Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Eléctrica

Tarea 1

MLP

Integrantes: Vincko Fabres
Profesor: Pablo Estevez.
Auxiliar: Ignacio Reyes Jainaga
Ayudantes: Andrés González
Bastián Andreas
Daniel Baeza
Francisca Cona
Javier Molina
Óscar Pimentel
Pablo Montero
Roberto Cholaky

Fecha de entrega: Sábado 2 de Octubre de 2021
Santiago, Chile

1. Parte teórica

1.1. ¿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?

El perceptrón al poseer neuronas le da una mayor flexibilidad a su frontera de decisión en comparación a la regresión logística, el número de capas ajusta la red de mejor manera que la utilización de un sólo perceptrón pudiendo llegar a generar fronteras de decisión más complejas, en el caso de la regresión logística genera un hiperplano separador, el MLP en cambio, puede generar zonas convexas o aún más complejas.

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

La capacidad de predicción del modelo está totalmente relacionada, a mayor número el aprendizaje es mejor, llegando a un límite en el cual la red de MLP memoriza los ejemplos llegando a un sobreajuste.

1.3. ¿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 corresponde al ponderador con el cual se avanza en la retropropagación del error para ajustar los pesos de la red, de este hiperparámetro depende mucho el desempeño de la red, ya que una tasa pequeña puede conducir a un minimo global de la funcion de costos, pero uno muy grande hará que oscile sin lograr realmente aprendizaje, es decir, controla la estabilidad y tasa de convergencia.
Para la elección de este parámetro existen en primera instancia una cota, $0 < \mu < \frac{1}{\lambda_{max}}$ para el método del gradiente, existe el método de momento adaptivo y también el método Delta-Barra-Delta.

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

Un mini batch corresponde a un subconjunto de muestra del conjunto de entrenamiento, el que sirve para entrenar la red para agilizar su entrenamiento.
Una iteración corresponde a un ciclo de entrenamiento y retropropagación de errores, mientras que una época, contiene un conjunto de iteraciones, las cuales se utilizan como referencia para mostrar los sobreajustes.

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

Cada una tiene una fórmula diferente, siendo accuracy o exactitud el porcentaje de clasificaciones correctas, mientras que precision apunta al número total de casos que fueron detectados cómo clase positiva, recall corresponde a la sensibilidad del clasificador; entregando la tasa de verdaderos positivos y F1 score corresponde a la combinación de precision y recall, entregando la media armónica.

2. Parte práctica

2.1.

Los resultados al utilizar la función de costo Entropía cruzada 5 veces son los siguientes:

```
Validation accuracy 0.976 +/- 0.002
[0.9777328  0.9757085  0.97469634 0.9787449  0.9716599 ]
Train accuracy 0.995 +/- 0.002
[0.99757326 0.99150646 0.99626654 0.9960799  0.9956132 ]
```

