Aprenentage Computacional

Cas Kaggle: Fake and Real news

Grau: Matemàtica Computacional i Analítica de Dades

Assignatura: Aprenentage Computacional

NIU de l'estudiant: 1570127

Nom de l'estudiant: Maria Graupera Rodón

Índex

1 Objectius

En aquesta pràctica analitzarem i estudiarem una base de dades per tractar d'assolir els següents objectius:

- Crear un repositori Github on s'expliquen els diversos passos realitzats per a la resolució d'un problema d'Aprenentatge Computacional.
- El projecte serà aplicat a una base de dades de la plataforma Kaggle la qual és Fake and Real news, i constaran de tres parts:
  - **1.**: una explicació dels atributs més importants de la base de dades i de l'atribut a classificar
  - **2.**: breu descripció del mètode d'aprenentatge computacional aplicat, juntament amb els paràmetres escollits.
  - **3.**: i una presentació dels resultats que s'han obtingut.

L'objectiu principal serà crear un classificador el qual ens permeti etiquetar si una noticia és falsa o no ho és, a partir del títol d'aquesta.

2 Exploració i visualització de la base de dades

La base de dades Fake and Real news consta de dos arxius que reben el nom de Fake.csv i True.csv. En l'arxiu Fake.csv es troben les noticies falses i en l'arxiu True.csv es troben les noticies que són certes. Ambdós arxius tenen la mateixa estructura i els mateixos atributs, els quals són els següents:

- Title: ens indica el títol de la notícia.
- Text: ens indica el contingut de la notícia.
- Subject: ens indica el tipus de notícia.
- Date: ens indica la data de la notícia.

Tots els atributs són de tipus objecte. Tanmateix s'han esborrat totes aquelles noticies que els hi mancava text o bé estaven duplicades.

En primer lloc com el que ens interessa saber si una noticia és falsa o no, per tant s'ha creat un atribut nou el qual s'ha anomenat label, aquest atribut és de tipus int, si hi ha un 0 implica que és falsa i si hi ha un 1 ens indica que és certa. Aquest serà el nostre atribut objectiu.

En segon lloc com que l'objectiu del classificador és classificar la noticia en certa o falsa a aprtir del su títol, els atributs que més ens interessen són el label i el title. La resta ens és irrellevant ja que el que volem és classificar en funció del títol.

En tercer lloc s'han fusionat els dos arxius en un de sol, per tal de poder separar la base de dades en train i test.

Finalment per tal de poder fer un classificador amb aquestes dades s'ha de convertir d'alguna manera les paraules a números. Per tal de fer aquesta conversió s'ha aplicat el métode nlp, aquest mètode ens permet passar del llenguatge natural a un vector amb un conjunt de números que ens permetrà realitzar el classificador. D'aquesta manera obtindrem l'atribut vector, el qual conté el vector de floats que surt d'aplicar nlp al'atribut title.

En la nostra base de dades no ens és necessari aplicar un PCA ni un TSNE ja que ja s'està treballant amb una base de dades amb poca dimensionalitat i no és necessari reduir el número d'atributs.

3 Selecció del model

Alhora d'escollir el model que més s'ajusta al nostre dataset, hem considerat provar els models següents: el regressor logístic, el svm, el svm lineal, el svm polinomial de grau 3 i 2, svm sigmoidal, el random forests gini, el random forests entropy amb ne=1000 i md=5, el KNN balltree, el KNN kdtree, el KNN brute, el KNN balltree pesos=distancia, el KNN kdtree pesos=distancia, el KNN brute pesos=distancia.

