Scikit Learn

Clasificación

CertiDevs

Índice de contenidos

1. Clasificación

La **clasificación** es una tarea central en el campo del aprendizaje automático y la ciencia de datos.

Es un tipo de **aprendizaje supervisado** que se enfoca en predecir una variable de salida **categórica** a partir de una serie de variables de entrada. La variable de salida categórica es a menudo referida como la **etiqueta** o **clase**, y las variables de entrada son conocidas como características o atributos.

En un problema de **clasificación**, se tiene un conjunto de observaciones o instancias, cada una de las cuales está asociada con una etiqueta de clase específica. El objetivo es construir un modelo que pueda determinar la etiqueta de clase de una observación basándose en sus características. Este modelo se construye utilizando un conjunto de datos de entrenamiento, que consiste en observaciones cuyas etiquetas de clase ya son conocidas.

El proceso de construcción del **modelo** involucra el uso de un algoritmo de aprendizaje automático para encontrar patrones en los datos de entrenamiento que relacionen las características de las observaciones con sus etiquetas de clase. El modelo resultante es una representación matemática de estos patrones. Una vez que el modelo ha sido **entrenado** de esta manera, puede ser utilizado para **predecir** la etiqueta de clase de nuevas observaciones cuyas características son conocidas, pero cuyas etiquetas de clase son desconocidas.

Hay muchos **algoritmos de clasificación disponibles**, cada uno con sus propios puntos fuertes y débiles. Algunos de los algoritmos más comunes incluyen la **regresión logística**, las **máquinas de vectores de soporte (SVM)**, los **árboles de decisión**, los **bosques aleatorios**, y las **redes neuronales**. La elección del algoritmo a utilizar puede depender de muchos factores, incluyendo la naturaleza de los datos, el número de características y observaciones, y el tipo de problema que se está tratando de resolver.

Los modelos de clasificación se utilizan en una amplia variedad de aplicaciones, desde el filtrado de spam en el correo electrónico y el reconocimiento de dígitos escritos a mano, hasta la detección de fraudes en transacciones financieras y la predicción de enfermedades en medicina. A pesar de su utilidad, también hay desafíos asociados con la clasificación, como el manejo de datos desbalanceados, la reducción de la dimensionalidad y la prevención del sobreajuste.

2. Funcionamiento

Datos de Entrenamiento: En el aprendizaje supervisado, los datos de entrenamiento consisten en observaciones para las que conocemos la "respuesta correcta". En la clasificación, cada observación en los datos de entrenamiento tiene una etiqueta que indica a qué clase pertenece. Por ejemplo, si estás construyendo un filtro de spam, tus datos de entrenamiento pueden consistir en una serie de correos electrónicos, cada uno de los cuales ha sido etiquetado como "spam" o "no spam".

Elección del Algoritmo: Diferentes problemas de clasificación pueden requerir diferentes tipos de algoritmos de clasificación. Algunos de los algoritmos de clasificación más comunes incluyen la regresión logística, las máquinas de vectores de soporte, los árboles de decisión, los bosques aleatorios, las redes neuronales, etc. Cada uno de estos algoritmos tiene sus propios puntos fuertes y débiles, y puede ser más adecuado para ciertos tipos de datos y problemas.

Entrenamiento del Modelo: Una vez que has elegido un algoritmo, el siguiente paso es entrenar tu modelo de clasificación en tus datos de entrenamiento. Durante el entrenamiento, el algoritmo ajustará sus parámetros internos para minimizar el error entre las predicciones del modelo y las etiquetas reales en los datos de entrenamiento. El objetivo es aprender un modelo que pueda predecir la etiqueta correcta a partir de las características de una observación.

Predicción: Una vez que tu modelo ha sido entrenado, puedes usarlo para predecir la clase de nuevas observaciones. Por ejemplo, si has construido un filtro de spam, puedes usar tu modelo de clasificación para predecir si un correo electrónico nuevo es spam o no.

Evaluación del Modelo: Para entender cómo de bueno es tu modelo, necesitas evaluarlo. En la clasificación, hay varias métricas de evaluación comunes, como la precisión (la proporción de predicciones correctas), el recall (la proporción de positivos reales que se identificaron correctamente) y el F1-score (una medida que combina precisión y recall).

