12/02/2024

19/02/2024

Predecir popularidad de tweets con un modelo de clasificación.

* Regresión: valores continuos
* Clasificación: valores discretos (número limitado de opciones)

Los problemas de regresión se pueden transformar en problemas de clasificación (para responder a preguntas sí o no).

Es importante recalcar que los modelos predicen una probabilidad, pero después hay que darle significado a esa probabilidad.

No es lo mismo predecir un ‘no’ cuando la respuesta es ‘si’ que predecir un ‘si’ cuando la respuesta es no (ejemplo detección enfermedades)

Datos

Hay dos conjuntos de datos que se han de relacionar.

No hay ninguna variable que sea un target como tal

Contexto

Eres una empresa que busca crecer su negocia a través a de publicar mensajes en redes sociales (Content marketing). Atraer clientes a través de un producto.

Quieren buscar tweets populares para crear contenido alrededor de ese tweet, lo antes posible.

Para ello nos piden crear un modelo que detecte que tweets van a ser populares lo antes posible.

Los datos son scrapeados de twitter, tantos de tweets como de usuariso.

Targets para clasificación

El target de la clasificación es una variable con un numero de variables discreto, el mas común es un target binario. 2 posibles valores.

Hay que crear un target a partir de datos continuos

* Se puede crear una variable a partir de un numero determinado de likes.

Los datos tienen muchísimos outliers, lo cual perjudica al modelo

Hay que tener cuidado con escoger un target demasiado alto con muy poca representación porque el modelo siempre va a predecir el mismo valor. Se arregla diciendo el coste de acertar y fallar, estableciendo un castigo dependiendo de porque falla el modelo

Stuff to consider

Para saber si el tweet es relevante para crear un articulo, hay que saber cuanto se tarda en crear un articulo, cuanto cobra un periodista por ese articulo y cuanto gana la empresa por visita aprox, cuanto se convierte de visita a tweet a visita a la pagina.

Se va a hacer contenido relacionado al tweet, por lo que importa es la visitas/interacion de la gente mas que los retweets/likes

Tener en cuenta el tiempo de los tweets?

**Target se podría considerar cualquier variable que cambia con el tiempo**

04/03/2024

Predicción

Predict vs predict\_proba

Predict\_proba da un array de probabilidades entre 0 y 1. Permite indicar un umbral para que se considere una clase a partir de dicho umbral.

El modelo te otorga una predicción pero tu tienes que decidir que hacer con esa predicción.

Modelos

* Regresión Logistica
* Modelos en árbol: los mas versátiles, existen el básico y luego optimizaciones de estos modelos donde se hace ensambling
* Redes neuronal

Cuanto mas complejo es un modelo, mas probabilidad hay de que ocurra overfitting. Cuando mas esecifico eres, puedes acertar mas pero también puedes fallar mas

11/03/2024

Optimización

Todos los modelos tienen una gran cantidad de parámetros para ajustar su comportamiento. Probar todas las posibles combinaciones de modelos y parámetros puede resultar una tarea imposible. Para dar saltos muy grandes de rendimiento es necesario cambiar la metodología, sin embargo, para mejorar ligeramente el modelo, esta técnica resulta útil.

La optimización se suele realizar una vez satisfecho con el modelo y metodología ya seleccionado.

Para realizar estas pruebas de optimización automáticamente, se puede utilizar la herramienta GridSearchCV. Entrena el modelo con todas las combinaciones y devuelve la mejor combinación.

Esta metodología se puede combinar con la herramienta Pipeline para mejorar sus resultados. Es un wrapper para ejecutar varios pasos uno detrás de otros

15/04/2024

Probabilidad vs target final. La probabilidad no tiene que tomar decisiones por nosotros mismos, nosotros somo los que utilizando esa probabilidad decidimos que hacer con el modelo

**Evaluación**

Como saber si nuestro modelo nos ayuda a cumplir nuestro objetivo.

I se utiliza un método de evaluación logarítmico, conviene clipear el valor al menos a 0.99 o 0.01 para evitar que te errores infinitos se falla

Simulaciones

**Caer en un umbral final**

06/05/2024

Variables avanzadas

* Conteo de actividad historica de usuario. Saber que ha hecho el usuario con anterioridad. Por ejemplo: si un usuario a cometido fraude a un seguro, es posible que vuelva a cometer fraude. Puede generar problemas a la hora de evaluación ya que se pueden meter datos del train en el test
* Análisis del contexto de un ejemplo: no solo usar el histórico de las personas, sino también el histórico del contexto. Si un día hubo muchos tweets populares en un día, es mas probable que un tweet sea popular
* Reducción de dimensiones: si tienes muchas variables, de esta forma se puede resumir variables en una sola. Es mejor trabajar con una sola variable que con muchas
* Análisis de texto e imágenes: los textos e imágenes se pueden introducir al modelo de manera directa o utilizar un modelo aparte para extraer características
  + Utilizar técnicas como tf-idf