Tarea 1 EL4106 MLP

Joaquín Zepeda

October 1, 2021

1 Parte Teórica

Responda las siguientes preguntas en no mas de uno o dos párrafos por pregunta.

1. ¿Cuál es la ventaja del perceptrón multicapa respecto a la regresión logística? ¿Cuál es la importancia de contar con capas ocultas en el modelo? ¿Qué diferencias tienen las fronteras de decisión de dichos modelos?

La ventaja del perceptrón multicapa es que puede generalizar cualquier función, en cambio la regresión logística presenta limitaciones para predecir cuando las relaciones entre las variables son muy complejas, en efecto esta ultima solo puede resolver problemas linealmente separables, lo cual en muchos casos es bastante útil, pero no siempre se tienen modelos que cumplen esta condición. La capa oculta es la que permite que el perceptron multicapa generalice cualquier función, estas capas permiten representar las características del modelo de mejor manera y así proporcionan grados de libertad a la red.
La regresión simple cuenta con un **hiperplano separador** como frontera de decisión, lo cual limita su capacidad como fue mencionado anteriormente, por lo que solo puede resolver problemas en modelos linealmente separables. Por otro lado el perceptrón multicapa cuenta con **regiones arbitrarias** como fronteras de decisión, lo cual permite adaptarse y aproximar cualquier tipo de función.

2. ¿Qué efecto tiene el numero de neuronas en la capa oculta del MLP sobre la capacidad del modelo? Explique el compromiso entre capacidad de un modelo y sobreajuste.

El número de neuronas en la capa oculta está directamente relacionado con la capacidad de la red (hasta cierto punto), a mayor es el número de neuronas ocultas mayor es la capacidad del modelo pero llega un punto donde si se aumenta demasiado el número de neuronas ocultas, existe el peligro de que la red se aprenda de memoria el conjunto de entrenamiento y pierda la capacidad de generalización buscada (overfitting o sobreajuste), si bien no hay un calculo para determinar el número óptimo de neuronas a tener en la capa oculta, pues varia mucho en cada modelo, es necesario hacer pruebas con distintos valores hasta encontrar un número óptimo, por lo que hay que seleccionar con cuidado este número y si es necesario podar la red en caso de que la capacidad de esta sea muy grande y ocurra este problema.

3. ¿Qué es la tasa de aprendizaje y cómo afecta el proceso de entrenamiento de una red neuronal? ¿Cómo se puede elegir la tasa de aprendizaje?

La tasa de aprendizaje es un hiper-parámetro de la red neuronal, esta regula que tanto se ajustan los pesos de la red en cada iteración o época. Mientras menor es la tasa de aprendizaje, menor es el "paso" que se hace cada vez que se ajustan los pesos y esta influye en que tan rápido o que tan lento converge el algoritmo. Generalmente se utilizan valores menores a uno como tasa de aprendizaje, pero esto no necesariamente tiene que ser así, se pueden utilizar valores muy grandes, pero esto puede provocar que nunca converja el algoritmo (o que se demore muchas iteraciones en converger), por otro lado se puede utilizar valores muy pequeños pero al igual que los valores grandes puede que con estos no converjan (o se demore mucho en converger) pues se cambia de forma muy mínima cada peso en las iteraciones o épocas. Para elegir una tasa de aprendizaje, generalmente se comienza con una tasa

de aprendizaje menor a 1 (0.1) y se registra su rendimiento, luego se van aumentando los valores de la tasa de aprendizaje analizando el rendimiento, tanto la rapidez de convergencia como la estabilidad y el Accuracy que más se adapte al problema para luego quedarnos con el mejor valor de esta tasa. Lamentablemente esta forma de elegir la tasa de aprendizaje no siempre es la mejor, pues requiere alto costo computacional y de tiempo en cada experimento, pero actualmente existen algoritmos los cuales pueden ayudar a resolver el problema de que tasa de aprendizaje elegir con un costo computacional menor.

4. ¿Qué es un mini-batch y para que sirve? ¿Cuál es la diferencia entre iteración y época?

