Examen Introducción al Machine Learning Aplicado al Audio: Teoría e Implementación.

Profesor: Rodolfo A. Lobo

Enero 2024

1. Explique cada parte involucrada en una clasificación binaria dentro de la función Binary Cross Entropy Loss (Pérdida de Entropía Cruzada Binaria):

$$\mathcal{L}(y, \hat{y}) = -\sum_{k=1}^{n} y \cdot \log(\sigma(x_i)) + (1 - y) \cdot \log(1 - \sigma(x_i)) \tag{1}$$

recuerde que en general, bajo nuestra notación $\hat{y} = \sigma(x)$. Utilice las clases $c \in \{0, 1\}$ para ejemplificar que ocurre en un problema de clasificación binaria y cómo la expresión anterior nos ayuda a "castigar"el modelo cuando comete errores.

- (1pt) Dibuje cada parte de la función, es decir, la porción que mide el error de la clase 1 y la porción que mide el error de la clase 0. Expliqué por qué y cómo funciona cada parte.

  Respuesta: Los gráficos de la función por partes dependen del error cometido dada la clase que queremos clasificar a través de nuestro modelo. Es decir:

  Si $y = 1$ Analizamos esta parte del sumatorio $y \cdot \log(\sigma(x_i))$
  Si $y = 0$ $(1 - y) \cdot \log(1 - \sigma(x_i))$ **(0.5 pts)**

  Para construir el gráfico debemos tomar valores importantes para el argumento de cada función logaritmo, es decir $\sigma(x_i) = 0$, $\sigma(x_i) = 1$ o límites $\sigma(x_i) \to 0$, $\sigma(x_i) \to 1$; recordando que $\sigma(x_i) = \dfrac{1}{1 + e^{-x_i}}$. Obteniendo:

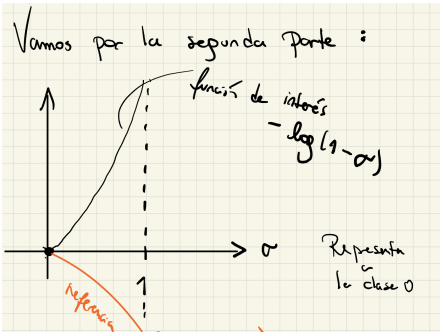


Figura 1: Para $y = 0$ el error es cero cuando la clase obtenida por $\sigma(x_i) = 0$ y crece asintoticamente en la recta $x = 1$ obteniendo valores que tienden a infinito **(0.25 pts)**.

- (1pt) Defina qué es Entropía, ejemplifique utilizando al menos un par de conjuntos (cómo lo visto en clases).

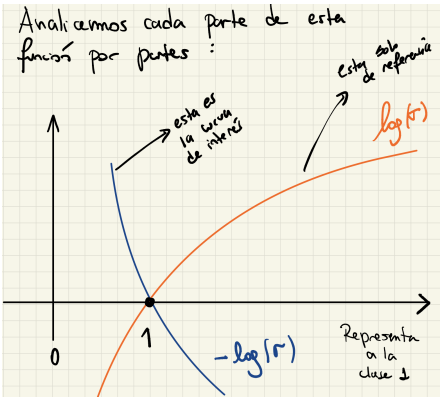


Figura 2: Para $y = 1$, de forma similar al caso anterior; para $\sigma(x_i) = 1$ el error cometido es igual a cero y cuando $\sigma(x_i) \to 0$ el error tiende al infinito **(0.25 pts)**.

Respuesta: La Entropía se define como la esperanza de la sorpresa. La sorpresa, definida de manera casi literal, es una medida inversamente proporcional a la probabilidad que permite medir el nivel de "sopresa"con que un evento puede aparecer en un experimento. Otra forma de pensar la definición de entropía es como una medida del desorden que tiene dicho espacio de eventos. Por ejemplo, si tenemos la misma cantidad de elementos de dos clases y todos con la misma probabilidad de suceder, la entropía es máxima pues todo nos sorprendería de la misma forma.
La Entropía se define matemáticamente por:

$$E[s] = \sum s_i p\left(x_i\right)$$

analogamente, dado que $s = log\left(\frac{1}{p(x)}\right)$ donde x_i es el evento

$$E[x_i] = \sum log\left(\frac{1}{p(x_i)}\right) p\left(x_i\right)$$

utilizando propiedades de logaritmo obtenemos:

$$E[x_i] = -\sum p\left(x_i\right) \cdot log\left(x_i\right) \quad \textbf{(0.5pts)}$$

