

Hoja de Trabajo 8.

Redes Neuronales Artificiales (RNA)

INTRODUCCIÓN:

Kaggle

Kaggle es una comunidad en línea de científicos de datos, propiedad de Google LLC. Permite a los usuarios encontrar y publicar conjuntos de datos, explorar y construir modelos en un entorno de ciencia de datos basado en la web, trabajar con otros científicos de datos e ingenieros de aprendizaje automático, y participar en competencias para resolver los desafíos de la ciencia de datos. Tuvo su inicio al ofrecer competencias de aprendizaje automático y ahora también ofrece una plataforma pública de datos, una mesa de trabajo basada en la nube para la ciencia de la información y educación en IA de formato corto. El 8 de marzo de 2017, Google anunció que estaban adquiriendo Kaggle.

Conjunto de datos a utilizar

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>

Notas:

- La hoja de trabajo se realizará en los mismos grupos de la hoja anterior.
- La nota de cada uno estará en correspondencia con los aportes que cada integrante del grupo haga a la hoja de trabajo.

INSTRUCCIONES

- Utilice el data set [House Prices: Advanced Regression Techniques](#) que comparte Kaggle. Debe hacer un análisis exploratorio para entender mejor los datos, sabiendo que el objetivo final es predecir los precios de las casas. Recuerde explicar bien cada uno de los hallazgos que haga. La forma más organizada de hacer un análisis exploratorio es generando ciertas preguntas de las líneas que le parece interesante investigar. Genere un informe en pdf con las explicaciones de los pasos que llevó a cabo y los resultados obtenidos. Recuerde que la investigación debe ser reproducible por lo que debe guardar el código que ha utilizado para resolver los ejercicios y/o cada uno de los pasos llevados a cabo si utiliza una herramienta visual.

ACTIVIDADES

1. Use los mismos conjuntos de entrenamiento y prueba que utilizó en las hojas anteriores.
2. Seleccione como variable respuesta la que creó con las categorías del precio de la casa.
3. Genere dos modelos de redes neuronales que sean capaz de clasificar usando la variable respuesta que categoriza las casas en baratas, medias y caras. Estos modelos deben tener diferentes topologías y funciones de activación.
4. Use los modelos para predecir el valor de la variable respuesta
5. Haga las matrices de confusión respectivas.

6. Compare los resultados obtenidos con los diferentes modelos de clasificación usando redes neuronales en cuanto a efectividad, tiempo de procesamiento y equivocaciones (donde el algoritmo se equivocó más, donde se equivocó menos y la importancia que tienen los errores).
7. Analice si no hay sobreajuste en los modelos. Use para esto la curva de aprendizaje.
8. Para el modelo elegido de clasificación tune los parámetros y discuta si puede mejorar todavía el modelo sin llegar a sobre ajustarlo.
9. Seleccione ahora el SalesPrice como variable respuesta.
10. Genere dos modelos de regresión con redes neuronales con diferentes topologías y funciones de activación para predecir el precio de las casas.
11. Compare los dos modelos de regresión y determine cuál funcionó mejor para predecir el precio de las casas.
12. Analice si no hay sobreajuste en los modelos. Use para esto la curva de aprendizaje.
13. Para el modelo elegido de regresión tune los parámetros y discuta si puede mejorar todavía el modelo sin llegar a sobre ajustarlo.
14. Compare la eficiencia del mejor modelo de RNA con los resultados obtenidos con los algoritmos de las hojas de trabajo anteriores. ¿Cuál es mejor para predecir? ¿Cuál se demoró más en procesar?
15. Compare los resultados del mejor modelo de esta hoja para clasificar con los resultados de los algoritmos usados para clasificar de las hojas de trabajo anteriores
16. Compare los resultados del mejor modelo para predecir el precio de venta con los resultados de los algoritmos usados para el mismo propósito de las hojas de trabajo anteriores.
17. Ahora que ha usado todos los modelos que hemos visto y aplicados al conjunto de datos llegue a conclusiones sobre cual es o cuales son los mejores modelos para clasificar dadas las características del conjunto de datos. ¿Cuál o cuáles son los mejores para predecir el precio de las casas? Una tabla de resumen con las métricas de los modelos le puede resultar muy útil para esto.
18. Genere un informe de los resultados y las explicaciones.

EVALUACIÓN

- **(10 puntos)** Creación de los dos modelos de clasificación usando diferentes topologías y funciones de activación. Queda claro cuales fueron las topologías y funciones de activación usadas en cada modelo.
- **(15 puntos)**. Matriz de confusión de ambos modelos para clasificar.
- **(20 puntos)**. Comparación de resultados y selección del mejor modelo de clasificación. Da una explicación amplia de sus resultados. Analiza si no tienen sobreajuste y tunea los parámetros del mejor modelo.
- **(10 puntos)**. Creación de los dos modelos de regresión usando diferentes topologías y funciones de activación. Queda claro cuáles fueron las topologías y funciones de activación usadas en cada modelo.

- **(20 puntos).** Comparación de resultados y selección del mejor modelo de regresión. Da una explicación amplia de sus resultados. Analiza si no tienen sobreajuste y tunea los parámetros del mejor modelo.
- **(15 puntos).** Comparación con la efectividad de los algoritmos de las hojas de trabajo anteriores.
- **(10 puntos).** Conclusiones. Debe seleccionar los mejores modelos de clasificación y regresión para el conjunto de datos.

MATERIAL A ENTREGAR

- Archivo .r o .py con el código y hallazgos comentados
- Link de Google docs con las conclusiones y hallazgos encontrados. Puede usar también Jupyter Notebooks o rmd.
- Vínculo del repositorio usado para trabajar la hoja de trabajo.