Figura 1: Accuracy con desviación estándar

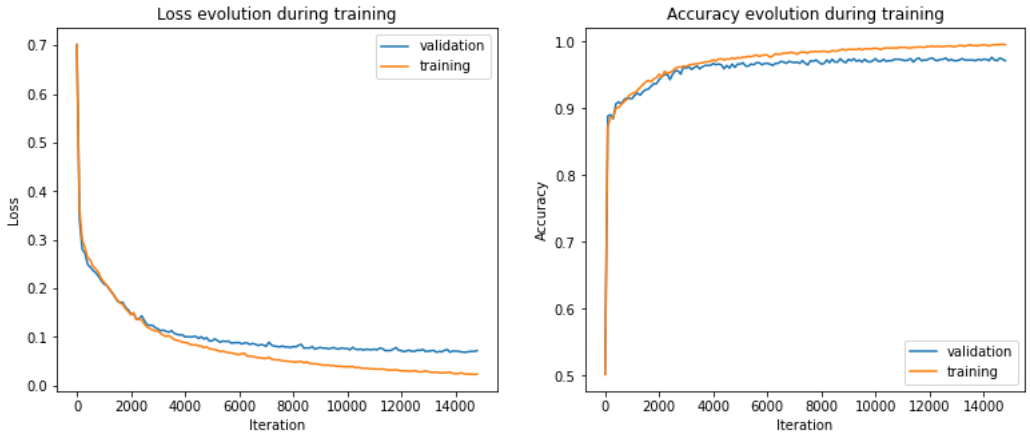


Figura 2: Curvas loss y accuracy

```
Confusion matrix, without normalization
[[956  18]
 [ 19 955]]
Training results:
TP: 5342, TN: 5325, FP: 32, FN: 15
99.5613% Accuracy (Porcentaje de clasificaciones correctas)
99.4045% Precision
99.7200% Recall

Validation results:
TP: 484, TN: 476, FP: 18, FN: 10
97.1660% Accuracy (Porcentaje de clasificaciones correctas)
96.4143% Precision
97.9757% Recall

Test results:
TP: 955, TN: 956, FP: 18, FN: 19
98.1006% Accuracy (Porcentaje de clasificaciones correctas)
98.1501% Precision
98.0493% Recall
```