El ejemplo visto en clases era simplemente observar el nivel de sorpresa al sacar de un conjunto de elementos uno aleatoriamente, dado ese nivel de sopresa y las probabilidades simples de cada evento (solo dos posibles, sacar un círculo o sacar una cruz) podíamos calcular la entropía en diferentes casos, por ejemplo **(0.5pts)**:

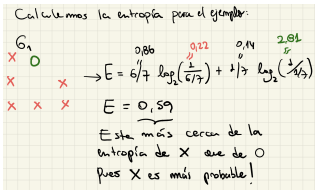


Figura 3: Dado un círculo y 6 cruces la probabilidad de obtener un círculo es $\frac{1}{7}$ y de obtener una cruz $\frac{6}{7}$ con esos dos números podemos calcular la Entropía para esa distribución de datos.

2. Explique cada parte del modelo LSTM:

- (1pt) ¿Cuál es el rol de cada sub modelo al interior de la célula?. Detalle los modelos, funciones de activación, dimensiones de entrada y salida.

  Respuesta: La célula se compone de 4 partes fundamentales. Las dimensiones de entrada son secuencias de tamaño d y las redes neuronales internamente poseen la misma dimensión de entrada y salida para conversar el tamaño la información o características. La unión de filas en el dibujo corresponde a concatenación de vectores, mientras que la bifurcación de las líneas en el dibujo corresponde a una copia del vector, mientras que los símbolos $\times$ y $+$ corresponden a operaciones elementales (multiplicación punto a punto y suma vectorial) **(0.20pts)**. Dadas por:

  (a) El estado de la célula representando las memorias de largo plazo, sufre pequeñas interacciones representadas por operaciones elementales (sumas y multiplicaciones punto a punto) que terminaran eliminando informacióno agregando información de las otras secciones **(0.20pts)** .

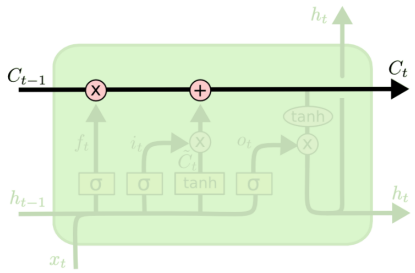


Figura 4: Estado de la célula.

La compuerta de "olvido"o forget gate. La misma notación sugiere la tarea de cada sección. En este caso f de forget. Se utiliza un función sigmoide y una red neuronal que aprenderá a dar pesos entre 0 y 1 a las componentes que debe recordar del vector de entrada **(0.20pts)**.

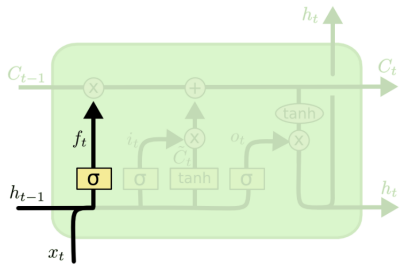


Figura 5: Compuerta de olvido o forget gate.

Puerta de entrada o input del modelo. La caja con el símbolo σ representa una red neuronal que dará pesos a aquellas componentes que deseamos aprender de este nuevo vector y la caja con la función tangente hiperbólica aprenderá información o será un nuevo estado que estará posteriormente ponderado por la salida del modelo σ o input. Este pasará al estado de la célula mediante una operacion elemental de suma vectorial **(0.20pts)**.

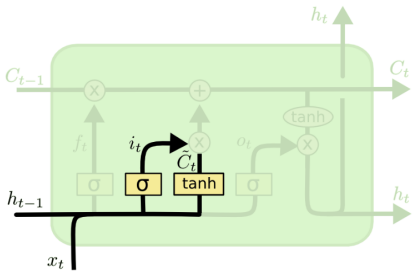


Figura 6: Compuerta de entrada o input gate.

Finalmente, la salida o el output del modelo que va a ponderar mediante una red neuronal la información combinada del estado de la célula a travás de la red neuronal con función de activación *tanh*, la ponderación dependerá de la información del vector de entrada **(0.20pts)**.