El mini-batch es un tipo de aprendizaje el cual consiste en realizar el ajuste a los pesos considerando pequeños conjuntos o mini-batches los cuales tienen un tamaño de una potencia de dos, el tamaño del batch en los mini batches se considera un nuevo hiper-parámetro y esta metodología es realmente útil cuando se tienen grandes bases de entrenamiento. Este se considera un punto intermedio entre el aprendizaje estocástico (por cada ejemplo se ajustan los pesos) y los batches (se usa todo el conjunto de ejemplos para ajustar los pesos). Por otro lado, una iteración es cuando se realiza una vez el algoritmo, cada vez que se realiza el algoritmo se considera una iteración más, en cambio una época es cuando se realiza cierto número de iteraciones, es decir, una época podría definirse como 10.

5. Explique los conceptos de accuracy, precision, recall y F1 score.

- Accuracy: mide la exactitud del sistema, este corresponde a la tasa de clasificaciones correctas.
- Precision: corresponde a la tasa de observaciones correctamente predichas como verdaderas (verdaderas positivas) con respecto al total de observaciones predichas verdaderas (las cuales considera las falsas positivas y las verdaderas positivas). Un valor de precision cercano a uno nos indica que el modelo es bueno, pues cuenta con un bajo porcentaje de falsos positivos.
- Recall: corresponde a la sensibilidad, es decir a la tasa de verdaderos positivos.
- F1 score: combina los estadísticos de la precision y el recall en uno solo, utilizando la media armónica de ambos, nos permite comparar el rendimiento combinado de los estadísticos recién mencionados.

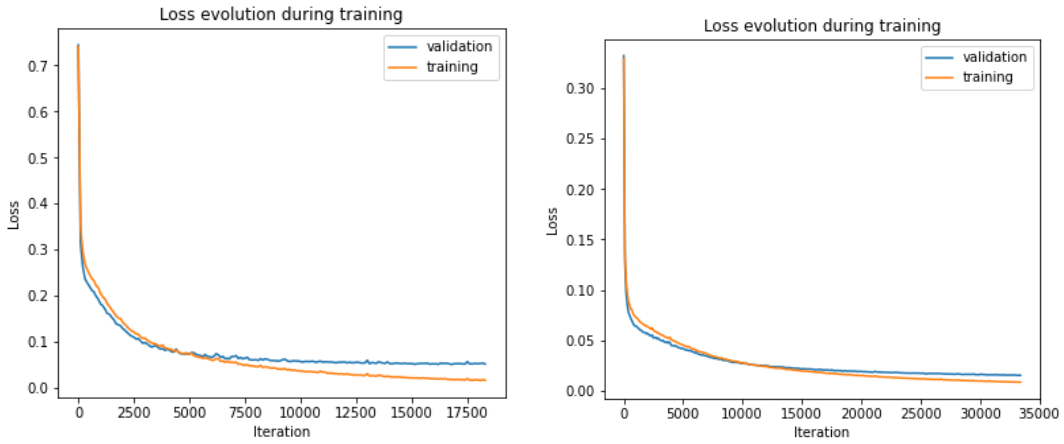
2 Parte Práctica

2.1 Pregunta 1

Una diferencia apreciable entre ambos casos es que en el caso 1 (xentropy) se necesita un número bastante menor de iteraciones para converger en comparación al caso 2 (MSE), el cual necesita casi el doble de iteraciones para converger, esto se puede apreciar en las curvas de aprendizaje en la figura 1, pues en la imagen de la izquierda de esta figura, la cual corresponde al entrenamiento utilizando la entropía cruzada, se puede observar que se realizaron aproximadamente 18000 iteraciones, por otro lado . Esto concuerda con los fundamentos teóricos, puesto que el método usando el error cuadrático medio es más lento que el método usando la entropía cruzada, esto debido al problema de desaparición del gradiente. Según los resultados obtenidos, los cuales se pueden observar en la tabla 1, el funcional más adecuado para entrenar la red es la entropía cruzada, puesto que se obtuvieron mayores tasas de clasificaciones correctas (Accuracy) tanto en entrenamiento y validación, además de esto se requieren menos recursos puesto que esta entrena el modelo en menos iteraciones, por otro lado analizando la matriz de confusión de ambos modelos, la cual corresponde a la figura 2, es posible apreciar que se obtienen mejores resultados utilizando la entropía cruzada como función de perdidas, pues se observan menores valores de falsos positivos y falsos negativos.