Mejora del Modelo: Basándote en tu evaluación del modelo, puedes querer mejorar tu modelo ajustando sus parámetros, eligiendo un algoritmo de clasificación diferente, o incluso recopilando más datos de entrenamiento.

3. Tipos de clasificación

3.1. Clasificación binaria

La **clasificación binaria**, como su nombre indica, es un tipo especial de clasificación en la que la variable objetivo tiene solo dos categorías posibles. El término "binario" se refiere al hecho de que hay exactamente dos clases en las que se pueden clasificar las observaciones. A menudo, estas clases se denotan como "1" (para positivos) y "0" (para negativos), aunque también podrían ser etiquetadas con cualquier par de categorías distintas, como "verdadero" y "falso", "sí" y "no", "éxito" y "fracaso", etc.

Uno de los ejemplos más comunes de clasificación binaria es la detección de spam en los correos electrónicos. En este caso, el algoritmo de clasificación binaria es entrenado con un conjunto de correos electrónicos que han sido previamente etiquetados como "spam" o "no spam". El algoritmo aprende de estas etiquetas y utiliza esa información para clasificar nuevos correos electrónicos en una de estas dos categorías. Las características que se utilizan para entrenar el modelo podrían ser la presencia o ausencia de ciertas palabras o frases, la frecuencia de ciertos caracteres, el formato del correo electrónico, etc.

Otro ejemplo común de clasificación binaria es la predicción de la aprobación de crédito en el sector bancario. En este caso, las dos clases podrían ser "crédito aprobado" y "crédito denegado". El algoritmo de clasificación binaria se entrena con los datos de los solicitantes anteriores, que incluyen características como la puntuación de crédito, el ingreso anual, el nivel de endeudamiento y la historia crediticia, junto con la decisión de si se aprobó o no el crédito. Luego, el modelo puede ser usado para predecir si un nuevo solicitante tendrá su crédito aprobado o denegado basándose en sus características.

En el campo de la medicina, un ejemplo de clasificación binaria podría ser la predicción de la presencia o ausencia de una enfermedad basándose en los resultados de varias pruebas médicas.

Por ejemplo, se podría utilizar un algoritmo de clasificación binaria para predecir si un paciente tiene o no cáncer basándose en características como la edad, el género, los antecedentes familiares, los resultados de las pruebas de sangre, etc.

Aunque la clasificación binaria puede parecer relativamente simple debido al hecho de que solo hay dos clases, en realidad puede presentar muchos desafíos. Por ejemplo, los datos pueden estar desbalanceados, con muchas más observaciones en una clase que en la otra. Además, la elección de las características correctas para usar en el modelo puede ser un proceso complejo que requiere un buen entendimiento tanto de los datos como del problema en sí. A pesar de estos desafíos, la clasificación binaria es una herramienta esencial en muchos campos y tiene una amplia gama de aplicaciones.

3.1.1. Puntos clave de la clasificación binaria:

1. **Definición:** La clasificación binaria es un tipo de clasificación en el aprendizaje automático supervisado en la que la variable objetivo tiene solo dos categorías posibles.
2. **Etiquetas de clase:** Las dos categorías en las que se pueden clasificar las observaciones a menudo se denotan como "1" (para positivos) y "0" (para negativos), aunque también podrían ser etiquetadas con cualquier par de categorías distintas.
3. **Modelado:** El objetivo de la clasificación binaria es construir un modelo que pueda determinar la etiqueta de clase de una observación basándose en sus características. Este modelo se construye utilizando un conjunto de datos de entrenamiento, cuyas etiquetas de clase ya son conocidas.
4. **Algoritmos:** Hay muchos algoritmos disponibles para la clasificación binaria, incluyendo la regresión logística, las máquinas de vectores de soporte, los árboles de decisión y los bosques aleatorios, entre otros.
5. **Aplicaciones:** La clasificación binaria se utiliza en una variedad de aplicaciones, incluyendo el filtrado de spam, la aprobación de crédito, la detección de enfermedades, y más.
6. **Desafíos:** Algunos de los desafíos en la clasificación binaria incluyen el manejo de datos desbalanceados, la elección de las características adecuadas, y la prevención del sobreajuste.
7. **Evaluación del rendimiento:** Las métricas comunes para evaluar el rendimiento de los modelos de clasificación binaria incluyen la precisión, la sensibilidad, la especificidad, el área bajo la curva ROC, entre otras. Estas métricas proporcionan diferentes formas de evaluar cómo de bien el modelo distingue entre las dos clases.
8. **Optimización del modelo:** Los modelos de clasificación binaria pueden ser optimizados a través de técnicas como la validación cruzada, la selección de características, y el ajuste de parámetros.

