τη μεταβλητή κλάσης (μπορείτε να το κάνετε αυτό κάνοντας κλικ στο κουμπί «Οπτικοποίηση όλων» (Visualize All). Καταγράψτε 3 χαρακτηριστικά που πιστεύετε ότι θα είναι χρήσιμα για την πρόβλεψη της μεταβλητής κλάσης, και με μία ή δύο προτάσεις, αιτιολογήστε το γιατί πιστεύετε ότι αυτά τα χαρακτηριστικά θα είναι χρήσιμα.

4. Εκτελέστε τον ταξινομητή Weka J48 στα αρχικά δεδομένα με την επιλογή ελέγχου (test option) ρυθμισμένη στο 66%, έτσι ώστε το 66% των δεδομένων να χρησιμοποιείται για εκπαίδευση και το υπόλοιπο για έλεγχο. Απαντήστε τα παρακάτω ερωτήματα:

   a. Ποια είναι η ακρίβεια του ταξινομητή στα δεδομένα ελέγχου;

   b. Πόσα φύλλα υπάρχουν στο δέντρο;

   c. Πόσος χρόνος χρειάστηκε για την κατασκευή του μοντέλου;

   d. Αντιγράψτε και επικολλήστε τον πίνακα σύγχυσης (confusion matrix).

5. Επαναλάβετε το προηγούμενο βήμα αλλά αυτή τη φορά διαμορφώστε τις επιλογές/ιδιότητες για τον ταξινομητή J48 ώστε το binarySplits να είναι αληθές (δηλαδή, όλα τα splits να είναι δυαδικά). Τότε απάντησε τις ίδιες τέσσερις ερωτήσεις (a - d) όπως στο προηγούμενο ερώτημα.

6. Επαναλάβετε το προηγούμενο βήμα 5, αλλά τώρα αλλάξτε και την επιλογή «unpruned» από «False» σε «True», ώστε να μην εκτελείται το κλάδεμα. Τότε απάντησε τις ίδιες τέσσερις ερωτήσεις (a - d) όπως στο προηγούμενο ερώτημα.

7. Συνοψίστε τις κύριες διαφορές μεταξύ των 3 εκτελέσεων του J48 όσον αφορά τις αλλαγές στην ακρίβεια (accuracy), τον αριθμό των φύλλων στο δέντρο απόφασης και τον χρόνο που χρειάζεται για να χτιστεί το μοντέλο.

8. Στο αρχικό σύνολο δεδομένων, εκτελέστε τον ταξινομητή ZeroR, τον οποίο μπορείτε να βρείτε στους ταξινομητές κανόνων (rules classifiers) στο WEKA. Απαντήστε τις ερωτήσεις a,

c, και d από την Ερώτηση 4 για αυτόν τον ταξινομητή και να χαρακτηρίσετε τις διαφορές μεταξύ αυτών των αποτελεσμάτων και εκείνων για την Ερώτηση 4. Τι πιστεύετε ότι κάνει αυτή η μέθοδος; Γιατί δεν μπορεί να οπτικοποιηθεί αυτό το δέντρο απόφασης;

9. Στο αρχικό σύνολο δεδομένων, εκτελέστε τη μέθοδο RandomForest τον οποίο μπορείτε να βρείτε στους ταξινομητές δέντρων (tree classifiers) στο WEKA. Αναφέρετε τις απαντήσεις στις ερωτήσεις a, c, και d από την Ερώτηση 4. Στη συνέχεια, απαντήστε στα δύο επόμενα:

    a. Πώς συγκρίνονται τα αποτελέσματα ακρίβειας αυτής της μεθόδου με αυτά του J48 με τις προεπιλεγμένες του παραμέτρους από την Ερώτηση 4;

    b. Περιγράψτε με λίγες προτάσεις πώς λειτουργεί η μέθοδος του τυχαίου δάσους (μπορείτε να το ψάξετε σε διάφορες πηγές).