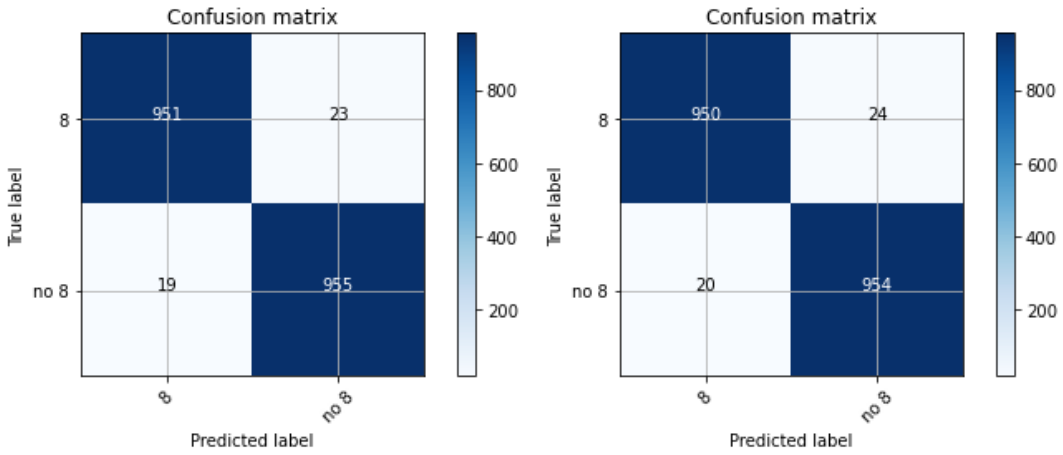
	Función de perdidas	Train Accuracy	Validation Accuracy
Experimento 1	Entropía cruzada	0.997 +/- 0.001	0.982 +/- 0.001
Experimento 2	Error cuadrático medio	0.993 +/- 0.000	0.979 +/- 0.001

Table 1: Tabla resultados accuracy pregunta 1



(a) Entropía cruzada como función de perdidas (b) Error cuadrático medio como función de perdidas

Figure 1: Curvas de aprendizaje



(a) Entropía cruzada como función de perdidas (b) Error cuadrático medio como función de perdidas

Figure 2: Matrices de confusión

2.2 Pregunta 2

La tasa de error al final del entrenamiento, como se puede observar en la figura 3 tiende a cero en los 4 casos, pero cada una converge con diferentes velocidades, diferentes números de iteraciones. En particular en el caso de la tasa de aprendizaje igual a 10, la tasa de error va cambiando de forma bastante irregular e inestable debido a su alto valor, en el caso del modelo utilizando la tasa de aprendizaje igual a 1, la tasa de error de este va convergiendo a cero durante el entrenamiento pero se encuentran puntos durante el entrenamiento donde se aprecian pequeños aumentos de la tasa de error, estos pequeños sobrepasos no son realmente significativos para este experimento pues logran llegar con muy pocas iteraciones a tasas de clasificaciones correctas muy altas, con aproximadamente 4 veces menos de iteraciones que el modelo usando tasa de 0.1, logra una Accuracy de entrenamiento y validación muy alto solo comparable con esta tasa recién mencionada, pero en su entrenamiento se usan menos recursos por lo que este modelo sería el mejor para este experimento, todo esto se puede observar en la tabla 2 y en la figura 3.