Figura 3: Resultados train, validation y test

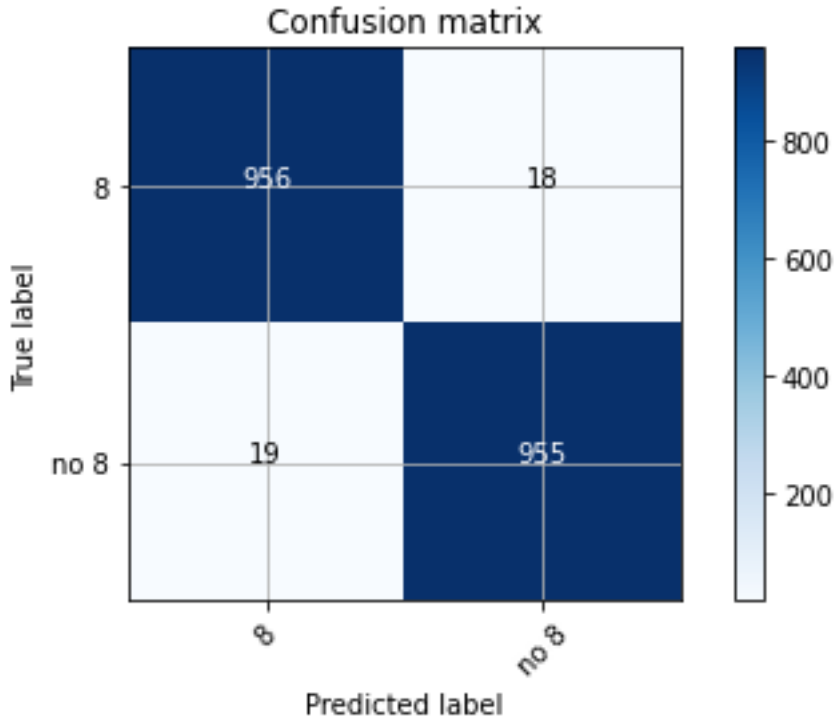


Figura 4: Matriz de confusión

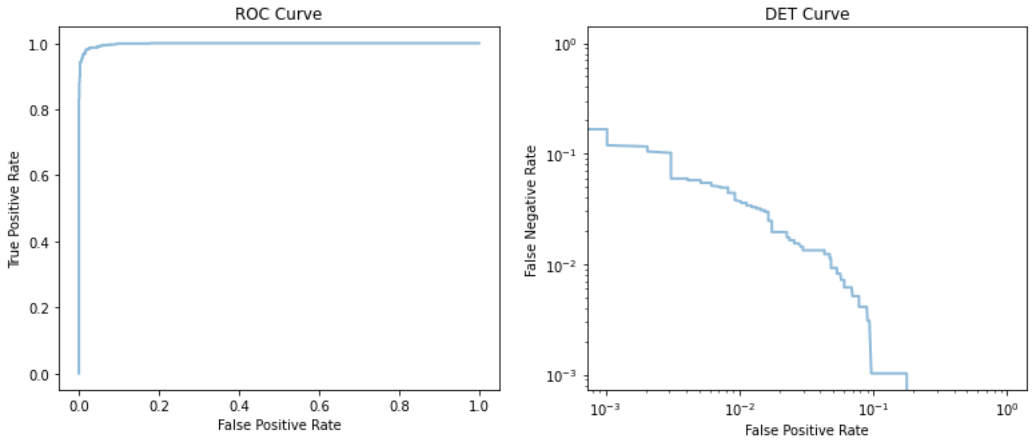


Figura 5: Curvas ROC y DET

Al realizar el mismo experimento con loss Mean Square Error se obtiene:

```
Validation accuracy 0.975 +/- 0.002
[0.9736842  0.9787449  0.9757085  0.9716599  0.97672063]
Train accuracy 0.994 +/- 0.001
[0.9937465  0.9944932  0.99383986 0.99281317 0.99337316]
```

Figura 6: Accuracy con desviación estándar

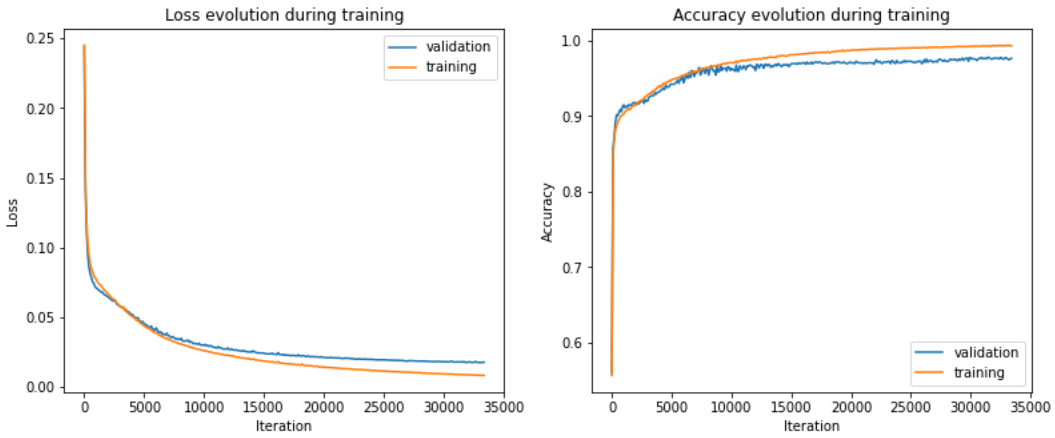


Figura 7: Curvas loss y accuracy

```
Confusion matrix, without normalization
[[956  18]
 [ 25 949]]
Training results:
TP: 5319, TN: 5324, FP: 33, FN: 38
99.3373% Accuracy (Porcentaje de clasificaciones correctas)
99.3834% Precision
99.2906% Recall

Validation results:
TP: 484, TN: 481, FP: 13, FN: 10
97.6721% Accuracy (Porcentaje de clasificaciones correctas)
97.3843% Precision
97.9757% Recall

Test results:
TP: 949, TN: 956, FP: 18, FN: 25
97.7926% Accuracy (Porcentaje de clasificaciones correctas)
98.1386% Precision
97.4333% Recall
```

