+34 691 225 633
Rev. Febrero 2021

Autor: Iván García Santillán

Semana (2)

Predicción del cáncer de mama usando Regresión logística y K-NN

Objetivo y alcance del trabajo

Esta práctica tiene el objetivo de realizar la predicción del tipo de cáncer de mama (maligno=1, benigno=0) utilizando los algoritmos de clasificación revisados de regresión logística y K-NN, así como realizar la evaluación del rendimiento de los algoritmos usando las métricas y gráficas respectivas. Al finalizar la práctica, los maestrantes podrán entender y manipular adecuadamente los diferentes parámetros e hiperparámetros de los algoritmos estudiados.

Explicación del Trabajo

Utilice el dataset de cáncer de mama (Diagnostic Wisconsin Breast Cancer Database, WBCD) donde tenemos los datos de 569 mujeres. Cada mujer está descrita por 32 atributos. El primero es un identificador, el segundo el tipo de cáncer (Maligno o Benigno) y el resto son el resultado de otros análisis clínicos. Se pretende aprender el tipo de cáncer de mama (maligno=1, benigno=0). En este conjunto de datos la distribución de clases es: 357 benignos y 212 malignos.

Fase I

Se pide al maestrante realizar las siguientes actividades:

- Descargue el dataset: http://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
- Cargue el conjunto de datos en el entorno de Python
- Realizar el preprocesamiento de datos y Análisis exploratorio de datos (estadísticas y visualización) como se indicó en la unidad anterior.
- Implementar, entrenar y afinar el algoritmo de **Regresión logística** usando Scikit-Learn.

- Evaluar el algoritmo usando las métricas y gráficas de rendimiento básicas.
- Realice un informe conciso en Word interpretando los resultados obtenidos en al menos 2 versiones (*baseline*, mejorado). A lo mejor se pueda experimentar eliminando datos atípicos, escalando datos, cambiando valores de hiperparámetros, eliminando variables irrelevantes basado en el análisis de correlación y de los coeficientes del modelo (importancia de características), etc.
- Suba el notebook de Python y/o el informe Word al LMS (plataforma).

Fase II

Se pide al maestrante realizar las siguientes actividades:

- Implementar, entrenar y afinar el algoritmo de **K-NN** usando Scikit-Learn.
- Evaluar el algoritmo usando las métricas y gráficas de rendimiento básicas.
- Compare los resultados de ambos algoritmos entrenados.
- Realice un informe conciso en Word interpretando los resultados obtenidos en al menos 2 versiones (*baseline*, mejorado). A lo mejor se pueda experimentar eliminando datos atípicos, escalando datos, cambiando valores de hiperparámetros, eliminando variables irrelevantes, etc.
- Suba el notebook de Python y/o el informe Word al LMS (plataforma).

ANEXOS

Rúbrica de Evaluación de la práctica

+34 691 225 633
Rev. Febrero 2021

Rúbrica de calificación de la práctica:

Criterio	Excelente	Bueno	Aceptable	Bajo
Análisis exploratorio de datos	El análisis exploratorio de datos es relevante, detallado y adecuado para la comprensión de los datos y el negocio incluyendo un preprocesamiento de datos, estadísticas y visualización de datos.	El análisis exploratorio de datos es adecuado para la comprensión de los datos y el negocio incluyendo un preprocesamiento de datos, estadísticas y visualización de datos.	El análisis exploratorio de datos es adecuado para la comprensión de los datos y el negocio, pero podría mejorarse en la variedad de técnicas, estadísticas y visualizaciones.	El análisis exploratorio de datos no es adecuado para la comprensión de los datos y el negocio y debe mejorarse significativamente.
Entrenamiento del modelo	El entrenamiento del modelo de clasificación es adecuado incluyendo todos los algoritmos revisados, el afinamiento y selección de los mejores parámetros.	El entrenamiento del modelo de clasificación es adecuado incluyendo varios algoritmos revisados, el afinamiento y selección de los mejores parámetros.	El entrenamiento del modelo de clasificación es adecuado incluyendo algunos de los algoritmos revisados, pero podría mejorarse considerando el afinamiento y selección de los mejores parámetros.	El entrenamiento del modelo de clasificación no es adecuado con métricas de evaluación bajos.
Elaboración de informe	El informe presentado es relevante, claro y conciso incluyendo todas las métricas y gráficas de evaluación del modelo entrenado.	El informe presentado es adecuado y conciso incluyendo varias métricas y gráficas de evaluación del modelo entrenado.	El informe presentado es adecuado incluyendo algunas métricas y gráficas de evaluación del modelo entrenado, pero podría	El informe presentado no es adecuado, faltando métricas y gráficas de evaluación del modelo entrenado.

+34 691 225 633

			complementarse con otras adicionales.	
Presentación del caso práctico	La presentación del caso práctico es clara, completa y concisa utilizando material visual en el tiempo asignado demostrando un alto nivel de creatividad.	La presentación del caso práctico es clara y concisa utilizando material visual en el tiempo asignado.	La presentación del caso práctico es clara utilizando material visual en el tiempo asignado, pero podría profundizarse en aspectos relevantes.	La presentación del caso práctico no es clara, no utiliza material visual adecuado y no respeta el tiempo asignado.