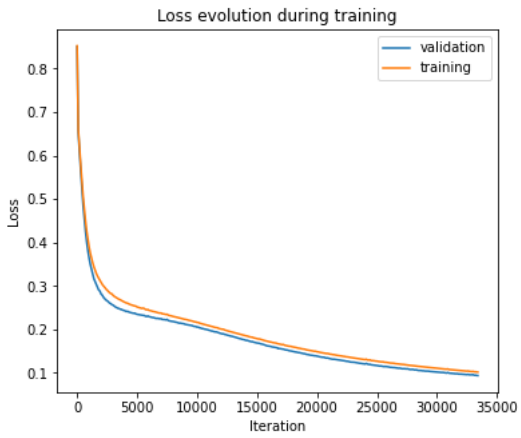
Por otro lado, si analizamos las matrices de confusión las cuales se puede observar en la figura 4, los errores correspondientes a falsos positivos y falsos negativos son mayores para las tasas de aprendizaje de 0.01 y 10, lo cual nos indica que estos modelos no son tan buenos en comparación con los modelos de las tasas de aprendizaje de 0.1 y 1, esto se debe a que se usaron valores muy pequeños en el caso del 0.01 y muy grandes para el caso de la tasa de aprendizaje de 10.

Como se puede apreciar, para las tasa más grandes la estabilidad se va perdiendo, pues no converge uniformemente a cero como lo hacen las tasas menores o iguales a 1, estas son inestables y la tasa de error en algunas iteraciones aumenta en gran manera.

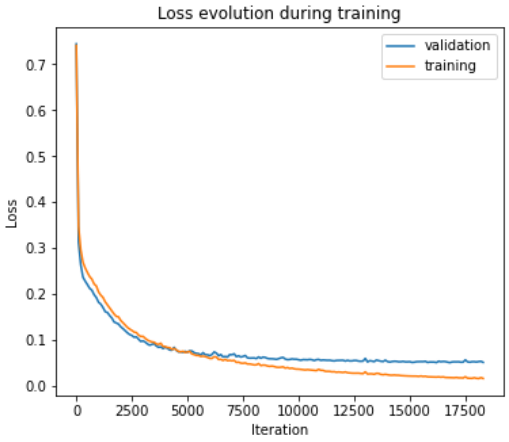
Las iteraciones que necesita cada modelo para converger, se encuentran en la tabla 2, como era de esperarse según los fundamentos teóricos, cuando las tasas son muy pequeñas se convierte en un proceso muy lento, por otro lado a mayor tasa se converge más rápido pero hay un umbral, cuando la tasa es muy grande, como lo es la tasa de valor igual a 10, en vez de converger más rápido necesita más iteraciones que las 2 tasas anteriores (1 y 0.1) puesto que tiene mucha inestabilidad. Además de esto, esta inestabilidad provoca una disminución del porcentaje de Accuracy siendo la menor tasa de Accuracy de los 4 modelos realizados, puesto que no va descendiendo las tasa de errores constantemente, los valores de accuracy se pueden observar en la tabla 2.

Tasa de aprendizaje	Train accuracy	Validation accuracy	N° de iteraciones aprox
0.01	0.967 +/- 0.001	0.964 +/- 0.001	35000
0.1	0.997 +/- 0.001	0.982 +/- 0.001	18000
1	0.998 +/- 0.001	0.981 +/- 0.001	4000
10	0.848 +/- 0.174	0.850 +/- 0.176	8000

