

Práctica 8

Introducción a los Sistemas Inteligentes

Entrega: durante la clase

Descargue el conjunto de datos [titanic.csv](https://www.kaggle.com/competitions/titanic/data) y lea la descripción del mismo en <https://www.kaggle.com/competitions/titanic/data>. Este conjunto de datos será usado en los siguientes puntos, los cuales deben ser resueltos usando Pandas y TensorFlow.

1. Prepare los datos.
 - a) Cargue los datos usando Pandas.
 - b) Seleccione las siguientes características del dataset: Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked.
 - c) Imprima las primeras filas del dataframe e inspeccione los datos. Qué variables son categóricas y cuáles son numéricas?
 - d) Verifique cuántos datos faltantes hay usando las funciones isnull and sum de Pandas
 - e) Utilice la función fillna de Pandas para manejar valores faltantes, use la mediana o la moda según aplique.
 - f) Convierta las columnas con valores categóricos a valores numéricos usando la función get_dummies de Pandas.
 - g) Cree dos arreglos de numpy, X y y , para los datos de entrada y la variable de clase (*Survived*)
2. Entrene un modelo de clasificación basado en redes neuronales.
 - a) Haga una partición del conjunto de datos, usando muestreo estratificado, en 60 % para entrenamiento, 20 % de validación y 20 % para test.
 - b) Defina un modelo en Keras que tenga tantas entradas como columnas en X , una capa oculta con 10 neuronas y función de activación sigmoide y una neurona de salida con función de activación sigmoide. Imprima un resumen de la arquitectura del modelo. ¿Cuántos parámetros tiene?
 - c) Compile el modelo especificando un optimizador Adam y una pérdida Binary Cross Entropy.
 - d) Entrene el modelo usando diferentes valores de la tasa de aprendizaje (learning rate): $10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$, etc. Use 30 epochs. Grafique curvas de aprendizaje de la pérdida de entrenamiento y la pérdida en validación. Qué tasa de aprendizaje es mejor? Explique.
 - e) Aplique el mejor modelo al conjunto de test.
 - f) Mida el desempeño del modelo calculando exactitud, error de clasificación, precisión, recall, curva ROC y matriz de confusión.
3. Exploración de hiperparámetros.
 - a) Varíe el número de neuronas en la capa oculta de 2 a 20. Determine cuál es la mejor configuración. Explique.

- b) Pruebe diferentes funciones de activación para la capa oculta. Muestre y discuta los resultados.
- c) Adicione una capa oculta adicional y pruebe varias configuraciones. Muestre y discuta los resultados.
- d) Haga una comparación detallada con diferentes métricas de desempeño entre el modelo del anterior punto y el mejor modelo encontrado en este punto.

La práctica debe enviarse como un Jupyter notebook a través del siguiente Dropbox file request, antes del final de la clase. El archivo debe nombrarse como `isi-practica8-unalusername.ipynb`, donde `unalusername` es el nombre de usuario asignado por la universidad. No incluya archivos adicionales.