Temas selectos de ciencia de datos
Tarea 1

Para entregar el 22 de septiembre de 2025

Instrucciones.

- Los ejercicios que no son computacionales se entregan en PDF-latex.
- Los ejercicios computacionales se entregan en un solo jupyter notebook por tarea. Incluye cualquier archivo, imágen, datos, etcétera, que se requiera para poder ejecutarse.
- Todos tus entregables debes nombrarlos con el siguiente formato:

  `TareaXXX_Nombre_Apellido.ipynb`, `TareaXXX_Nombre_Apellido.pdf`,

  `TareaXXX_Nombre_Apellido.csv`, etcétera.
- Si se ejecuta en Colab, asegúrate de dar acceso a los archivos necesarios para ejecutar el código
- Las tareas se entregan en la plataforma Moodle
- Las tareas son individuales, a menos que se especifique lo contrario
- Las indicaciones respecto al uso de IA y demás aspectos éticos a considerar en las tareas y proyectos se dieron al inicio del curso. La calificación se sujetará a tales indicaciones.

1. Calcula lo siguiente:

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 2 & 3 \end{pmatrix} + \begin{pmatrix} 7 & 9 \end{pmatrix}$$

   Usa broadcasting de tal forma que la operación esté bien definida. Antes, averigua y describe qué es broadcasting, en el contexto de numpy.

2. Considera las redes multicapa (con funciones de activación lineal) que se muestran en la Figura 1:

   a) Describe una ventaja (al menos) de la red A sobre la red B

   b) Describe una ventaja (al menos) de la red B sobre la red A

3. Considera la regularización l_1 de una función de costo L que asumimos es continuamente diferenciable:

$$\tilde{L}(\mathbf{w}; \mathbf{X}, \mathbf{y}) = L(\mathbf{w}; \mathbf{X}, \mathbf{y}) + \alpha \sum_i |w_i|.$$

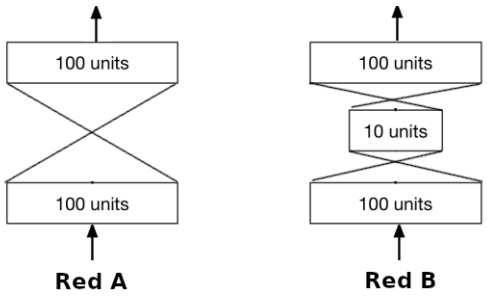

Figura 1: Esquema de dos redes neuronales.

a) Como en clase, considera una aproximación de segundo orden alrededor de $\mathbf{w}^*$, y muestra que la aproximación regularizada de $\tilde{L}$ es

$$\hat{L}(\mathbf{w}) = L(\mathbf{w}^*) + \sum_i \left(\frac{1}{2} H_{ii}(w_i - w_i^*)^2 + \alpha |w_i| \right),$$

donde se asume que los datos están decorrelacionados (*blanqueados*), tal que $\mathbf{H}$ es una matriz diagonal con $H_{ii} > 0$.

b) Muestra que $\hat{L}$ se minimiza en

$$w_i = \text{signo}(w_i^*) \, \text{máx} \left\{ |w_i^*| - \frac{\alpha}{H_{ii}}, 0 \right\}.$$

¿En qué casos tendremos soluciones *sparse* (donde $w_i = 0$)?

4. Considera un problema de clasificación multiclase y una red neuronal densamente conectada con una capa oculta, como se muestra en la Figura 2. Considera también la función sigmoide como activación de las unidades ocultas, la función softmax para las estimaciones en la capa de salida y cross-entropy como función de costo.

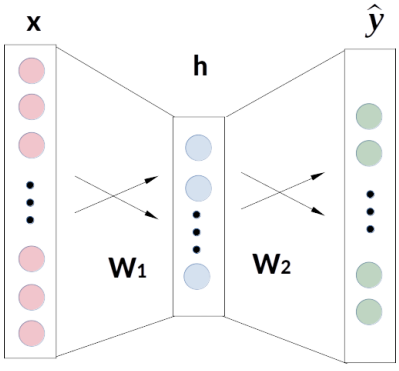

Figura 2: Red neuronal densamente conectada con una sola capa oculta.

a) Muestra que softmax es invariante a traslaciones (constantes) del vector de entrada, es decir, para cualquier vector $\mathbf{x}$ y cualquier constante c:

$$\text{softmax}(\mathbf{x}) = \text{softmax}(\mathbf{x} + c),$$

donde la operación $\mathbf{x} + c$ se realiza con broadcasting. Recuerda que

$$\text{softmax}(\mathbf{x})_i = \frac{e^{\mathbf{x}_i}}{\sum_j e^{\mathbf{x}_j}}.$$