10. Στο αρχικό σύνολο δεδομένων, χρησιμοποιήστε το J48 για να δημιουργήσετε έναν ταξινομητή για να προβλέψετε μία άλλη κλάση από την προεπιλεγμένη κατηγορία. Μπορείτε να κάνετε αυτό χρησιμοποιώντας την καρτέλα «Ταξινόμηση» (Classify) για να επιλέξετε τον J48, ως συνήθως, και στη συνέχεια, μπορείτε να μεταβείτε στην αριστερή στήλη στο κουμπί κάτω από το κουμπί «Περισσότερες επιλογές…» (More Option) και χειροκίνητα να αλλάξτε τη μεταβλητή class/target. Όπως πριν, χρησιμοποιήστε το 66% για δεδομένα εκπαίδευσης. Απαντήστε στις ερωτήσεις a - d της Ερώτησης 4. Στη συνέχεια απαντήστε στις ακόλουθες ερωτήσεις:

    a. Με βάση τα αποτελέσματα στον πίνακα σύγχυσης, καθορίστε τον αριθμό των στιγμιοτύπων της κλάσης τόσο ως μετρήσεις (ακέραιοι αριθμοί), όσο και ως ποσοστά.

    b. Αν έπρεπε να φτιάξετε έναν απλό ταξινομητή για την πρόβλεψη του πιο συνηθισμένου τύπου κλάσης, ποια θα ήταν η ακρίβεια του ταξινομητή;

* Σε όλα τα βήματα, μπορείτε να προσθέσετε εικόνες και γραφήματα όπου πιστεύετε πως αξίζει.

<u>Β. Συσταδοποίηση (Β' σύνολο δεδομένων)</u>

1. Περιγράψτε εν συντομία πώς δημιουργήσατε το αρχείο arff για το επιλεγμένο σύνολο δεδομένων.

2. Πόσα στιγμιότυπα υπάρχουν, και πόσα χαρακτηριστικά; Περιγράψτε τον τύπο των χαρακτηριστικών (π.χ. αριθμητικού, συνεχές, με διακριτές τιμές).

3. Από την καρτέλα «Cluster» εφαρμόσετε τρεις αλγορίθμους συσταδοποίησης (ο ένας να είναι ο SimpleKMeans και ένας HierarchicalClusterer) στα φορτωμένα δεδομένα. Για κάθε έναν από τους αλγορίθμους αναφέρετε το πλήθος συστάδων, καθώς και τα στατιστικά στοιχεία που δίνουν τη μέση και τυπική απόκλιση για καθένα από τα χαρακτηριστικά.

4. Για κάθε έναν από τους αλγορίθμους επιλέξτε «Οπτικοποίηση συστάδων» (Visualize cluster assignments) και αναλύστε το αποτέλεσμα σε δύο-τρεις προτάσεις.

5. Για την περίπτωση του SimpleKMeans (ή όπου αλλού είναι εφικτό) αλλάξτε τον αριθμό των συστάδων σε 3, 5, και 10 και αναλύστε το αποτέλεσμα.

    a. Ποια συσταδοποίηση είναι καλύτερη;

    b. Τι είναι το «άθροισμα τετράγωνων σφαλμάτων εντός συστάδας»;

* Σε όλα τα βήματα, μπορείτε να προσθέσετε εικόνες και γραφήματα όπου πιστεύετε πως αξίζει.

Παραδοτέα

1. Ένα έγγραφο κειμένου όπου θα περιγράφονται:
    a. οι θεματικές περιοχές και τα datesets που επιλέξατε.
    b. οι απαντήσεις σας στα ανωτέρα σημεία, με τις απαραίτητες εικόνες.
2. Τα αρχεία arff.

Χρήσιμες πηγές

- How to use Kaggle (Documentation)
- Getting started on Kaggle (συλλογή videos στο YouTube)
- Σημειώσεις και διαφάνειες του μαθήματος (διαθέσιμα στο GUNET)
- Οδηγός WEKA (online κεφάλαιο βιβλίου στον Κάλλιπο)

Ημερομηνία παράδοσης: 14/04/2024 (αποκλειστικά μέσω GUNET)

Το παραδοτέο θα είναι ένα συμπιεσμένο αρχείο με την εξής ονομασία:
{Project-SURNAME.zip} ή {Project-SURNAME.rar}