ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ
ΓΡΑΠΤΗ ΕΡΓΑΣΙΑ 6
Support Vector Machines

Σκοπός της συγκεκριμένης εργασίας είναι η εξοικείωση με την υλοποίηση του αλγορίθμου Support Vector Machines (SVM), μέσω της αντίστοιχης μεθόδου που παρέχει η βιβλιοθήκη sci-kit learn της Python.

Δημιουργήθηκαν διάφορα SVM, δίνοντας διάφορες τιμές στις χαρακτηριστικές παραμέτρους που επηρεάζουν την απόδοσή τους. Η εφαρμογή τους έγινε με χρήση των δεδομένων που υπάρχουν στο παρεχόμενο αρχείο 'creditcard.csv'. Τα δεδομένα, πριν εισαχθούν στα N.N. υφίστανται προεπεξεργασία (preprrocessing), μέσω της κλάσης MinMaxScaler() της βιβλιοθήκης sklearn, η οποία τα κανονικοποιεί μετατρέποντάς τα στην κλίμακα (0,1), στοιχείο κρίσιμο για την ορθή λειτουργία του αλγορίθμου. Η αξιολόγηση και η σύγκριση της απόδοσης των διαφορετικά παραμετροποιημένων SVM έγινε, σύμφωνα με τις οδηγίες της εκφώνησης, με χρήση των ακόλουθων μετρικών (στον υπολογισμό των μετρικών επιλέχθηκε τιμή παραμέτρου average='macro'):

- **precision**: Ορίζεται ως TP / (TP+FP)
- **recall**: Ορίζεται ως TP / (TP+FN)
- **f1**: Ορίζεται ως 2*precision*recall / (precision+recall)

Α. Περιγραφή και μελέτη του αλγορίθμου

Σύμφωνα με τις απαιτήσεις της εκφώνησης δημιουργήθηκαν διάφορα SVM διαφοροποιούμενα στις τιμές των παρακάτω παραμέτρων:

- **C**: Κρίσιμη παράμετρος που καθορίζει την αποφασισιμότητα του αλγορίθμου. Μικρές τιμές οδηγούν στη δημιουργία πιο ομαλού υπερεπιπέδου με μεγαλύτερα περιθώρια, επομένως με μεγαλύτερα ποσοστά λαθεμένων κατατάξεων. Αντίθετα μεγαλύτερες τιμές αυξάνουν τα επίπεδα επιτυχίας, αυξάνοντας όμως πολύ και τον υπολογιστικό χρόνο.
- **kernel**: καθορίζει τον τρόπο που ο αλγόριθμος μετατρέπει την αναπαράσταση των δεδομένων από το υπερεπίπεδο σε υψηλότερες διαστάσεις, για να μπορέσει να τα διαχωρίσει.
- **degree**: Αντιστοιχεί στο βαθμό του χρησιμοποιούμενο πολυωνυμίου όταν επιλεγεί ο πολυωνυμικός kernel.
- **gamma**: Χρησιμοποιείται ως συντελεστής στην kernel function (εκτός του γραμμικού).

Για τα διάφορα SVM επιλέχθηκαν τιμές παραμέτρων, όπως αυτές ορίστηκαν στο συνοδευτικό αρχείο *SVM_Results.xlsx*, που μας παραδόθηκε, καθώς και ορισμένων δικών μου επιλογών για να διευκρινίσω την επίδραση των παραμέτρων. Τα διάφορα παραμετροποιημένα μοντέλα εκτελέστηκαν διαδοχικά, με παράλληλη καταγραφή των τιμών των μετρικών. Ο κώδικας υλοποίησης και εκτέλεσης του αλγορίθμου υπάρχει στο συνοδευτικό αρχείο *Iliapikos_ge6_SVM.py*. Η μεταβολή των τιμών των ζητούμενων μετρικών φαίνεται στο συνοδευτικό αρχείο *SVM_Results.xlsx*.

