

Lecture 6: Bayesian Learning

Project 5

Να φορτώσετε το dataset **fetch_20newsgroups** και να εφαρμόσετε τη μέθοδο **Naive Bayes** (συνιστάται η χρήση της βιβλιοθήκης **scikit-learn** για Python) για την ταξινόμηση κειμένου σε κατηγορίες.

Να χρησιμοποιηθεί ο αλγόριθμος Naive Bayes με την υπόθεση ότι τα δεδομένα ακολουθούν πολωνυμική κατανομή (και συνεπώς να δοκιμαστούν οι διάφορες τιμές της παραμέτρου α). Να υπολογιστούν οι τιμές των μετρικών **Accuracy**, **Recall**, **Precision** και **F1**, και να δημιουργηθεί το γράφημα θερμότητας (heatmap) του πίνακα σύγχυσης, όπως φαίνεται στο παρακάτω παράδειγμα. Τα αποτελέσματα των μετρικών να φαίνονται στον τίτλο του γραφήματος.

Multinomial NB - Confusion matrix ($\alpha = 0.10$) [Prec = 0.78047, Rec = 0.72416, F1 = 0.72268]

alt.atheism	52	0	0	0	0	0	0	1	1	0	1	2	0	6	0	43	0	6	0	0
comp.graphics	0	99	8	8	3	8	2	0	0	0	0	4	1	7	3	4	0	0	0	0
ip.os.ms-windows.misc	0	5	99	15	1	6	2	0	0	0	0	3	3	5	0	1	0	0	0	0
ip.sys.ibm.pc.hardware	0	6	7	110	6	2	5	0	0	0	0	2	7	2	1	0	0	0	0	0
omp.sys.mac.hardware	0	3	7	14	102	1	4	0	1	0	0	1	6	8	0	1	0	0	1	0
comp.windows.x	0	7	7	2	0	135	1	0	0	0	1	0	1	2	1	2	0	0	0	0
misc.forsale	0	0	0	8	3	0	96	2	0	0	1	4	4	5	1	3	4	0	0	0
rec.autos	0	1	1	2	0	3	4	113	6	0	0	1	3	7	2	7	8	0	0	0
rec.motorcycles	1	0	0	0	0	3	2	8	121	4	1	1	0	3	2	8	7	0	1	0
rec.sport.baseball	0	1	1	0	0	1	4	0	1	126	5	2	0	5	0	1	1	0	0	0
rec.sport.hockey	0	0	0	0	0	0	0	1	2	4	135	0	0	4	0	2	1	0	1	0
sci.crypt	0	1	1	0	0	2	0	0	0	0	1	134	3	4	0	1	5	1	2	0
sci.electronics	0	2	3	10	5	0	3	4	3	0	1	7	99	3	4	1	2	0	0	0
sci.med	1	1	0	0	0	1	0	0	0	0	0	1	0	116	2	7	1	0	1	0
sci.space	0	2	1	0	0	2	0	4	1	0	2	4	1	8	124	3	1	0	1	0
soc.religion.christian	1	0	0	1	1	1	0	0	0	0	0	0	1	4	1	142	0	3	0	0
talk.politics.guns	0	1	1	0	0	0	0	2	1	0	1	5	1	10	2	5	112	1	2	0
talk.politics.mideast	2	1	0	0	0	0	0	0	0	0	2	1	0	4	1	12	2	118	1	0
talk.politics.misc	3	0	0	0	0	0	1	0	0	1	0	6	0	2	2	17	7	5	64	0
talk.religion.misc	12	0	0	0	0	0	0	0	0	0	0	0	0	8	2	43	9	5	1	7
atheism		graphics	ows.misc	hardware	hardware	/indows.x	sc.forsale	rec.autos	itorcycles	.:baseball	rt:hockey	sci.crypt	lectronics	sci.med	sci.space	.:christian	itics:guns	s.mideast	itics.misc	gion.misc

Σημείωση 1: Ένας σημαντικός (αλλά προαιρετικός) σκοπός είναι να επιτευχθεί τιμή για την **F1** κάτω από **70%** (ή περίπου εκεί)!

Σημείωση 2: Κατά την υποβολή της εργασίας, είναι απαραίτητο να συμπεριληφθεί και ο κώδικας που χρησιμοποιήθηκε, πέραν του γραφήματος.