Figura 8: Resultados train, validation y test

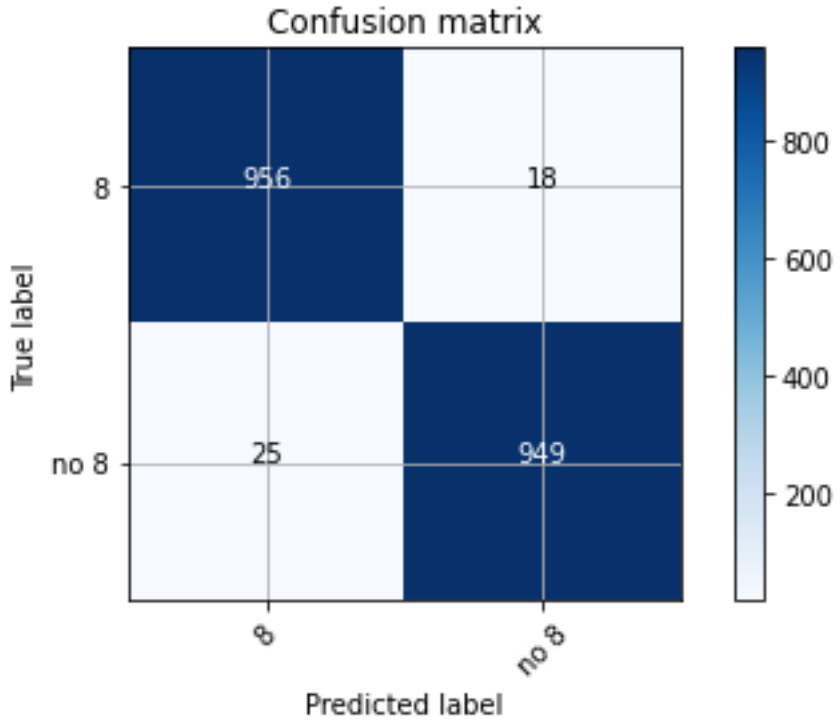


Figura 9: Matriz de confusión

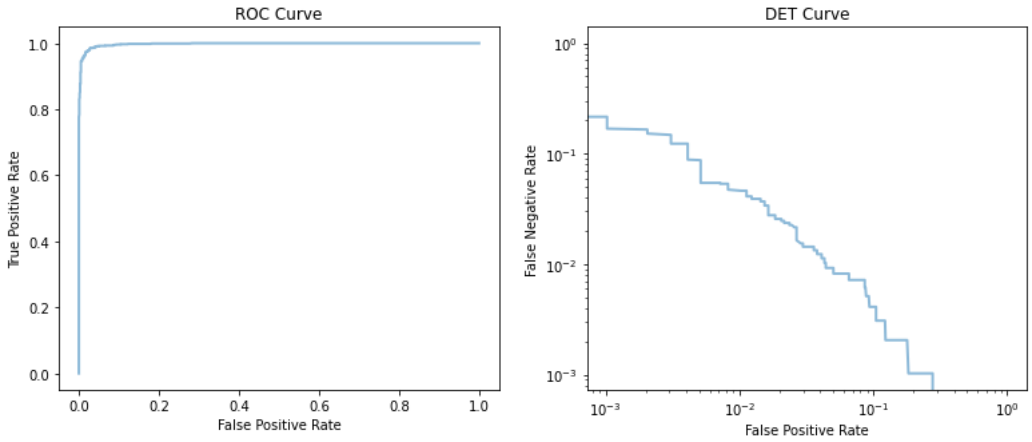


Figura 10: Curvas ROC y DET

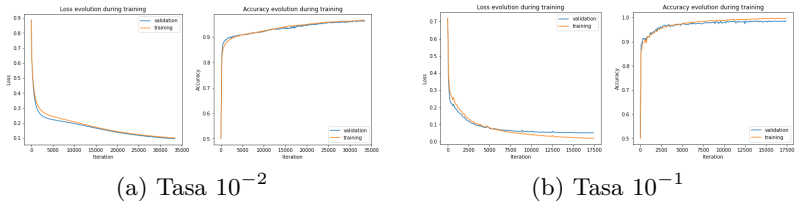
Dados los resultados es posible apreciar una leve diferencia, siendo mejor la utilización de entropía cruzada para el entrenamiento de la red.
Para apreciar el desempeño total basta con ver los accuracy, en el cual los rendimientos de ambos funcionales se ven casi igualados, con una diferencia ínfima de performance, razón por la cual se opta revisar el rendimiento en las matrices de confusión, donde es posible apreciar el porqué cross entropy es mejor configuración; se debe a la detección de verdaderos negativos. Al entender el funcionamiento de ambas funciones de error es plausible encontrar el fundamento teórico. El funcionamiento de la

entropía cruzada es entregar información de cuán cercana es la predicción, por lo que resultados cercanos afectan poco mientras que para grandes diferencias el error aumenta, lo cual beneficia este entrenamiento donde el resultado es categórico.