Β. Σύγκριση των παραλλαγών του αλγορίθμου

Βλέποντας συγκριτικά τα αποτελέσματα, παρατηρώ τα παρακάτω αναφορικά με την επίδραση των διαφόρων παραμέτρων στην επίδοση του αλγορίθμου:

- **C**: Όπως ήταν λογικό, αύξηση της τιμής της παραμέτρου αυξάνει και την απόδοση του αλγορίθμου, εκτός από την περίπτωση χρήσης sigmoid kernel, όπου δεν παρατηρείται καμία απολύτως επίδραση.
- **gamma**: Ελάχιστη ή και καθόλου επίδραση στην απόδοση του αλγορίθμου, όταν χρησιμοποιείται kernel που τη χρησιμοποιεί. Η βελτίωση που παρατηρείται με kernel rbf, οφείλεται αποκλειστικά στην μεταβολή της τιμής της C, όπως γίνεται φανερό με τη χρήση ενός επιπλέον SVM με τιμές παραμέτρων C=10 και γαμμα=0.3.
- **degree**: Χρησιμοποιείται μόνο με πολυωνυμικό kernel. Για να διαχωρίσω την επίδρασή του από αυτή της C, χρησιμοποίησα επιπλέον δείγματα με τιμή C=10 και degree=(2, 3, 4). Από τα αποτελέσματα φαίνεται ξεκάθαρα ότι η αύξηση στη τιμή της degree μειώνει δραστικά την απόδοση του αλγορίθμου.

Σημαντική παρατήρηση: Σε ορισμένους συνδυασμούς τιμών παραμέτρων, ειδικά όταν C=0.1, παρατήρησα ότι εμφανίζονταν μηνύματα που με πληροφορούσαν ότι κάποια κατηγορία δεν εμφανίζεται καθόλου στις προβλέψεις του συνόλου ελέγχου, προκαλώντας πρόβλημα στο υπολογισμό κάποιων μετρικών. Η κλάση αυτή ήταν η 1.

Πραγματοποιώντας ποιοτικό έλεγχο στα δεδομένα εισόδου είδα ότι αυτά ήταν εντελώς ανισοβαρή αναφορικά με την κατανομή των δειγμάτων στις 2 κατηγορίες. Συγκεκριμένα το 99.8% των δεδομένων (n=284.315) ανήκαν στην κατηγορία 0 και μόλις το 0.2% (n=492) στην κατηγορία 1. Οπότε είναι φυσικό κατά την εκπαίδευση, υπό συγκεκριμένες συνθήκες τιμών παραμέτρων, η επικρατούσα αριθμητικά κατηγορία να υποσκελίζει εντελώς την άλλη. Σύμφωνα με τις υποδείξεις στην τεκμηρίωση του αλγορίθμου, σε αυτές τις περιπτώσεις προτείνεται η χρήση της παραμέτρου **class_weight**='balanced', η οποία καθορίζει για κάθε κατηγορία ένα βάρος αντιστρόφως ανάλογο της πληθικότητάς του. Αυτό όμως αύξησε δραματικά τον υπολογιστικό χρόνο που απαιτούταν, οπότε δεν μπόρεσα να ολοκληρώσω την εκτέλεση και να παρουσιάσω κάποια αποτελέσματα.

Συμπερασματικά, τις καλύτερες αποδόσεις τις παρουσιάζει ο αλγόριθμος όταν χρησιμοποιείται kernel polynomial ή rbf σε συνδυασμό με μια σχετικά μεγάλη τιμή της C=10 (παρόμοιες και στις δύο περιπτώσεις). Αντίθετα η απόδοση του αλγορίθμου στο συγκεκριμένο dataset και όταν χρησιμοποιείται kernel sigmoid, είναι σταθερά κακή και φαίνεται ότι δεν επηρεάζεται από τις τιμές των παραμέτρων. Η ίδια εικόνα διατηρήθηκε ακόμα και όταν συμπεριέλαβα στη μελέτη και την παράμετρο **coef0**, που επιδρά άμεσα μαζί με την gamma στην sigmod kernel function (τα δεδομένα δεν δίνονται).

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Υλικό μαθήματος.

2. Müller, A.C & Guido, S. - Introduction to Machine Learning with Python. A Guide for Data Scientists, O' Reilly, 2017.

3. Τεκμηρίωση από τον ιστότοπο της βιβλιοθήκης Sklearn.