Lo anterior es útil cuando se escoge $c = -\,\text{máx}(\mathbf{x})$, es decir, quitando el valor mayor en todos los elementos de $\mathbf{x}$, para estabilidad numérica.

b) Para un escalar x, muestra que el gradiente de la función sigmoide es

$$\sigma(x)(1 - \sigma(x))$$

c) Muestra que el gradiente en la capa de salida es

$$\frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{z}} = \hat{\mathbf{y}} - \mathbf{y},$$

donde $\hat{\mathbf{y}} = \text{softmax}(\mathbf{z})$, para algún vector $\mathbf{z}$ que proviene de la capa de salida. ¿Qué interpretación puedes dar a ésa expresión?

La función de costo, como mencionamos al inicio, es la cross-entropy:

$$L(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_i y_i \log(\hat{y}_i),$$

donde $\mathbf{y}$ es un vector *one-hot* de las clases y $\hat{\mathbf{y}}$ es el vector de probabilidades estimadas.

d) Considerando los incisos anteriores, obtén los gradientes respecto a los parámetros del modelo calculando

$$\frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{x}},$$

para obtener de ésta forma, las ecuaciones de backpropagation de la red.

Recuerda que el paso forward calcula las activaciones: $\mathbf{h} = \sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)$ y $\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_2\mathbf{h} + \mathbf{b}_2)$. Recuerda también que la función de activación en un vector, se aplica entrada por entrada.

5. En el `moodle` del curso, encontrarás un conjunto de datos que corresponden a un extracto del Free Music Archive (FMA) [1], que es una base de datos muy extensa de archivos de audio usada para diversas tareas de MIR. [1].

   Encontrarás varios conjuntos de datos en formato `csv` que corresponden a dos grupos: uno de entrenamiento (archivos que empiezan con el prefijo `tracks_fma_train`) y uno de prueba (archivos que empiezan con el prefijo `tracks_fma_test`). Para cada uno hay un archivo con metadatos (`..._metadata.csv`) que incluye la variable de respuesta `track.genre1`[2] **excepto** para los datos de test.

[1] Este subconjunto de datos lo construí basado en los archivos originales y haciendo mucho preproceso de los mismos. En general, traté de incluir la información relevante para éste ejercicio, quitando valores nulos que no podían estimarse, entre otras cosas. Los datos originales son considerablemente mas grandes.

[2] El género que contiene la columna track.genre1 fue obtenido siguiendo el esquema jerárquico descrito en el paper original del dataset [1].

También, se proporcionan diversos archivos que contienen características de audio y de la señal (`..._features.csv`) en forma de indicadores o estadísticas de la señal, por lo que las escalas pueden variar [3].

a) Usando los datos de `train`, realiza un análisis exploratorio de los datos, usando métodos de reducción de dimensión que consideres apropiados. ¿Qué características tiene la variable de respuesta? ¿Puedes identificar los géneros, o al menos, algún subconjunto de ellos?

Utiliza las características de audio y de la señal que creas convenientes y repórtalo junto con todos tus hallazgos.

b) Usando también los datos de entrenamiento, ajusta una red neuronal con `pytorch` para predecir el género. Genera un conjunto de datos de entrenamiento y validación para verificar su ajuste. Define la arquitectura de la red, y reporta todos los detalles del modelo y el preprocesamiento de los datos que hayas usado (características de audio y señal usadas, reducción de dimensión, regularización, optimizador, estandarización, escalamiento, etcétera). Reporta las métricas de desempeño y gráficas que creas conveniente.

c) Usando la red neuronal ajustada en el paso previo, predice el género de los datos de `test`. Debes entregar un archivo csv con el `track_id` de cada registro y el género que predijo tu modelo. El que obtenga el mejor resultado, recibirá puntos extra.

6. Leé y haz un resúmen corto del ensayo "*Una nota personal sobre música, sonido y electrónica*", de Daphne Oram. Incluye tu opinión personal sobre el mismo.

[3] Las características de audio se generaron en su momento, por los autores del dataset, mediante la API de Spotify (antes Echonest). Las características de la señal se obtuvieron también por los autores mediante diferentes análisis de la señal de audio de los tracks, y se presentan mediante estadísticos (promedio, desviación estándar, mediana, mínimos, máximos, kurtosis, entre otros) calculados en ciertas porciones de la señal. Más adelante en el curso, hablaremos de éstas características con mayor detalle.

Bibliografía

[1] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference*, 2017.