Semana 6
CI-0134: Investigación en Ciencias de la Computación
Paula Monge - B85066

Actividades de la semana:

- ☐ Se investiga sobre cual es una cantidad adecuada de datos para un algoritmo de clasificación, ya que en los trabajos relacionados no se habla de cantidad de datos.

La cantidad de datos que requiere un modelo de ML dependerá de cuantas características, tamaño y variabilidad de la salida esperada se tenga que considerar, entre más complejo sea más datos necesita ingresar.

Se tiene una guia de cuales son los aspectos que dictan la cantidad óptima de datos:

- Complejidad del algoritmo: los algoritmos más complejos requieren una mayor cantidad de datos, si se hace uso de un algoritmo de aprendizaje automático estándar que use aprendizaje estructurado, una cantidad menor de datos será suficiente [1], los algoritmos escogidos para esta investigación entran en esta categoría por lo que no es necesaria mucha cantidad de datos.
- Necesidades de etiquetado: según la necesidad que se vaya a clasificar se define la cantidad de datos, entre más representaciones tenga que aprender para cumplir la clasificación más datos requiere[1], en este caso se maneja una clasificación de Positivo/Negativo y al usar PLN se facilita el procesamiento de dicha clasificación.
- Margen de error aceptable: según el tipo de proyecto se identifica su tolerancia a errores, entre más crítico sea su uso más datos puede ocupar para obtener resultados más precisos[1], si bien la clasificación de los sentimientos que transmite un texto es importante, es tolerante a errores ya que no arriesga la integridad de nada ni nadie.
- Diversidad de entrada: En algunos casos, se debe enseñar a los algoritmos a funcionar en situaciones impredecibles, como lo es un asistente virtual, este debe entender lo que los usuarios le escriben para responder, cuanto más descontrolado esté el entorno, más datos necesitará para su proyecto[1], si bien lo que clasifica son textos de usuarios de Twitter, este no debe generar respuestas sobre lo recibido, sino solo clasificarlos según las palabras que contiene.

Según estos factores se puede definir el tamaño de los conjuntos de datos que necesita para lograr un buen rendimiento del algoritmo y resultados confiables. Ahora profundicemos más y encontremos una respuesta a nuestra pregunta principal: ¿cuántos datos se requieren para el aprendizaje automático?

La forma más común de definir si un conjunto de datos es suficiente es aplicar una regla de 10 veces, esta regla significa que la cantidad de datos de entrada (es decir, la cantidad de ejemplos) debe ser diez veces mayor que la cantidad de parámetros que tiene un modelo. Por ejemplo, si su algoritmo distingue imágenes de gatos de imágenes de perros según 1000 parámetros, necesita 10 000 imágenes para entrenar el modelo, dicha regla sólo es útil en modelos pequeños.

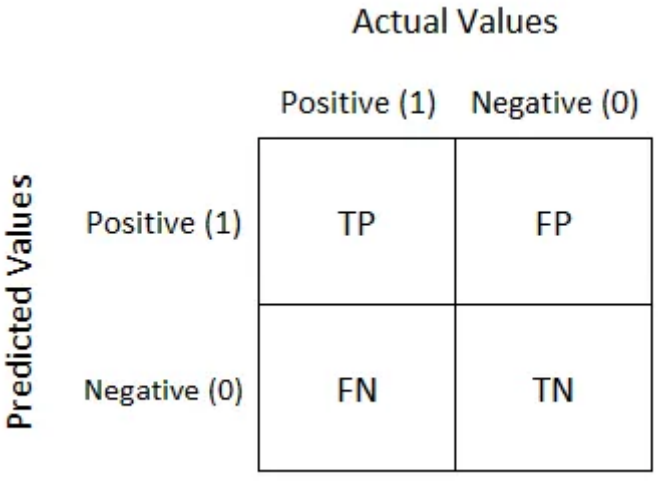
Es bajo este razonamiento que consideramos que una cantidad de 9.000 datos es suficiente para el entrenamiento de los diversos modelos que se utilizan durante la investigación.

- ☐ Reunión de equipo el dia 25/04 para probar el comportamiento de los datos con los hiperparametros fijos

- ☐ Se realiza una investigación sobre las diversas métricas.utilizadas en ML, asi como tambien un reporte del análisis de las mismas en los algoritmos

Cuando se quiere medir el comportamiento de un modelo el uso de la matriz de confusión es ideal ya que esta es una herramienta que permite visualizar el desempeño de un algoritmo. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real.

En nuestro caso se están usando dos clases: Positivo y Negativo {1 y 0}, a continuación se muestra una imagen que muestra la composición de la matriz, tomada de [2]



Explicando los términos que rellenan la tabla con nuestro caso de textos positivos y negativos:

- True Positive (TP): tweets positivos clasificados como positivos
- False Positive (FP): tweets negativos, clasificados como positivos
- False Negative (FN): tweets positivos clasificados como negativos

- True Negative (TN): tweets negativos, clasificados como negativos

Nota: los valores de la diagonal principal son aquellos que el modelo clasificó de manera correcta.

Una vez explicada la matriz, vamos a ver las métricas relacionadas a esta:

- **Recall**: esta métrica nos indica de todas las clases positivas, cuántas predijimos correctamente.
  Fórmula:

$$Recall = \frac{TP}{TP + FN}$$

- **Precision**: indica de todas las clases que hemos predicho como positivas, cuántas son realmente positivas.
  Fórmula:

$$Precision = \frac{TP}{TP + FP}$$

- F1-Score: ayuda a medir la recuperación y la precisión al mismo tiempo
  Fórmula:

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Accuracy**: de todas las clases (positivas y negativas), cuántas de ellas hemos acertado
  Fórmula:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

- ☐ Reunión de equipo el día 26/04 para experimentar con diversos hiperparametros y así ver cuales de ellos dan los mejores resultados en las métricas.
  Se adjunta un reporte de resultados de los cambios en los algoritmos

Actividades para la próxima semana:

Todas estas actividades pueden ser realizadas por mi compañero o por mi.

1. Crear la presentación para la presentación del Avance 2
2. Corregir lo escrito en el Avance 1, si ya se proporcionó retroalimentación

3. Crear cualquier material que sirva para la presentación: gráficos, tablas, etc
4. Experimentación con datos en inglés para tener un contraste de los resultados en los distintos idiomas.

Referencias

[1] "How Much Data Is Required for Machine Learning?" postindustria_. https://postindustria.com/how-much-data-is-required-for-machine-learning/ (accedido el 24 de abril de 2023).

[2] "Understanding Confusion Matrix". Towards Data Science. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 (accedido el 26 de abril de 2023).