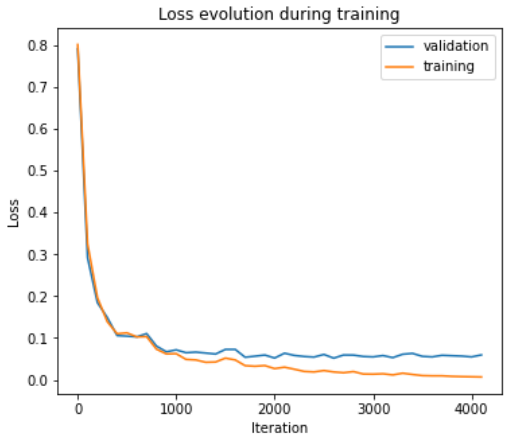
Table 2: Tabla resultados pregunta 2



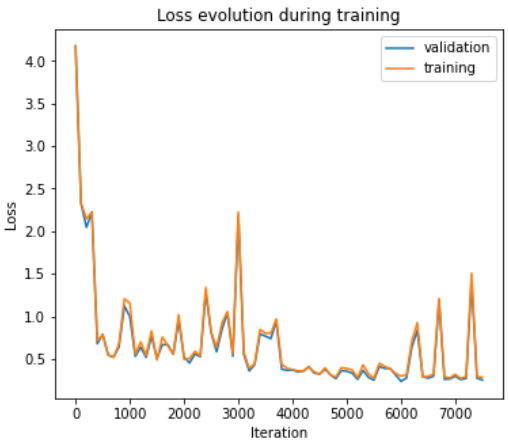
(a) Tasa de aprendizaje de 0.01



(b) Tasa de aprendizaje de 0.1

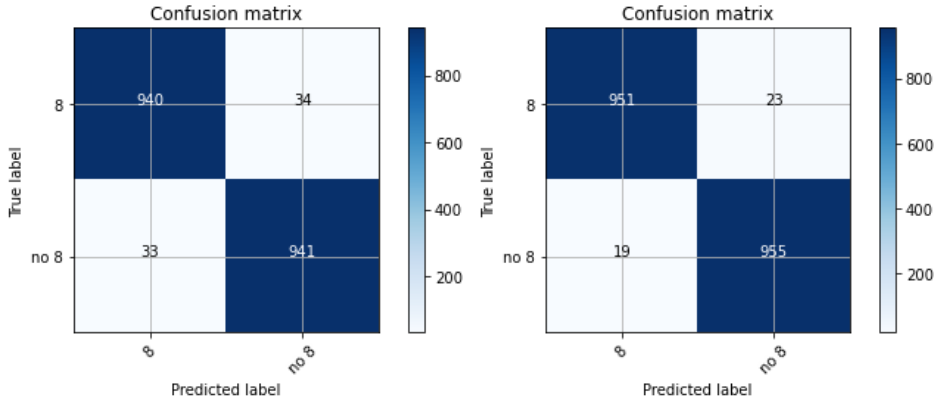


(c) Tasa de aprendizaje de 1



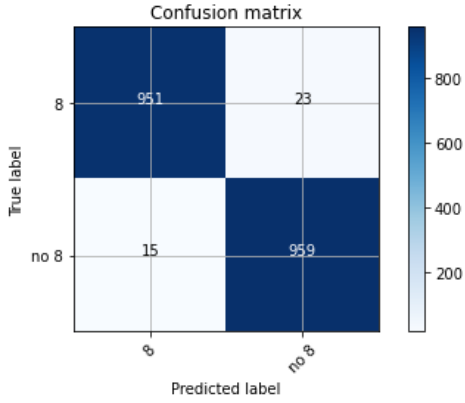
(d) Tasa de aprendizaje de 10

Figure 3: Curvas de aprendizaje para cada tasa de aprendizaje

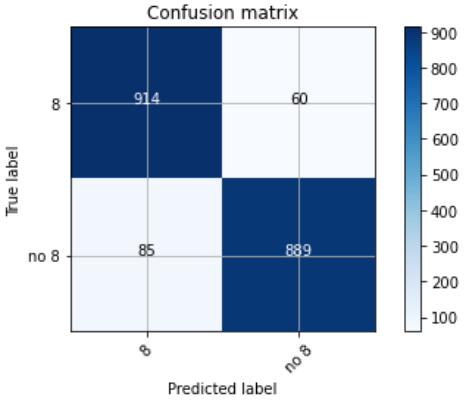


(a) Tasa de aprendizaje de 0.01

(b) Tasa de aprendizaje de 0.1



(c) Tasa de aprendizaje de 1



(d) Tasa de aprendizaje de 10

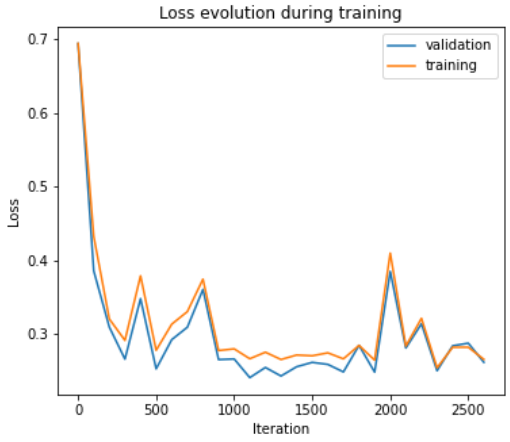
Figure 4: Matrices de confusión para cada tasa de aprendizaje

