Fundamentos de Deep Learning: Proyecto - entrega # 1

Docente: *Raúl Ramos Pollán*
Semestre: *2023-01*
Fecha: *02-04-2023*

Autor: *Frank Sánchez Restrepo*
Correo electrónico: *frank.sanchezr@udea.edu.co*
Repositorio: *https://github.com/frank-zr/PD-DeepLearning.git*

Título

Predicción de la progresión de la enfermedad de Parkinson

Introducción

La enfermedad de Parkinson (PD) es un trastorno neurodegenerativo incapacitante que afecta los movimientos, la cognición, y otras series de funciones normales. En la actualidad no existe una cura para esta enfermedad y los tratamientos disponibles se enfocan en la identificación temprana de los síntomas para su apropiado manejo y reducción. Las investigaciones en el tema proyectan un incremento en la población que se verá afectada por la enfermedad de Parkinson, implicando un desafío a nivel de la salud pública y a nivel económico.

Las investigaciones indican que anomalías, cambios y/o presencia de ciertas proteínas o péptidos desempeñan un papel importante en la aparición y el desarrollo de la enfermedad. Con la ayuda de la ciencia de datos se puede obtener una mejor comprensión de esto, y podría proporcionar pistas importantes para el desarrollo de nuevas terapias farmacológicas para retrasar la progresión de la enfermedad de Parkinson.

Objetivos de machine learning

El objetivo es predecir las puntuaciones de la escala UPDR (Unified Parkinson's Disease Rating), que mide la progresión de la enfermedad de Parkinson. Es una evaluación clínica integral de los síntomas motores y no motores asociados con la enfermedad de Parkinson. Se utilizarán datos de los niveles de abundancia (mediciones) de proteínas y péptidos a lo largo del tiempo en los pacientes. Es decir, para cada paciente se tienen varias visitas en las cuales han sido registrados los niveles de proteínas/péptidos y las puntuaciones de la escala UPDR.

Datos (Dataset)

El conjunto de datasets utilizados pertenecen a la competencia de Kaggle AMP-Parkinson's Disease Progression Prediction (Actualmente abierta). Datos con las mediciones de proteínas

y péptidos de pacientes con la enfermedad de parkinson para predecir la progresión de la enfermedad. Los diferentes dataset son un conjunto de archivos .csv con la información de relevancia:

train_proteins.csv: (232742 filas/observaciones)
Frecuencias de la expresión de proteínas agregadas a partir de los datos de los niveles de los péptidos.

> ***visit_id:*** Código de identificación para la visita del paciente
> ***visit_month:*** El mes de la visita, relativa a la primera visita del paciente
> ***patient_id:*** Código de identificación del paciente
> ***UnitProd:*** Código internacional para la proteína.
> ***NPX:*** Expresión normalizada de proteínas (Medición). La frecuencia de aparición de la proteína en la muestra. Puede no tener una relación 1:1 con los péptidos componentes.

train_peptides.csv: (981835 filas/observaciones)
Datos de los niveles de abundancia de los péptidos. Las proteínas están conformadas por agregaciones de péptidos, es decir, estos últimos son las subunidades que componen las proteínas.

> ***visit_id:*** Código de identificación para la visita del paciente
> ***visit_month:*** El mes de la visita, relativa a la primera visita del paciente
> ***patient_id:*** Código de identificación del paciente
> ***UnitProd:*** Código internacional para la proteína. Hay varios péptidos por proteína.
> ***Peptide:*** Identificación del péptido
> ***PepideAbundance:*** Medición de abundancia

train_clinical_data.csv: (2616 filas/observaciones)
Datos de la evaluación clínica relacionada con la severidad de la enfermedad en los pacientes, según escala Unified Parkinson's Disease Rating Scale - UPDRS, que se encuentra divida en cuatro métricas o índices.

> ***visit_id:*** Código de identificación para la visita del paciente
> ***patient_id:*** Código de identificación del paciente
> ***visit_month:*** El mes de la visita, relativa a la primera visita del paciente
> ***updrs_[1-4]:*** La puntuación del paciente para la parte N (N: 1..4) de la Escala de valoración unificada de la enfermedad de Parkinson (Unified Parkinson's Disease Rating Scale - UPDRS). Los números más altos indican síntomas más graves. Cada subsección cubre una categoría distinta de síntomas, como el estado de ánimo y el comportamiento para la Parte 1 (updrs_1) y las funciones motoras para la Parte 3 (updrs_3).

updrs_23b_clincial_state_on_medication: Si el paciente estaba tomando o no medicamentos durante la evaluación UPDRS. Se espera que afecte principalmente las puntuaciones de la Parte 3, función motora, (updrs_3).

Métricas de desempeño

La evaluación del desempeño del modelo, será por medio de la métrica de error porcentual absoluto medio simétrico (SMAPE), la cual es una medida de precisión basada en errores porcentuales (o relativos). Definida como:

$$\text{SMAPE} = \frac{100}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

Donde *At* es el valor real y *Ft* es el valor predicho. Y *n* siendo el número de puntos ajustados. No obstante, durante el desarrollo del proyecto se pueden incluir otras métricas para la evaluación del desempeño del modelo.

Para cada visita del paciente en la que se tomó una muestra de proteína/péptido, se deberá estimar sus puntajes UPDRS para esa respectiva visita, y predecir los puntajes UPDRS para cualquier visita potencial, ej. 6, 12 y 24 meses después.

Referencias

AMP -Parkinson's Disease Progression Prediction. Kaggle. (2023). Retrieved April 2, 2023, https://www.kaggle.com/competitions/amp-parkinsons-disease-progression-prediction/data