Un cop executats tots els models obtenim el següent output:

```
Correct classification Logistic         0.5 % of the data:  0.9975764919721296
Correct classification SVM              0.5 % of the data:  0.9872008482278097
Correct classification SVML             0.5 % of the data:  0.9978794304756134
Correct classification SVMP deg3        0.5 % of the data:  0.9981823689790973
Correct classification SVMP deg2        0.5 % of the data:  0.9978794304756134
Correct classification SVMS             0.5 % of the data:  0.5348379279006362
Correct classification RFC              0.5 % of the data:  0.980914874280521
Correct classification RFC etpy         0.5 % of the data:  0.9884126022417449
Correct classification KNN BT           0.5 % of the data:  0.9930324144198728
Correct classification KNN KD           0.5 % of the data:  0.9930324144198728
Correct classification KNN BRT          0.5 % of the data:  0.9930324144198728
Correct classification KNN BT wd=d      0.5 % of the data:  0.9930324144198728
Correct classification KNN KD wd=d      0.5 % of the data:  0.9930324144198728
Correct classification KNN BRT wd=d     0.5 % of the data:  0.9930324144198728

Correct classification Logistic         0.7 % of the data:  0.9987378518238041
Correct classification SVM              0.7 % of the data:  0.989902814590433
Correct classification SVML             0.7 % of the data:  0.9979805629180866
Correct classification SVMP deg3        0.7 % of the data:  0.9987378518238041
Correct classification SVMP deg2        0.7 % of the data:  0.9979805629180866
Correct classification SVMS             0.7 % of the data:  0.5408304934999368
Correct classification RFC              0.7 % of the data:  0.9824561403508771
Correct classification RFC etpy         0.7 % of the data:  0.9893979553199546
Correct classification KNN BT           0.7 % of the data:  0.9941941183894989
Correct classification KNN KD           0.7 % of the data:  0.9941941183894989
Correct classification KNN BRT          0.7 % of the data:  0.9941941183894989
Correct classification KNN BT wd=d      0.7 % of the data:  0.9941941183894989
Correct classification KNN KD wd=d      0.7 % of the data:  0.9941941183894989
Correct classification KNN BRT wd=d     0.7 % of the data:  0.9941941183894989

Correct classification Logistic         0.8 % of the data:  0.998485422188565
Correct classification SVM              0.8 % of the data:  0.9888299886406664
Correct classification SVML             0.8 % of the data:  0.9977281332828474
Correct classification SVMP deg3        0.8 % of the data:  0.9982960999621355
Correct classification SVMP deg2        0.8 % of the data:  0.9982960999621355
Correct classification SVMS             0.8 % of the data:  0.5357819007951533
Correct classification RFC              0.8 % of the data:  0.9827716773949262
Correct classification RFC etpy         0.8 % of the data:  0.9878833775085195
Correct classification KNN BT           0.8 % of the data:  0.9929950776221128
Correct classification KNN KD           0.8 % of the data:  0.9929950776221128
Correct classification KNN BRT          0.8 % of the data:  0.9929950776221128
Correct classification KNN BT wd=d      0.8 % of the data:  0.9929950776221128
Correct classification KNN KD wd=d      0.8 % of the data:  0.9929950776221128
Correct classification KNN BRT wd=d     0.8 % of the data:  0.9929950776221128
```

Figura 1: Output dels diferents models

Podem observar que tots els mètodes classifican molt bé ha exepció del SVMS, que és l'únic que no fa be l classificació.

4 Cross-validation

Apliquem la tècnica de cross-validation amb l'objectiu de garantir que el resultat del nostre model són realment independents a la partició de les dades d'entrenament i proba.

Per fer la cross-validation separem el nostre dataset en dos, la part de test i la part de train. Hem de seleccionar bé aquests dos conjunts. Ja que si agafem un conjunt de test massa gran podríem obtenir overfitting, però si aquest es massa petit tindríem underfitting.

Per evitar això podem fer us de la tècnica anomenada K-fold, on fem k particions de les dades, una d'aquestes k particions s'utilitzarà per a la part de test i les altres per fer el training del model. Amb això aconseguim fer k models i per mesurar la eficàcia es fa la mitjana dels resultats d'aquests k models.

Al nostre cas, primer hem separat el dataset fent que el conjunt de train sigui un 70% i el test un 30%. També hem fet us del K-fold que té la llibreria sklearn amb k = 50 ja que aquest ja ens dona uns bons resultats de la nostra base de dades.

Al aplicar cross-calidation obtenim un accuracy de 0.998 +/- 0.002.