2.3 Pregunta 3

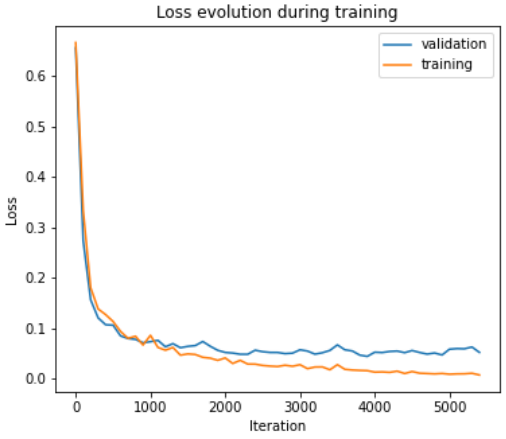
La mejor tasa de aprendizaje encontrada de la pregunta 2 corresponde a la tasa igual a 1, esto pues tiene los mejores resultados de porcentajes de acierto y tiene una gran rapidez, pues necesita aproximadamente 4 veces menos iteraciones para converger que la segunda mejor tasa en porcentajes de acierto, además de esto converge prácticamente sin irregularidades a cero por lo que es lo suficientemente estable.
Experimentalmente se determina que el número óptimo de N según el conjunto de validación corresponde a $N = 25$, pues es el que tiene el mayor porcentaje de Accuracy, esto se puede observar en la tabla 3.

Número de neuronas en la capa oculta	Train accuracy	Validation accuracy	Validation accuracy %
1	0.914 +/- 0.005	0.912 +/- 0.005	91.2 +/- 0.5
10	0.996 +/- 0.003	0.976 +/- 0.006	97.6 +/- 0.6
25	0.998 +/- 0.001	0.981 +/- 0.001	98.1 +/- 0.1
100	0.998 +/- 0.002	0.978 +/- 0.004	97.8 +/- 0.4