2.2.

Los resultados de cada tasa son:

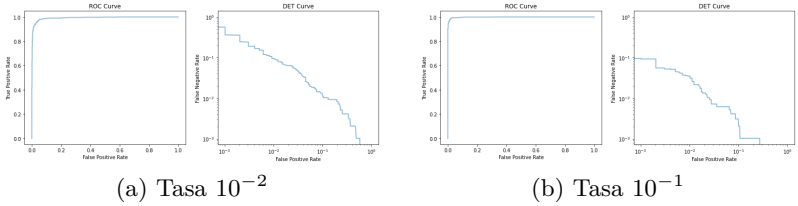
Tasa de aprendizaje	Validation accuracy	Train accuracy
10^{-2}	0.962 +/- 0.002	0.967 +/- 0.001
10^{-1}	0.979 +/- 0.005	0.994 +/- 0.003
10^{0}	0.983 +/- 0.002	0.999 +/- 0.001
10^{1}	0.668 +/- 0.206	0.665 +/- 0.203



(a) Tasa 10^{-2} (b) Tasa 10^{-1}

(c) Tasa 10^{0} (d) Tasa 10^{1}

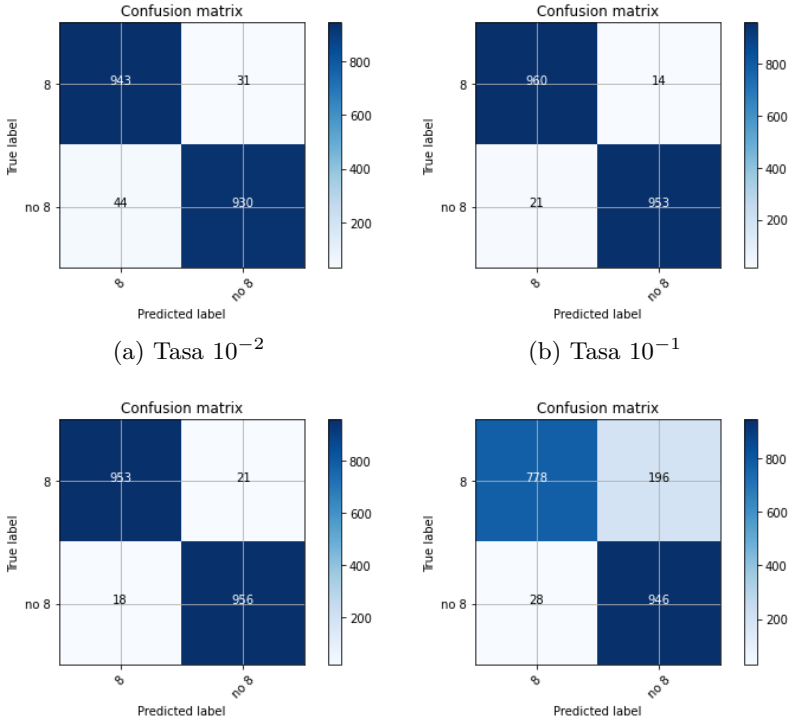
Figura 11: Curvas de aprendizaje



(a) Tasa 10^{-2} (b) Tasa 10^{-1}

(c) Tasa 10^{0} (d) Tasa 10^{1}

Figura 12: Tasas de acierto



(a) Tasa 10^{-2} (b) Tasa 10^{-1}

(c) Tasa 10^{0} (d) Tasa 10^{1}

Figura 13: Matrices de confusión

Dados los resultados de las curvas de loss es posible apreciar que la tasa de aprendizaje afecta a este, ya que si este es muy pequeño las diferencias son mínimas lo que reduce el aprendizaje con ruido en la curva, por otra parte, si la tasa es muy grande la estabilidad de la red se ve afectada con variaciones abruptas. Las iteraciones necesarias para cada tasa son las siguientes:

Tasa de aprendizaje	Iteraciones
10^{-2}	35000
10^{-1}	17500
10^{0}	4000
10^{1}	8000

2.3.

