

Hoja de Trabajo 5. Naive Bayes

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.
- La hoja no se calificará si no pertenece a ningún grupo de los creados en canvas para esta hoja.

INSTRUCCIONES

Utilice el data set House Prices: Advanced Regression Techniques que comparte Kaggle. Puede usar el análisis exploratorio que hizo en hojas anteriores. Genere un informe con las explicaciones de los pasos que llevó a cabo y los resultados obtenidos en la generación y aplicación de los modelos. 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 dos hojas anteriores.
- 2. Elabore un modelo de bayes ingenuo (naive bayes) utilizando el conjunto de entrenamiento y explique los resultados a los que llega. El experimento debe ser reproducible por lo que debe fijar que los conjuntos de entrenamiento y prueba sean los mismos siempre que se ejecute el código.
- 3. El modelo debe ser de clasificación, use la variable categórica que hizo con el precio de las casas (barata, media y cara) como variable respuesta.
- 4. Utilice el modelo con el conjunto de prueba y determine la eficiencia del algoritmo para clasificar.



- 5. Haga un análisis de la eficiencia del algoritmo usando una matriz de confusión. Tenga en cuenta la efectividad, donde el algoritmo se equivocó más, donde se equivocó menos y la importancia que tienen los errores.
- 6. Analice el modelo. Explique si hay sobreajuste (overfitting) o no.
- 7. Haga un modelo usando validación cruzada, compare los resultados de este con los del modelo anterior. ¿Cuál funcionó mejor?
- 8. Compare la eficiencia del algoritmo con el resultado obtenido con el árbol de decisión (el de clasificación). ¿Cuál es mejor para predecir? ¿Cuál se demoró más en procesar?

EVALUACIÓN

- (25 puntos) Análisis de los modelos generados. Recuerde explicar los razonamientos.
- (25 puntos) Análisis de las variables a incluir en el modelo. Pruebas de normalidad, correlación, etc.
- (10 puntos) Aplicación de los modelos al conjunto de prueba.
- (20 puntos) Matriz de confusión de cada modelo. Explicación de los resultados obtenidos.
- (20 puntos) Comparación del método de naive bayes con el árbol de clasificación.

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.