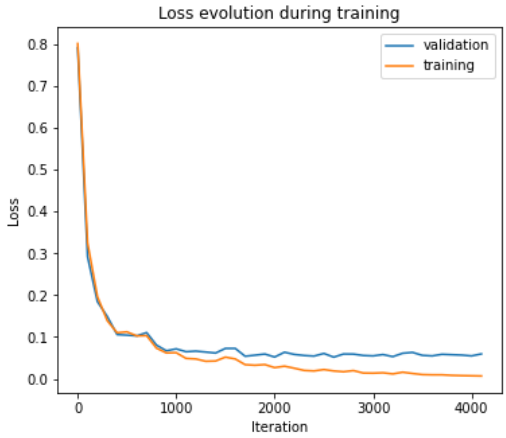
Table 3: Tabla resultados pregunta 3



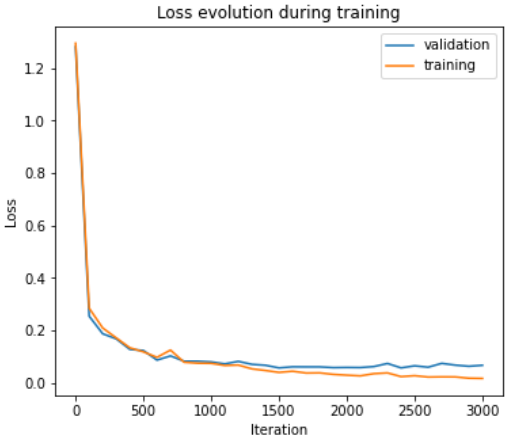
(a) 1 neurona en la capa oculta



(b) 10 neurona en la capa oculta

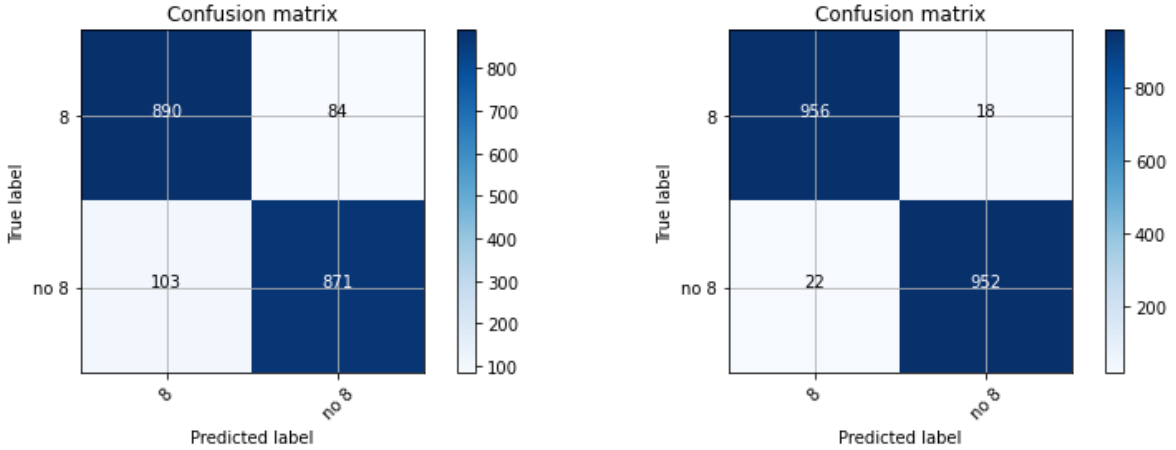


(c) 25 neurona en la capa oculta



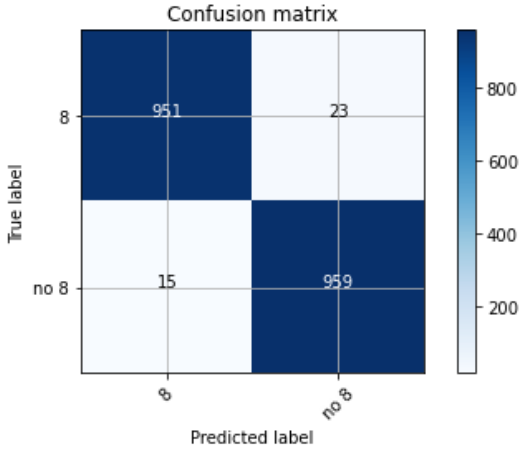
(d) 100 neurona en la capa oculta

Figure 5: Curvas de aprendizaje para cada experimento usando distinto número de neuronas en la capa oculta

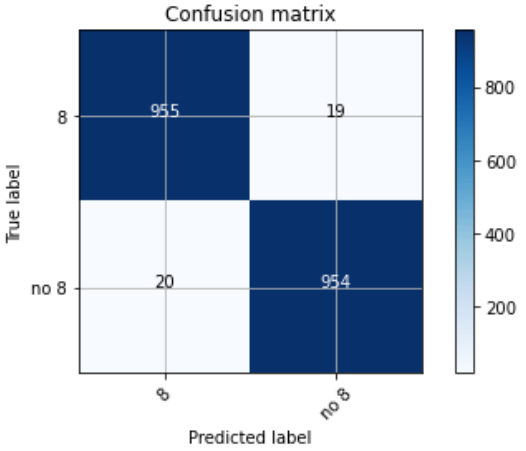


(a) 1 neurona en la capa oculta

(b) 10 neurona en la capa oculta



(c) 25 neurona en la capa oculta



(d) 100 neurona en la capa oculta

Figure 6: Matrices de confusión para cada experimento usando distinto número de neuronas en la capa oculta

Cambiando el valor de "early_stopping" a 1000, es posible observar el sobreajuste, esto se observa en los datos de la tabla 4, en donde se puede apreciar que los datos de entrenamiento tienen un 100% de accuracy, lo cual nos indica que la red se ha aprendido de memoria los datos de entrenamiento, lo cual afecta al rendimiento de la red para generalizar nuevos datos, la curva de aprendizaje se puede observar en la figura 7 en donde se puede observar que la tasa de error en el entrenamiento se va a cero.