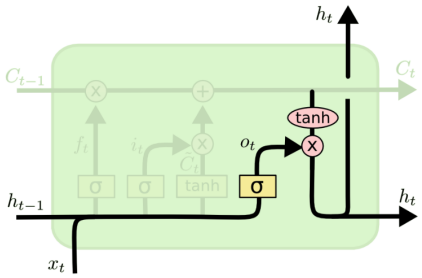


Figura 7: Compuerta de salida u output gate.

- (1pt) Explique cuál es la principal diferencia entre este modelo y las redes neuronales convencionales. De al menos 3 posibles ejemplos de uso para este modelo.

  Respuesta:

  (a) Las redes neuronales convencionales no pueden trabajar con secuencias de vectores. Es decir, no permiten un tratamiento adecuado para trabajar con información temporal.

  (b) Una red neuronal equivale a un modelo de neurona artifical repetido varias veces formando capas. En este caso, tenemos dos modelos diferentes interactuando a través de operaciones elementales. **(0.75pts)**

  (c) Aplicaciones: clasificación de series temporales, predicción de series temporales, generación de texto. **(0.25pts)**

3. Dado el modelo lineal:

$$g_\theta(x) = \theta_1 x + \theta_0, \tag{2}$$

y la función de pérdida:

$$\mathcal{L}(\theta, x_i, y_i) = \frac{1}{n}\sum_{k=1}^{n} \left(g_\theta(x_i) - y_i\right)^2 \tag{3}$$

y dados lo siguientes parámetros de inicialización:

- $\alpha = 0.1$
- $\theta_0 = 0$
- $\theta_1 = 0$

- (1pt) Escriba las ecuaciones de Gradiente Descendiente, calcule las derivadas de forma explícita.

  Respuesta: En este caso tenemos tres puntos en nuestro dataset, por lo tanto $n = 3$:

$$\nabla_\theta \mathcal{L} = \begin{bmatrix} \frac{\partial}{\partial \theta_0}\mathcal{L}(\theta_0) \\ \frac{\partial}{\partial \theta_1}\mathcal{L}(\theta_1) \end{bmatrix} = \begin{bmatrix} \frac{2}{3}\sum(\theta_1 x_i + \theta_0 - y_i)\cdot 1 \\ \frac{2}{3}\sum(\theta_1 x_i + \theta_0 - y_i)\cdot x_i \end{bmatrix} \quad \textbf{(0.5 pts)}$$

  Luego, las ecuaciones de gradiente descendiente de forma explícita son dadas por:

$$\begin{bmatrix} \theta_0^{k+1} \\ \theta_1^{k+1} \end{bmatrix} = \begin{bmatrix} \theta_0^{k} \\ \theta_1^{k} \end{bmatrix} - 0.1 \cdot \begin{bmatrix} \frac{2}{3}\sum(\theta_1^k x_i + \theta_0^k - y_i)\cdot 1 \\ \frac{2}{3}\sum(\theta_1^k x_i + \theta_0^k - y_i)\cdot x_i \end{bmatrix} \quad \textbf{(0.5 pts)}$$

- (1pt) Realice una iteración del modelo de gradiente descendiente para todo el conjunto de entrenamiento (no estocástico):

x	y
1	1
2	0.5
3	2

Tabela 1: Datos de entrenamiento

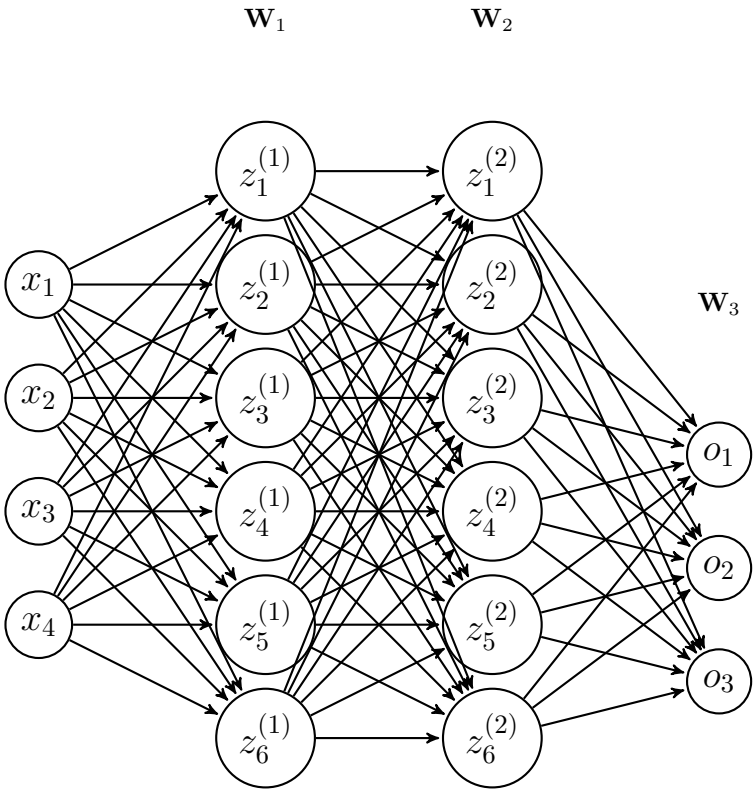
  Respuesta: Sustituyendo los valores de θ_0 y θ_1 inicial en la ecuación para $k = 0$:

$$\begin{bmatrix} \theta_0^{1} \\ \theta_1^{1} \end{bmatrix} = \begin{bmatrix} \theta_0^{0} \\ \theta_1^{0} \end{bmatrix} - 0.1 \cdot \begin{bmatrix} \frac{2}{3}\sum(\theta_1^0 x_i + \theta_0^0 - y_i)\cdot 1 \\ \frac{2}{3}\sum(\theta_1^0 x_i + \theta_0^0 - y_i)\cdot x_i \end{bmatrix} \quad \textbf{(0.5pts)}$$

  Es decir:

$$\begin{bmatrix} \theta_0^{1} \\ \theta_1^{1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} \frac{2}{3}\sum(0 \cdot x_i + 0 - y_i)\cdot 1 \\ \frac{2}{3}\sum(0 \cdot x_i + 0 - y_i)\cdot x_i \end{bmatrix} = \begin{bmatrix} -0.1 \cdot \frac{2}{3}(-1 - 0.5 - 2) \\ -0.1 \cdot \frac{2}{3}(-1 - 1 - 6) \end{bmatrix} \approx \begin{bmatrix} 0.23333 \\ 0.53333 \end{bmatrix} \quad \textbf{(0.5pts)}$$

4. Dado el siguiente diagrama de una red neuronal donde x_j con $j = 1, \ldots, 4$ siendo las componentes de entrada o información de entrada y además, donde el sesgo o bias es igual a cero para cada capa, es decir $\mathbf{b}_i = 0$, $\forall i = 1, \ldots,$ `numerodecapas`:



Responda las siguientes preguntas:

(a) (2pt) Cuántas capas tiene este modelo.

Respuesta: El modelo tiene 3 capas. Una de entrada W_1, una oculta W_2 y una de salida W_3.

(b) (2pt) Cuántos parámetros tiene este modelo.

Respuesta: La matriz W_1 tiene 4×6 parámetros; la matriz W_2 tiene 6×6 parámetros y la matriz de salida tiene 6×3 parámetros. Obtenemos un total de: $24 + 36 + 18 = 78$ parámetros.

(c) (2pt) Sea $z_i^{(j)}$ la salida de la neurona i en la capa j. Cuál es la dimensión de las matrices involucradas si los modelos dentro de cada potencial de activación son modelos lineales de la forma $W_i\mathbf{x} + \mathbf{b}_i$.

Respuesta: $W_1 \in \mathbb{R}^{6\times4}$ con $b_1 \in \mathbb{R}^{6\times1}$, $W_2 \in \mathbb{R}^{6\times6}$ con $b_2 \in \mathbb{R}^{6\times1}$ y $W_3 \in \mathbb{R}^{6\times3}$ con $b_3 \in \mathbb{R}^{3\times1}$

(d) (2pt) Escriba las ecuaciones que representan a este modelo (ya sea de forma anidada o como un sistema de ecuaciones).

Respuesta: Como sistema de ecuaciones

$$\begin{aligned}\mathbf{z}^1 &= \sigma\left(W_1\mathbf{x} + \mathbf{b}_1\right)\\ \mathbf{z}^2 &= \sigma\left(W_2\mathbf{z}^1 + \mathbf{b}_2\right)\\ \mathbf{o} &= \sigma\left(W_3\mathbf{z}^2 + \mathbf{b}_3\right)\end{aligned}$$

5. En relación a las redes convolucionales:

(a) (2pt) Describa la arquitectura de una red neuronal convolucional (CNN) utilizada para la clasificación de imágenes. Utilice un dibujo de ser necesario.

