Εργασία 6 - SVM - Bayes

Εργασία 6a – Support Vector Machines (SVM)

Ύστερα από πρόσφατες φυσικές καταστροφές, η κυβέρνηση αποφάσισε να βελτιώσει το σύστημα πρόληψης και εκκένωσης με τη βοήθεια αναρτήσεων από το X (Twitter). Πιο συγκεκριμένα, το σύστημα θα εξετάζει αναρτήσεις και σε περίπτωση που κρίνει ότι κάποια φυσική καταστροφή (target=1) βρίσκεται σε εξέλιξη, θα αποστέλλει κατάλληλο μήνυμα μέσω του 112.

Ερωτήματα:

1. Κατεβάστε τα σύνολα δεδομένων εκπαίδευσης/δοκιμής train.csv και χωρίστε το dataset σε train-test (90-10%) με την επιλογή stratify και random_state=0. https://www.kaggle.com/competitions/nlp-getting-started. Στη συνέχεια, δημιουργήστε bar-plot με το πλήθος των target στα train, test αντίστοιχα. Είναι ισορροπημένο το dataset?
2. Επεξεργαστείτε τα δεδομένα train/ test ως εξής:
    a. Αφαιρέστε τα χαρακτηριστικά id, location.
    b. Συμπληρώστε τις ελλιπής τιμές με το keyword “null”.
    c. Ενώστε τις στήλες keyword και text ως: df[‘inputs’] = df[‘keyword’] + ‘ ‘ + df[‘text’].
3. Επεξεργαστείτε τα κείμενα των tweets (inputs) στα train, test, με σκοπό να αφαιρέσετε όλες τις περιττές πληροφορίες (π.χ. σύμβολα, emoji, κλπ). Χρησιμοποιείστε τις έτοιμες υλοποιημένες συναρτήσεις που θα βρείτε στο παρακάτω repository: https://github.com/Deffro/text-preprocessing-techniques/tree/master?tab=readme-ov-file#0-remove-unicode-strings-and-noise. Να περιγράψετε περιληπτικά (σε 1 γραμμή) την κάθε τεχνική που χρησιμοποιήσατε και να αιτιολογήσετε το λόγο που επιλέξατε τις συγκεκριμένες τεχνικές.
4. Χωρίστε τα δεδομένα σε inputs, targets (numpy arrays).
5. Να μετατρέψετε τα inputs ως εξής: Για κάθε ανάρτηση (post), υπολογίστε το tf-idf κάθε λέξης. Χρησιμοποιήστε την κλάση TfIDFVectorizer της scikit-learn: https://scikit-learn.org/1.5/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html. Θα χρειαστεί να εφαρμόσετε fit_transform για το train set και σκέτο transform για το test set.
6. Εκπαιδεύστε ταξινομητή SVM με linear Kernel (LinearSVC) https://scikit-learn.org/dev/modules/generated/sklearn.svm.LinearSVC.html στο train και μετρήστε για τα σύνολα train και test: Accuracy, F1, Precision, Recall.
   Ποια μετρική είναι πιο σημαντική? Αιτιολογείστε.
7. Να εφαρμόσετε μετασχηματισμό PCA στα δεδομένα, κρατώντας το 95% της χρήσιμης πληροφορίας (προσοχή όχι των συνιστωσών!). Στη συνέχεια, να εκπαιδεύσετε τους LinearSVC και SVC με RBF kernel https://scikit-learn.org/dev/modules/generated/sklearn.svm.SVC.html και συγκρίνετε τις αποδόσεις τους.
8. Ποια είναι η τιμή της παραμέτρου γ (Gamma) που χρησιμοποιείται ως default ‘scale’ και ποια η τιμή ‘auto’? Να αλλάξετε την παράμετρο γ σε ‘auto’ και να επανεκπαιδεύστε τον SVC.

Τέλος, να δημιουργήσετε bar-plot, συγκρίνοντας τις μετρικές accuracy, F1, precision, recall των τεσσάρων μοντέλων που εκπαιδεύτηκαν.

Εργασία 6b – Bayes

Να επαναλάβετε την ερώτηση 6 (πριν εφαρμοστεί ο μετασχηματισμός PCA) με τη χρήσει Gaussian Naive Bayes:
https://scikit-learn.org/dev/modules/generated/sklearn.naive_bayes.GaussianNB.html