Tarea 1

Septiembre 2022
Entrega: 5 de octubre 2022

Profesor: Ángel Jiménez M.
Auxiliar: Rafael De la Sotta V.
Ayudantes: Gabriela Mora M., Patricio Ortiz V., Camila Pulgar F., Maximiliano Rosadio Z., Sebastián Urbina G.

P1.- (3 ptos.) **Backpropagation del error cuadrático medio** [1]**.** Considere una red neuronal feedforward de una capa oculta y una capa de salida, con datos de entrenamiento $T = \{(x_1, y_1), (x_2, y_2)\} = \{(0.05, 0.01), (0.1, 0.99)\}$, dos neuronas en la capa oculta y variable de salida $\widehat{Y} = (\widehat{y}_1, \widehat{y}_2)$. Sea $W^{(1)}$ la matriz de parámetros de la capa oculta y $w^{(1)}_{j,i}$ el parámetro de la conexión entre la entrada x_i y la neurona j de la capa oculta, tal que $w^{(1)}_{1,1} = 0.15$, $w^{(1)}_{1,2} = 0.2$, $w^{(1)}_{2,1} = 0.25$ y $w^{(1)}_{2,2} = 0.3$, mientras que $w^{(1)}_{j,0}$ corresponde al parámetro de sesgo de la neurona j, tal que $w^{(1)}_{1,0} = w^{(1)}_{2,0} = 0.35$. Por otro lado sea U la matriz de parámetros de la capa de salida y $u_{k,j}$ el parámetro de la conexión entre la la neurona j de la capa oculta y la componente k de la variable de salida, tal que $u_{1,1} = 0.4$, $u_{1,2} = 0.45$, $u_{2,1} = 0.5$ y $u_{2,2} = 0.55$, mientras que $w_{k,s}$ corresponde al parámetro de sesgo de la componente k de la variable de salida, tal que $w_{1,s} = w_{2,s} = 0.6$. Tanto la capa oculta como la capa de salida se activan con una función sigmoid $\sigma(a) = \frac{1}{1+e^{-a}}$. Considere como función de pérdida del entrenamiento al error cuadrático medio dado por:

$$L = \frac{1}{2} \sum_{k=1}^{2} (y_k - \widehat{y}_k)^2 \tag{1}$$

1. (0.5 ptos.) Dibuje un grafo de computación para la red neuronal feedforward. Utilice el algoritmo forward propagation para computar manualmente los valores $\widehat{y}_1$, $\widehat{y}_2$ y de la pérdida L.

2. (2 ptos.) Utilice el algoritmo de backpropagation para computar manualmente el gradiente de L con respecto a cada uno de los parámetros de la red, es decir, $\frac{\partial L}{\partial w^{(1)}_{1,1}}$, $\frac{\partial L}{\partial w^{(1)}_{1,2}}$, $\frac{\partial L}{\partial w^{(1)}_{2,1}}$, $\frac{\partial L}{\partial w^{(1)}_{2,2}}$, $\frac{\partial L}{\partial u_{1,1}}$, $\frac{\partial L}{\partial u_{1,2}}$, $\frac{\partial L}{\partial u_{2,1}}$ y $\frac{\partial L}{\partial u_{2,2}}$.

3. (0.5 ptos.) Considere una tasa de aprendizaje $\eta = 0.5$ y actualice todos los parámetros de la red de acuerdo al algoritmo del descenso del gradiente, luego compute nuevamente los valores $\widehat{y}_1$, $\widehat{y}_2$ y de la función de pérdida L. Concluya.

[1] Esta pregunta es larga y mecánica, pero es algo que hay que hacer por lo menos una vez en la vida; se recomienda partir con anticipación.

P2.- (3 ptos.) **Análisis de riesgo de crédito.** En este problema utilizaremos una base de datos de riesgo crediticio, que contiene 3000 registros de variables categóricas y numéricas, para determinar si un cliente bancario es un mal pagador o no.

Considere el archivo *credito.xlsx* compuesto por las siguientes características:

- **ID:** número de cliente.
- **AgnosDirec:** años que el cliente lleva viviendo en la misma dirección.
- **AgnosEmpleo:** años de antigüedad del cliente en su actual empleo.
- **DeudaExt:** deuda del cliente externa al banco.
- **DeudaInt:** deuda del cliente en el banco. Por ejemplo, otros creditos.
- **Edad:** edad del cliente.
- **Ingreso/Ingreso2:** ingreso del cliente (recopilados de dos fuentes diferentes).
- **Nacionalidad:** nacionalidad del cliente.
- **NivelEdu:** nivel educacional del cliente.
- **VarObj:** variable objetivo. Si el cliente es un mal pagador ("cae en bancarrota") en caso de recibir el crédito. Si el cliente cae en bancarrota toma el valor de S y N en el caso contrario.

En base al problema descrito y sus variables, desarrolle los siguientes pasos para clasificar si un cliente es un mal pagador o no:

1. (1.0 pto.) Preprocese adecuadamente y realice un análisis exploratorio de los datos según lo aprendido en clases auxiliares, discutiendo sus hallazgos. Dentro de su análisis, tenga en consideración las siguientes sugerencias, pues posteriormente le permitirán obtener mejores desempeños en sus modelos:

    - Impute valores perdidos en los atributos numéricos cuando estos sean menores al 5 % del total de los datos, reemplazando por la mediana o la media.
    - Construya y agregue un atributo adicional a la base de datos que represente el endeudamiento real del cliente por medio de la expresión DeudaIng = (DeudaInt + DeudaExt)/Ingreso.
    - Para atributos con distribución asimétrica, es decir cuya media está muy cerca de los valores extremos, aplique la función logaritmo para expandir la distribución y asemejarla a una distribución normal (sumar 1 evita que se indefina la función en caso de haber ceros).
    - Transforme los atributos categóricos a variables binarias.
    - Descarte el o los atributos muy concentrados. Por ejemplo, si la desviación estándar de un atributo es cero, debe eliminarse.

- Visualice la matriz de correlaciones bivariadas para identificar atributos que estén muy correlacionados entre sí, dejando sólo uno (variables redundantes no aportan información nueva)

2. (1.0 pto.) Proponga una red neuronal feedforward que permita clasificar si un cliente bancario es mal pagador o no. Según lo aprendido en cátedras, ¿qué funciones de activación (en las capas ocultas y de salida), y qué función de pérdida propone utilizar y por qué? Discuta las **métricas de desempeño apropiadas** para este problema y justifique su selección.

3. (1.0 pto.) Implemente la red neuronal con diferentes optimizadores, utilizando fundadamente métodos de mejoramiento y regularización según lo visto en clases auxiliares. Entregue un diagrama del modelo final, aplique la red sobre los datos de test y compare los resultados obtenidos con distintas configuraciones.

Evaluación de la tarea

La tarea se realiza en grupos. Para la pregunta 1 se puede entregar la digitalización del manuscrito o su edición en algún procesador. Para la pregunta 2 se debe entregar el código reproducible en un cuaderno de Jupyter o en Google Colab, explicando de manera ordenada los pasos realizados. Debe incluir sus supuestos, justificaciones, procedimientos bien detallados y discusiones en el mismo cuaderno. Cite acórdemente si utiliza alguna fuente de información para realizar su trabajo.

Dado que el problema se puede abordar de diferentes maneras, se evaluará la implementación de modelos, utilidad del análisis estadístico, la coherencia del informe y las respuestas entregadas según los resultados obtenidos.

Recuerden que el foro siempre está abierto para preguntas.