

PEC2: Actividad y Debate

Machine Learning

La PEC2 consta de una **actividad** (70%) y de un **debate** (30%).

Actividad

Descripción

Los datos a tratar corresponde a un experimento en ratones relacionado con la predicción de 8 clases a partir de los niveles de expresión de proteínas/modificaciones de proteínas que produjeron señales detectables en la fracción nuclear de la corteza.

Cada clase se basa en el genotipo, comportamiento y tratamiento. Según el genotipo, los ratones pueden ser controles (c) o trisómicos (t). Según el comportamiento, algunos ratones han sido estimulados para aprender (context-shock(CS)) y otros no (shock-context(SC)) y con el fin de evaluar el efecto del fármaco memantina (m) en la recuperación de la capacidad de aprendizaje en algunos ratones se ha inyectado la droga y a otros se ha inyectado suero salino (s) como placebo.

Las 8 clases son:

- 1: *c-CS-s*: control-context-shock-salino, 2: *c-CS-m*: control-context-shock-memantina
- 3: *c-SC-s*: control-shock-context-salino, 4: *c-SC-m*: control-shock-context-memantina
- 5: *t-CS-s*: trisómico-context-shock-salino, 6: *t-CS-m*: trisómico-context-shock-memantina
- 7: *t-SC-s*: trisómico-shock-context-salino, 8: *t-SC-m*: trisómico-shock-context-memantina

Los datos para esta actividad se encuentran en dos ficheros. El fichero `data3.csv` y el fichero `clase3.csv`. El fichero `data3.csv` con los niveles de proteínas, tiene 1080 filas (muestras) y 73 columnas. La primera columna es el código de la muestra y el resto de columnas son proteínas. El fichero `clase3.csv` con los valores de la clase de 1 a 8, tiene 1080 filas (muestras) y una columna con el valor de la clase.

El objetivo planteado es la implementación y evaluación de una red neuronal basada en capas densas (fc) para la clasificación de las 8 clases.

Enunciado

Se pide:

- 1- Cargar los datos (`data3.csv` y `clase3.csv`).
- 2- Realizar un estudio exploratorio de los datos con gráficos y tablas.
- 3- Normalizar las expresiones con la transformación **minmax**.
- 4- Separar los datos en **train** (2/3) y **test** (1/3).
- 5- Definir el modelo 1, que consiste en una red neuronal con una capa oculta densa de 35 nodos, con activación **relu**. Añadir un 20% de **dropout**. Proporcionar el **summary** del modelo y justificar el total de parámetros de cada capa.

- 6- Ajustar el modelo 1 con un 20% de validación, mostrando la curva de aprendizaje de entrenamiento y validación con 50 épocas.
- 7- Obtener la tabla de clasificación errónea en test. Y las métricas usuales de evaluación.
- 8- Definir el modelo 2, que consiste en una red neuronal con dos capas ocultas densas de 35 nodos y 15 nodos, con activación **relu**. Añadir un 20% de **dropout** en ambas capas. Proporcionar el **summary** del modelo y justificar el total de parámetros de cada capa.
- 9- Ajustar el modelo 2 con un 20% de validación, mostrando la curva de aprendizaje de entrenamiento y validación con 50 épocas.
- 10- Comparar en test, mediante las métricas de evaluación, los dos modelos.

Informe de la actividad

Se entregaran dos ficheros:

1. Fichero ejecutable (.ipynb) que incluya un texto explicativo que detalle los pasos implementados en el script y el código de los análisis.
2. Informe (pdf) resultado de la ejecución del fichero (.ipynb) anterior.

Puntuaciones de los apartados

Apartado 1 (5%), Apartado 2 (10%), Apartado 3 (5%), Apartado 4 (5%), Apartado 5 (20%), Apartado 6 (10%), Apartado 7 (10%), Apartado 8 (20%), Apartado 9 (10%) y Apartado 10 (5%)

Referencia de la actividad

Chollet, Francois. 2021. Deep Learning with Python, Second Edition. Manning Publications.

Debate

El debate parte principalmente del artículo de revisión de Lecun et al. (ver abajo) como referencia básica sobre Deep Learning. Para dar unas pautas en el debate se plantean las siguientes preguntas:

1. ¿Qué es el “Representation learning”?
2. ¿Cuál es la utilidad del “stochastic gradient learning”?
3. ¿Cuál es la función ReLU? y ¿Cuál es su utilidad?
4. En la práctica, ¿Son un problema los mínimos locales en el entrenamiento de grandes Neural Networks?
5. ¿Qué son las Convolutional Neural Networks? Descripción. Propósito y Aplicaciones.
6. ¿Qué son las Autoencoders Neural Networks? Descripción. Propósito y Aplicaciones.
7. ¿Qué son las Recurrent Neural Networks? Descripción. Propósito y Aplicaciones.
8. ¿Qué son las Memory Neural Networks? Descripción. Propósito y Aplicaciones.
9. Algoritmos generativos tales como Generative Adversarial Networks (GAN), Variational AutoEncoders (VAE)
10. Transformers.
11. Big data.
12. Métodos bioinspirados.
13. Otros temas en Inteligencia Artificial en Biociencias.

Las aportaciones se harán en el foro de discusión del reto 2 donde cada pregunta se corresponde a un foro. Así que se tiene 13 foros para el debate. La última pregunta la podéis usar para comentar aspectos que no estén incluidos en las anteriores. Recordar que lo importante es hacer aportaciones efectivas, no de relleno.

Es suficiente con hacer **5** contribuciones que tenga información/sugerencias/comparación/. . . RELEVANTE en el debate que incluya las referencias bibliográficas que son base o extensión de la contribución aunque siempre es mejor aportar el mayor número de contribuciones que se pueda.

Además del artículo de Lecun et al., se indican dos artículos sobre Deep learning en bioinformática (ver abajo). Seguro que si buscáis encontraréis más artículos que os pueden servir. Es material complementario para vuestra formación y puede dar origen a preguntas diferentes a las propuestas que podéis formular y contestar en la sección “Otros temas en Inteligencia Artificial en Biociencias.”

Referencias del debate

- Yann LeCun, Yoshua Bengio & Geoffrey Hinton (2015) Deep Learning. Nature Reviews, vol 521, 436-444
- Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. Molecular systems biology, 12(7), 878.
- Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. Briefings in bioinformatics, 18(5), 851-869.

Finalmente, ¿Qué hay que entregar?

Hay que subir a la entrega de la PEC2 un fichero zip con:

1. Un documento pdf con las aportaciones realizadas en el debate indicando a que secciones corresponden.
2. Los dos ficheros correspondientes a la actividad.