UNIVERSIDAD DE ANTIOQUIA
1803

Entrega 1

Introducción a la inteligencia artificial

Andres David Medina Herrera, Miguel Arias Londoño

cc1003562902, cc1017272204

Facultad de ingeniería, Universidad de Antioquia Sede Medellín

andres.medina2@udea.edu.co - miguel.ariasl@udea.edu.co

26/09/2023

1. Descripción del problema predictivo a resolver.
   En base a un conjunto de datos de entrenamiento y de prueba, se busca predecir la interacción magnética entre dos átomos en una molécula es decir, la constante de acoplamiento escalar,la predicción de esta constante para entre pares de átomos en moléculas,el tipo de acoplamiento y cualquier característica que se pueda crear a partir de los archivos dados de estructura de la molécula,es el objetivo de este reto.

   No se predecirá todos los pares de átomos en cada molécula, sino solo los pares que se enumeran explícitamente en los archivos de entrenamiento y prueba,además se excluye cualquier predicción de la constante escalar que involucre átomos de flúor.

2. Dataset que va a utilizar.
   Mi equipo encontró en Kaggle “Predicting Molecular Properties”, la cual es una competición que presenta un dataset con 131 mil archivos y 47 columnas.
   Se encuentran los archivos de:
   - train.csv: el conjunto de entrenamiento,donde se encuentran: el nombre de la molécula donde se origina la constante de acoplamiento,índices atómicos del par de átomos que crean el acoplamiento y la constante de acoplamiento escalar que queremos poder predecir.
   - test.csv: el conjunto de prueba.
   - Structures.zip/Structures.csv: contiene archivos sobre la estructura molecular (xyz), con número de átomos de la molécula y las coordenadas cartesianas X, Y y Z:

   Link de la competición: Predicting Molecular Properties | Kaggle

3. Métricas de desempeño requeridas (de machine learning y de negocio).

   El desempeño del modelo será evaluado con base en el logaritmo del error absoluto medio, se calculará para cada tipo de acoplamiento escalar y luego se promedian entre los tipos, de modo que una disminución del 1 % en MAE para un tipo proporciona la misma mejora en la puntuación que una disminución del 1 % para otro tipo.

$$score = \frac{1}{T}\sum_{t=1}^{T}\log\left(\frac{1}{n_t}\sum_{i=1}^{n_t}|y_i - \hat{y}_i|\right)$$

   Dónde:

   - T es el número de tipos de acoplamiento escalar
   - n_t es el número de observaciones de tipo t
   - y_i es la constante de acoplamiento escalar real para la observación
   - $\hat{y}_i$ es la constante de acoplamiento escalar predicha para la observación

   Para este caso un métrica de negocio no tiene explicación aparente,debido a que se está trabajando con estructuras moleculares y su predicción las aplicaciones para este tipo de modelos afectan desde las ciencias farmacéuticas hasta la ingeniería de alimentos o materiales,por ende se considera que si bien tiene un efecto económico no es posible definir una métrica de negocio relacionable aparentemente

4. Un primer criterio sobre cuál sería el desempeño deseable en producción.

   Para la métrica,utilizada (MAE) para cualquier grupo tiene un piso de 1e-9, de modo que la puntuación mínima, es decir,la mejor posible para predicciones perfectas es aproximadamente de -20,7232, se tiene que los ganadores del concurso tuvieron un puntaje de -3.23968, por ende nuestra única métrica es el criterio más importante,se podría espera un desempeño dentro del rango entre -1 y el puntaje ganador,si bien la aplicación de este competencia tiene fines investigativos,alcanzar valores dentro de ese rango significan una buena implementacion de metodologias para este tipo de proyectos de aprendizaje.