Aprenentatge Bayesià: Naïve Bayes

Observacions:

En aquesta pràctica es presenta el problema de l'aprenentatge bayesià. El problema s'organitza en tres nivells de dificultat: A (sobre 10), el qual és la dificultat màxima, B (sobre 8), que és la dificultat mitjana, i C (sobre 6), que és la dificultat més baixa.

Per aprovar la pràctica és requisit necessari completar satisfactòriament la part C del problema, demostrant així una comprensió fonamental de la matèria. La superació de la mateixa estarà condicionada a la presentació d'una documentació i una defensa adequades. La màxima puntuació s'aconsegueix resolent els problemes de tots els nivells (A, B, C). S'ha de tenir en compte que no es pot fer la part A sense haver fet abans la part B.

Objectius de la pràctica:

Els objectius d'aquesta pràctica són:

- Aplicar l'aprenentatge bayesià en el context d'una aplicació de classificació de tweets segons el seu sentiment.
- Apendre a utilitzar els mètodes de validació explicats a teoria.
- Ser capaç d'aplicar la teoria en problemes reals.
- Fomentar la capacitat per presentar resultats de forma adequada davant d'altres persones.

Materials per a la sessió:

1. Fitxer amb la base de dades de tweets per poder realitzar la pràctica.

Enunciat

Per a aquest exercici teniu a la vostra disposició un fitxer anomenat FinalStemmedSentimentAnalysisDataset.csv on hi trobareu un conjunt de tweets (>1M) ja processats. Es tracta de crear un filtre per decidir si els tweets són positius o negatius. Per poder dur a terme aquesta tasca, s'utilitzaran les xarxes bayesianes.

Per tal de tenir un conjunt de train i un de test, dividirem la base de dades en dos conjunts diferents de manera aleatòria, tal com s'ha explicat a teoria. S'ha de fer un estudi de com afecta a la xarxa el fet d'entrenar amb més o menys quantitat de dades.

Podeu trobar tota la base de dades de tweets sense processar en l'arxiu SentimentAnalysisDataset.csv.

Recordeu balancejar els conjunts de train i test per tal que tots dos conjunts continguin el mateix percentatge de casos de les diferents classes de dades. Penseu que si no ho feu d'aquesta manera podria ser que estiguéssiu entrenant amb totes les dades d'una sola classe i després no classifiquéssiu bé la resta de classes.

Per poder tractar més fàcilment els tweets, aquests han estat processats amb el Lancaster Stemmer de python. Aquest algorisme el que fa és reduir les paraules a la seva arrel. Si creieu que podeu aplicar alguna mena de processat millor que el que us hem facilitat, podeu fer-ho sobre la base de dades sense processar que us facilitem,

sempre i quan ho documenteu correctament. A més a més, s'han eliminat tots els signes de puntuació que podien introduir soroll.

L'arxiu .csv conté quatre columnes diferents en les quals s'indica l'identificador del tweet, el tweet, la data en la qual es va realitzar el tweet i l'etiqueta de la classe a la qual pertany. Així doncs un exemple de tweet seria:

16; I fell in love again; 02/12/2015; 1

El 16 seria l'identificador, "I fell in love again" seria el text, després vindria la data i, per últim, el 1 indicaria que aquest tweet és positiu.

Les etiquetes dels tweets són:
0: negatiu
1: positiu

Avaluació sessió de seguiment

En la sessió de seguiment caldrà tenir:

- Analitzat i codificat l'estructura de dades.
- Dissenyat el conjunt de proves a realitzar i les mesures que s'usaran per avaluar els algorismes
- Tenir construïts el/s diccionari/s
- Tenir definit i construïda la taula d'extracció de P(w_i|c_j) c_j={positiu, negatiu}

Exercici 1: (C)

En aquest exercici es demana implementar una xarxa bayesiana per determinar si els tweets del test són positius o negatius. Per aquesta primera part, s'utilitza tota la base de dades de tweets i el diccionari que obtingueu de la part del train.

A l'informe cal que quedi clar els següents punts, a més a més d'explicar com heu resolt el problema i analitzar-ne el resultats.

Generació de diccionaris (estructura i contingut). Com genereu el/s diccionari/s?
Si hi ha múltiples diccionaris, perquè i com?
Justificació del mètode de validació (cross-validation, leave-one-out, etc.)
Justificació de la mètrica?
Resultats i anàlisi

Exercici 2: (B)

En aquest exercici es demana avaluar la xarxa utilitzant conjunts de train i de diccionari de diferents mides.

1. Ampliar el conjunt de train (podeu determinar vosaltres l'interval). A l'augmentar el nombre de tweets, el diccionari també canviarà de mida.

2. Fixar el conjunt de train, però utilitzar diferents mides de diccionari, d'aquesta manera es pot valorar com afecta la mida del diccionari a l'hora de l'entrenament.

3. Utilitzar sempre la mateixa mida de diccionari, però modificant el conjunt de train, per veure com afecta això a l'entrenament.

    1. Què passa en el primer cas, quan es va ampliant el nombre de tweets i el diccionari per fer el train?
    2. Com afecta la mida del diccionari?
    3. Com afecta la mida del conjunt de tweets de test?

Exercici 3: **(A)**

En aquest exercici es demana avaluar la xarxa utilitzant conjunts de train i de diccionari de diferents mides tal i com es fa en l'apartat B, però en aquest cas, implementant 'Laplace smoothing'.

- Com afecta el 'Laplace smoothing'