Respuesta: Una red convolucional se compone de capas convolucionales, dadas por filtros o kernels, además tiene etapas de normalización, funciones de activación y pooling. Cuando observamos una arquitectura de red convolucional los bloques representan las dimensiones de salida de cada capa y no el modelo en si mismo. En el caso de clasificar, necesitamos como capa de salida una red completamente conectada y una función softmax de activación para obtener probabilidades de salida. Un dibujo podría estar dado por:

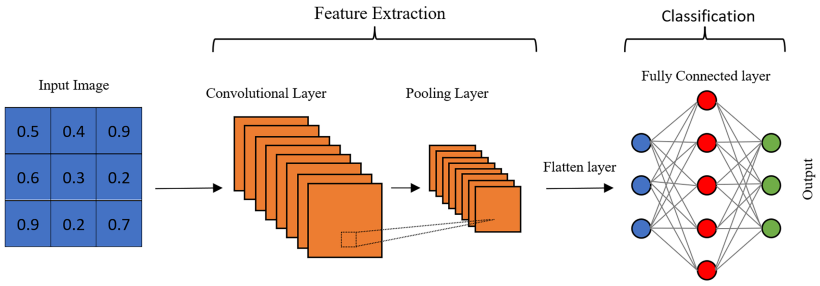


Figura 8: Versión resumida de la arquitectura de una red convolucional.

(b) (2pt) Describa brevemente los componentes clave como capas convolucionales, capas de pooling, capas completamente conectadas y el propósito de cada una. Además, explique cómo el tamaño de la imagen de entrada afecta a la arquitectura.

Respuesta:

i. Capas convolucionales: Estas capas son el núcleo de las redes convolucionales. Utilizan un conjunto de filtros aprendibles que se deslizan sobre la imagen de entrada para detectar características como bordes, texturas y patrones complejos. Cada filtro produce un mapa de características que resalta la presencia de dichas características en la imagen. **(0.5 pts)**

ii. Capas de Pooling: Su función es reducir la dimensionalidad espacial de los mapas de características, disminuyendo así el número de parámetros y la cantidad de cómputo requerido en la red. Esto se hace usualmente a través de operaciones como el max pooling, que toma el valor máximo dentro de una ventana deslizante sobre el mapa de características.**(0.5 pts)**

iii. Capas completamente conectadas: Estas capas, ubicadas generalmente hacia el final de la red, tienen la función de clasificar las características detectadas en categorías específicas.**(0.5 pts)**

iv. Tamaño de las imágenes: Imágenes más grandes pueden aumentar la cantidad de parámetros y el cómputo necesario, especialmente en las primeras capas convolucionales. **(0.5 pts)**

(c) (1pt) Explique con un ejemplo el concepto fundamental de convolución en una CNN.

Respuesta: La convolución se realiza a traves de matrices pequeñas que operan sobre la imagen punto a punto, multiplicando los valores de la matriz filtro y sumando esos valores, por ejemplo:

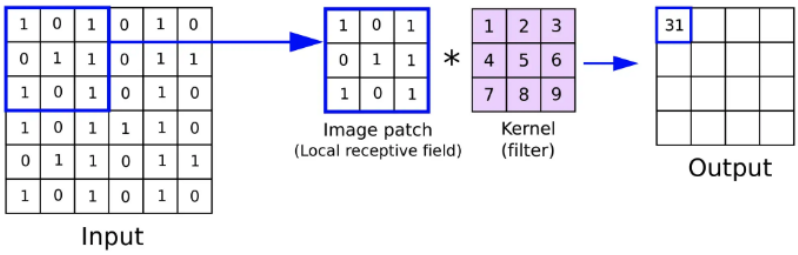


Figura 9: Filtro 3×3 con valores entre 1 y 9 operando sobre una matriz binaria.

(d) (1pt) ¿Por qué sirven para trabajar con audio?

Respuesta: Pues el audio puede representarse a través de imágenes en escalas de grises (espectrogramas o mel-spectrogramas), logrando que el modelo aprenda representaciones visuales del audio. **(1 pt)**