Com que el logistic és un dels classificadors que millor classifica les nostres dades, hem utilitzat aquest classificador per extreure la matriu de confusió següent:

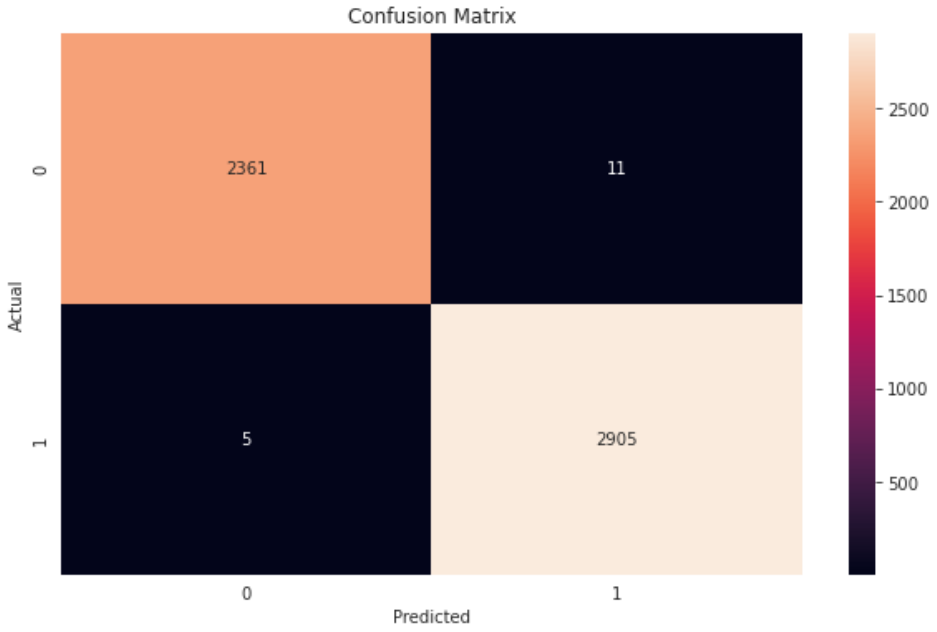
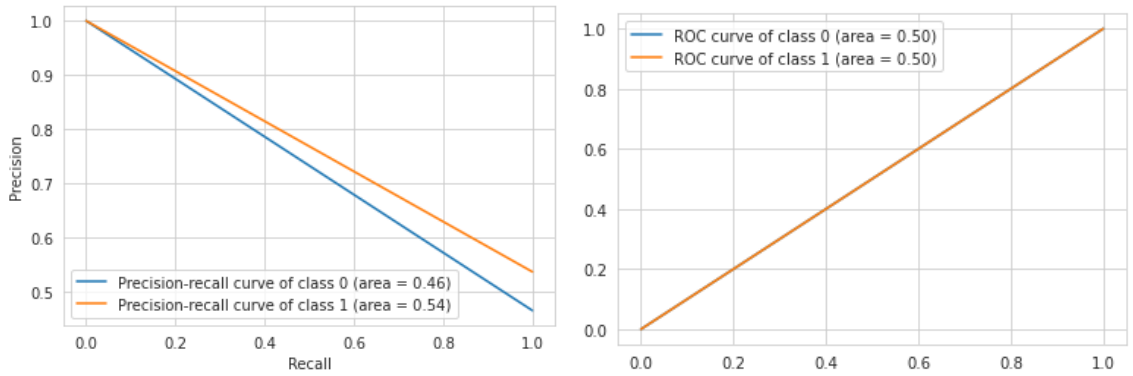


Figura 2: Matriu de confusió

D'aquesta matriu podem extreure que de les 2366 noticies certes que hi ha en total, el nostre classificador classifica 5 com a falsos negatius, i de les 2916 noticies falses 11 són falsos positius.

5 Anàlisi de les mètriques

Per poder graficar les corbes del Precision-Recall i del ROC hem fet servir “average_precision_score”. Tots els resultats són de 1.00 a exepció de la SVMS el qual ens han sortit les següents gràfiques:



(a) PR svcs (b) ROC svcs

Figura 3: PR i ROC de svcs

6 Conclusions

S’ha pogut veure que és pot utilitzar qualsevol mètode de classificació ha excepció del SVMS ja que és l’únic amb un accuracy inferior a 0.98. Tanmateix podem observar que el mètode NLP ens ha sigut molt útil per obtenir auests resultats tan bons.