Número de neuronas en la capa oculta	Train accuracy	Validation accuracy
25	100.0000%	98.0769%

Table 4: Tabla resultados pregunta 3 con early_stopping igual a 1000

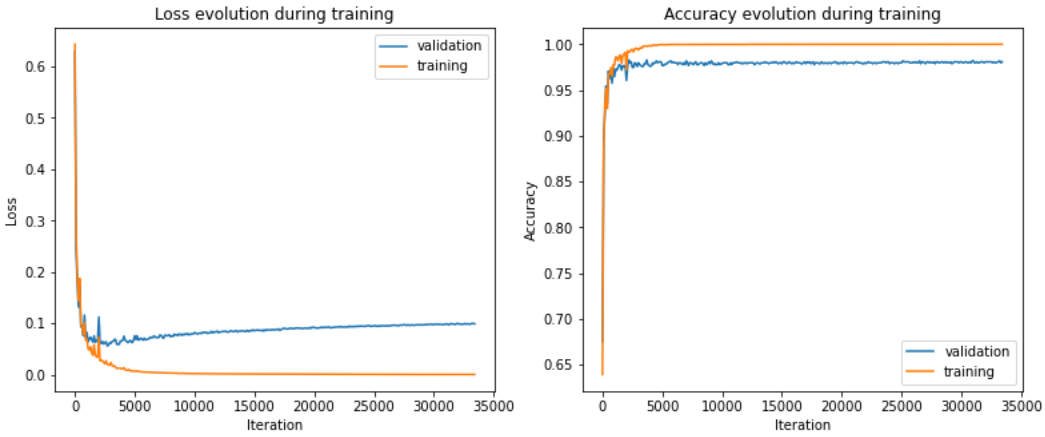


Figure 7: Tasa de aprendizaje y la evolución del Accuracy durante el entrenamiento

2.4 Pregunta 4

Indique cual es dicho valor de umbral y el punto asociado en la curva ROC, explicando como llego a esta eleccion. En este punto de operacion, ¿cual es la cantidad esperada de casos negativos que son erroneamente etiquetados por el modelo como positivos?
El mover el umbral tiene un impacto directo en los casos predichos como positivos, mientras más alto este umbral, menores son los valores predichos como verdaderos positivos y también disminuyen los valores predichos como falsos positivos, estos tienden a cero, por otro lado, el valor de los falsos negativos aumenta comparando con la situación de un umbral menor, esto se puede observar en la tabla 5.

El valor del umbral utilizado para que los casos predichos como positivos sean aproximadamente 400 fue un threshold o umbral de 0.999955, en efecto con este valor de umbral se obtuvieron 408 True positives (TP), lo cual es bastante cercano a 400, elegí este umbral usando haciendo pruebas, aumentando de a poco el valor del umbral pretendiendo encontrar un valor cercano a 400 en los valores predichos como positivos.
El punto asociado en la punta roc esta en la zona marcada por el circulo rojo de la figura 8, punto en donde se encuentra la menor tasa de falsos positivos y la mayor tasa de verdaderos positivos, esto se puede concluir puesto que con el nuevo umbral se obtienen 0 False Positives y 408 TP los cuales si bien no corresponden al todos los True positives por lo que la tasa no alcanza el 1.0, si se obtiene una gran cantidad de estos, por lo que el punto debería tender a estar dentro de la zona dicha anteriormente, pero no en el punto más alto, pues ese punto se alcanza cuando se tienen todos los valores de True positives.

Umbral	True positives (TP)	True Negatives (TN)	False Positives (FP)	False Negatives (FN)
0.5	483	483	11	11
0.999955	408	494	0	86

Table 5: Tabla pregunta 4

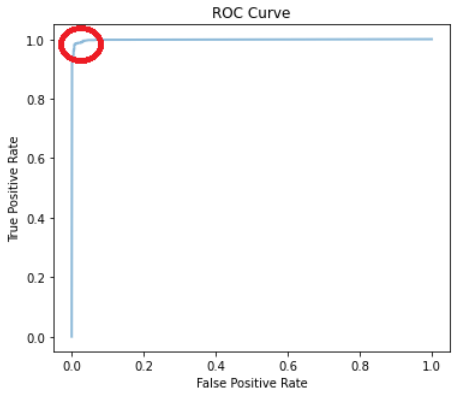


Figure 8: Curva ROC