Αναγνώριση Προτύπων-Μηχανική Μάθηση

ΠΡΟΑΙΡΕΤΙΚΉ ΕΡΓΑΣΊΑ 2019

ΌΝΟΜΑ: ΑΝΑΣΤΑΣΙΟΣ ΑΝΤΩΝΟΠΟΥΛΟΣ
Α.Μ.: 1115201400014

Εισαγωγή

Στην εργασία αυτή υλοποιήθηκαν βασικοί αλγόριθμοι μηχανικής μάθησης. Στην συνέχεια εφαρμόστηκαν πάνω στο σύνολο δεδομένων Wisconsin Breast Cancer από τον ιστότοπο του UCI Repository. Τέλος αξιολογήθηκαν με βάση τα κριτήρια της Ακρίβειας(Accuracy), της Ευαισθησίας (Sensitivity) και της Ειδικότητας (Specificity).

Η εργασία υλοποιήθηκε στο περιβάλλον του MATLAB R2018.

Περιγραφή

Στα αρχεία της εργασίας περιλαμβάνονται τα εξής αρχεία:

- wdbc.data
- read_data.m
- normalize_data.m
- return_targets.m
- K_Nearest_Neighbor.m
- Naive_Bayes.m
- SupportVectorMachine.m
- DecisionTree.m

Το αρχείο wdbc.data περιλαμβάνει τα δεδομένα του προβλήματος όπως αυτά υπάρχουν στον ιστότοπο του UCI Repository. Αρχικά εκτελούμε το αρχείο read_data.m το οποίο διαβάζει τα δεδομένα του προβλήματος χρησιμοποιώντας τις συναρτήσεις που υπάρχουν στα αρχεία normalize_data.m και return_targets.m. Μετά την εκτέλεση έχουμε δημιουργήσει ένα αρχείο με ονομασία datasets.mat το οποίο περιλαμβάνει ένα πίνακα x με τα χαρακτηριστικά του προβλήματος κανονικοποιημένα στο εύρος [0,1] και ένα πίνακα t που περιέχει τα Targets για κάθε περιστατικό(1 -> κακοήθης , 0-> καλοήθης). Το αρχείο αυτό θα χρησιμοποιηθεί στην συνέχεια από τα υπόλοιπα για την εκτέλεση των ταξινομητών. Για την εκτέλεση κάθε ταξινομητή επιλέγουμε το αρχείο με το αντίστοιχο όνομα και το εκτελούμε. Σε κάθε αρχείο εκτελείται ο ίδιος αλγόριθμος ταξινόμησης σε δύο διαφορετικές εκδοχές, δηλαδή με διαφορετικές παραμέτρους, και εκτυπώνονται στο Command Window του MATLAB τα αποτελέσματα για κάθε περίπτωση.

Τέλος όλοι οι αλγόριθμοι εκτελούνται με cross-validation χρησιμοποιώντας την συνάρτηση crossval(), με K-Fold = 10. Η απόδοση μετρήθηκε χρησιμοποιώντας την συνάρτηση classperf().

Παρακάτω ακολουθούν αναλυτικοί πίνακες με τα αποτελέσματα και τις παραμέτρους για κάθε αλγόριθμο ταξινόμησης.

Ταξινομητής Κ-Κοντινότερων Γειτόνων (K-Nearest Neighbor)

Model 1 parameters:

- BreakTies -> smallest
- Distance -> Euclidean
- DistanceWeight -> Equal
- NumNeighbors -> 10

Model 2 parameters:

- BreakTies -> random
- Distance -> seuclidean
- DistanceWeight -> inverse
- NumNeighbors -> 50

Ενδεικτικά Αποτελέσματα

	Accuracy	Sensitivity	Specificity
Model 1	92.44	86.79	95.80
Model 2	95.43	88.68	99.44

Ταξινομητής Bayes

Model 1 parameters:

- DistributionNames -> kernel
- Kernel -> normal
- Support -> unbounded

Model 2 parameters:

- DistributionNames -> kernel
- Kernel -> triangle
- Support -> unbounded

Ενδεικτικά Αποτελέσματα

	Accuracy	Sensitivity	Specificity
Model 1	94.20	92.45	95.24
Model 2	92.44	87.74	94.96

Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

Model 1 parameters:

- KernelFunction -> rbf
- KernelScale -> 1

Model 2 parameters:

- KernelFunction -> gaussian
- KernelScale -> auto

Ενδεικτικά Αποτελέσματα

	Accuracy	Sensitivity	Specificity
Model 1	89.46	73.58	98.88
Model 2	92.44	85.85	96.36

Δένδρα Απόφασης (Decision Tree)

Model 1 parameters:

- AlgorithmForCategorical -> exact
- MergeLeaves -> on
- MinParentSize -> 10

Model 2 parameters:

- AlgorithmForCategorical -> PullLeft
- MergeLeaves -> off
- MinParentSize -> 40

Ενδεικτικά Αποτελέσματα

	Accuracy	Sensitivity	Specificity
Model 1	92.44	90.09	93.84
Model 2	90.51	84.43	94.12

Συμπέρασμα

Όπως γνωρίζουμε η αξιολόγηση με βάση μόνο την Ακρίβεια δεν οδηγεί πάντα σε σωστά συμπεράσματα. Για να είμαστε πιο σίγουροι για το ποιος αλγόριθμος ταξινόμησης είναι καλύτερος για το πρόβλημα μας χρησιμοποιούμε και άλλους δείκτες μέτρησης της απόδοσης όπως είναι η Ευαισθησία και η Ειδικότητα.

Από τους ορισμούς τους η Ευαισθησία είναι το ποσοστό που μας δείχνει ποιοι από αυτούς που βγήκαν θετικοί στο αποτέλεσμα έχουν όντως την ασθένεια και η Ειδικότητα είναι το ποσοστό που μας δείχνει ποιοι από αυτούς που βγήκαν αρνητικοί στο αποτέλεσμα όντως δεν νοσούν.

Άρα θεωρώ ότι για το συγκεκριμένο πρόβλημα ο δείκτης της Ειδικότητας είναι πιο σημαντικός. Έτσι σε συνδυασμό και με την Ακρίβεια ο αλγόριθμος ταξινόμησης που θα επέλεγα είναι ο Ταξινομητής Κ-Κοντινότερων Γειτόνων με παραμέτρους BreakTies -> random, Distance -> seuclidean, DistanceWeight -> inverse, NumNeighbors -> 50. Αφού στο συγκεκριμένο παράδειγμα ο ταξινομητής αυτός παρουσιάζει μεγάλα ποσοστά Ακρίβειας αλλά και μεγάλα ποσοστά Ειδικότητας.