Torneo SNA 2025

Predicción de la lipofilicidad de una molécula

Introducción

En esta parte de la práctica entrenaremos una red neuronal para predecir la lipofilicidad, afinidad de una molécula para disolverse en grasas, lípidos y disolventes no polares. La lipofilicidad es un valor real positivo (afinidad por disolventes no polares) o negativo (afinidad por disolventes polares). Aunque en teoría puede tener cualquier valor de la escala real, la mayoría de los compuestos conocidos están entre el -3 y el 7. Para el ejercicio se usará un dataset de MoleculeNet modificado dónde ciertas etiquetas han sido eliminadas y tendrán que predecirse para el torneo.

Dataset

Cada grafo representa una molécula, los nodos son elementos químicos y las aristas enlaces entre ellos. Los grafos tienen los siguientes atributos.

- x[?, 9] : tensor con los atributos de los nodos como peso atómico, orbitales sp2 y sp3, carga relativa ... etc. Cada grafo tiene un número de nodos diferentes, la mediana está 27 nodos.
- edge_index[2, ?]: tensor con las aristas del grafo. Cada grafo tiene un número de aristas distinto, la mediana está en 60.
- y[1]: tensor con la lipofilicidad del grafo.
- nid[1]: identificador de cada grafo, sirve para enviar tus soluciones al servidor y que se evalúe tu solución.
- train_idx[3360]: tensor con las posiciones de los grafos del dataset de los que se tiene etiqueta y se pueden usar para entrenar/validar.
- test_idx[840]: tensor con las posiciones de los grafos de los que no se tiene etiqueta y por lo tanto no se pueden usar para entrenar/validar. ESTOS NODOS SERÁN LOS QUE SE EVALÚEN EN EL TORNEO.

Entrega

Para esta segunda parte de la práctica de la asignatura se debe entregar el código con el proceso de diseño, entrenamiento y validación, así como una memoria que documente dicho código. Puedes usar el código que usaste en el torneo o entregar una versión mejorada.

Estructura de la memoria

La memoria debe contener las siguientes secciones

- **Portada**: con el nombre del alumno.
- **Diseño de la GNN**: descripción de la arquitectura usada (update, mensaje, agregación).
- **Proceso de entrenamiento**: descripción del proceso de entrenamiento incluyendo cualquier técnica de data augmentation o de sampleo utilizada.
- **Proceso de validación**: descripción del proceso de validación utilizado para validar la arquitectura propuesta.
- **Resultados**: Resultados de tu modelo sobre el conjunto de test (se publicará después del torneo).
- **Conclusiones y trabajo futuro**: Resumen del trabajo realizado, valoración de la GNN presentada, limitaciones y posibles mejoras.
- **Bibliografía**