

Multinomial Naive Bayes


La base de datos *BuzzFeed-Webis Fake News Corpus 2016* posee diferentes artículos periodísticos de una semana cercana a las elecciones estadounidenses de ese año. Se desea entrenar un algoritmo Multinomial Naive Bayes capaz de clasificar los artículos en: “*mayormente falso*”, “*mayormente verdadero*”, “*mezcla de verdadero y falso*” y “*sin contenido factual*”.

(a) *Exploración de datos:*

- Descargar la base de datos en <https://zenodo.org/record/1239675/files/articles.zip?download=1>.

- Construir la base de datos. : Puede usar el siguiente código:

```
import xml.etree.ElementTree as ET
data = {"mainText": [], "orientation": [], "veracity": []}
for filename in os.listdir("articles/"):
    root = ET.parse(f"articles/{filename}").getroot()
    for elem in root:
        if elem.tag in data.keys():
            data[elem.tag].append(elem.text)
data = pd.DataFrame(data)
data = data[data.notna().all(axis="columns")]
```

- Utilice el comando `train_test_split` (sklearn) para definir dos conjuntos de datos. El conjunto de entrenamiento debe contener el 80 % de las muestras, el resto serán de testeo.
- Utilizando `CountVectorizer` (sklearn) pre-procesar los datos del texto principal de los artículos. : Se recomienda convertir el texto a minúscula, utilizar como *Stop Words* las palabras estándar del idioma inglés, eliminar las palabras que aparecen en más del 60 % de los documentos y descartar las palabras vistas en menos de 3 documentos.


(b) *Entrenamiento:* Implementar un MNB de $\alpha = (1, 1, \dots, 1)$ que prediga la veracidad de un artículo a partir de su texto principal (pre-procesado). El código debe estar estructurado de la siguiente manera:

```
class MNB:
    # Inicializar atributos y declarar hiperparámetros
    def __init__(self, ...

    # Etapa de entrenamiento
    def fit(self, X, y):

    # Etapa de testeo soft
    def predict_proba(self, X):

    # Etapa de testeo hard (no repetir código)
    def predict(self, X):
```

(c) *Inferencia:* **Implementar** un método a la clase anterior que calcule el *accuracy* y la *Macro-F1*. Evaluar dichas métricas en el conjunto de testeo. ¿Por qué dan tan diferentes? : Para el cálculo de la F1 debe considerar el caso de *precision* y *recall* nulas.

(d) *Orientación:* Repetir el ejercicio pero para clasificar la orientación política del portal donde fue publicada la noticia (izquierda, derecha o mainstream) a partir del texto principal preprocesado. ¿Siguen siendo válidas las conclusiones extraídas anteriormente? Justificar.