La mejor tasa de aprendizaje dados los resultados de las curvas ROC y DET, junto con la matriz de confusión dan por ganador a la tasa de 10^{-1}.

Los resultados por número de neuronas son:

Numero de neuronas	Validation accuracy	Train accuracy
1	0.912 +/- 0.002	0.917 +/- 0.005
10	0.972 +/- 0.003	0.991 +/- 0.002
25	0.982 +/- 0.001	0.997 +/- 0.001
100	0.978 +/- 0.002	0.998 +/- 0.001

Para el conjunto de validación los resultados son los siguientes:

Numero de neuronas	Accuracy	Precision	Recall
1	90.7895 %	95.0783 %	86.0324 %
10	97.3684 %	97.3684 %	97.3684 %
25	97.9757 %	98.3673 %	97.5709 %
100	97.3684 %	96.8000 %	97.9757 %

Por lo que el número óptimo de neuronas es 25.

Una vez modificados el parámetro early stopping se obtiene accuracy 1 para conjunto de entrenamiento y 0.984 para validaión, con la siguiente curva de aprendizaje:

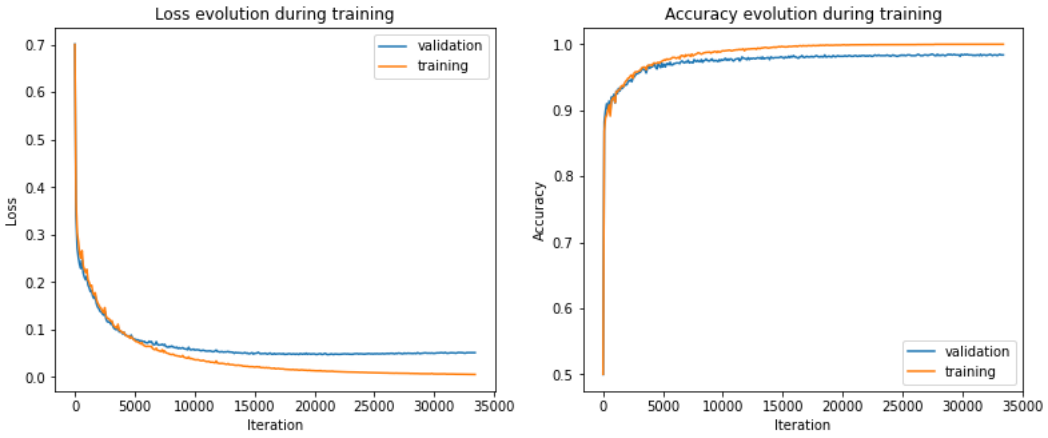


Figura 14: Curva de aprendizaje para N óptimo y earky stopping = 1000

2.4.

Al mover el umbral el clasificador las predicciones se sesgan hacia una clase, siendo el umbral 0 orientado totalmente a la clase positiva y por otra parte, un umbral de 1 encasilla todos los inputs como clase negativa.
El umbral que experimentalmente cataloga según lo solicitado posee un valor de $treshold = 0.994$, este punto al ser asociado a la curva ROC corresponde a la intersección del eje y, True Positive Rate = 0.8 y la curva, esta elección toma como base que la distribución de los datos es con las clases balanceadas, por lo que de 1000 datos 500 son de la clase positiva y se quieren identificar 400 correctamente, por lo cual la tasa de aciertos positivos corresponde al 80 % de verdaderos positivos. En

este caso la cantidad esperada de casos negativos errónamente etiquetados o falsos negativos, dado que es el complemento corresponde al 20 %, es decir, 100 casos.