3.2. Clasificación multiclase

La **clasificación multiclase**, también conocida como clasificación multietiqueta o multinomial, es un tipo de clasificación supervisada en la que la variable objetivo puede tener más de dos categorías.

En otras palabras, en lugar de predecir simplemente si una observación cae en una de las dos

categorías, como en la clasificación binaria, los problemas de clasificación multiclase implican predecir una de las tres o más categorías.

Ejemplos:

- Un ejemplo de clasificación multiclase se puede encontrar en el campo de la **visión por computadora**, donde un modelo podría estar entrenado para reconocer varias categorías de imágenes. Por ejemplo, podrías tener un modelo que está diseñado para clasificar imágenes en varias categorías como "perro", "gato", "coche", "casa", etc. Las características podrían incluir patrones de píxeles, colores, formas, texturas, etc.
- Otro ejemplo de clasificación multiclase es en el campo de la inteligencia artificial para **juegos**, donde se puede entrenar un modelo para predecir los movimientos de un jugador basándose en su estado actual en el juego. Las categorías en este caso podrían ser diferentes movimientos o estrategias, como "avanzar", "retroceder", "saltar", "agacharse", etc.
- En el ámbito de la **medicina**, un modelo de clasificación multiclase podría ser utilizado para diagnosticar diferentes tipos de enfermedades basándose en ciertos síntomas o resultados de pruebas. Por ejemplo, un modelo podría ser entrenado para clasificar a los pacientes en categorías como "sin enfermedad", "diabetes", "enfermedad cardíaca", "enfermedad renal", etc., basándose en características como los niveles de azúcar en sangre, la presión arterial, los niveles de creatinina, etc.

Los modelos de clasificación multiclase a menudo requieren técnicas más complejas que los modelos de clasificación binaria.

Hay varias estrategias para lidiar con la clasificación multiclase. Algunas técnicas, como la regresión logística multinomial o las redes neuronales, se extienden naturalmente al caso multiclase. Otras técnicas, como las máquinas de vectores de soporte, se diseñan inicialmente para la clasificación binaria, pero se pueden adaptar a la clasificación multiclase mediante enfoques como "uno contra uno" o "uno contra todos". En el enfoque "uno contra uno", se construye un clasificador para cada par de clases y la clase que recibe la mayoría de los votos de los clasificadores se selecciona como la predicción final. En el enfoque "uno contra todos", se construye un clasificador para cada clase que distingue esa clase de todas las demás clases, y la clase con la mayor confianza se selecciona como la predicción final.

Cada uno de estos enfoques tiene sus ventajas y desventajas y la elección entre ellos puede depender de varios factores, incluyendo la naturaleza de los datos, el número de características y observaciones, el número de clases y el tipo de problema que se está tratando de resolver.

3.2.1. Puntos clave de la clasificación multiclase:

1. **Definición:** La clasificación multiclase es un tipo de clasificación supervisada en el aprendizaje automático en la que la variable objetivo tiene más de dos categorías posibles.
2. **Número de clases:** A diferencia de la clasificación binaria, que solo tiene dos posibles resultados, la clasificación multiclase tiene tres o más categorías posibles.
3. **Algoritmos:** Algunos algoritmos de aprendizaje automático, como la regresión logística multinomial y las redes neuronales, se extienden naturalmente al caso multiclase. Otros, como las máquinas de vectores de soporte, se adaptan al caso multiclase mediante enfoques como

"uno contra uno" o "uno contra todos".

4. **Modelado:** La meta de la clasificación multiclase es construir un modelo que pueda determinar la etiqueta de clase de una observación basada en sus características. Este modelo se construye utilizando un conjunto de datos de entrenamiento, cuyas etiquetas de clase ya son conocidas.
5. **Aplicaciones:** La clasificación multiclase tiene una amplia gama de aplicaciones, incluyendo el reconocimiento de imágenes, el diagnóstico médico, la predicción de movimientos en juegos, entre otros.
6. **Desafíos:** Algunos de los desafíos en la clasificación multiclase incluyen el manejo de un gran número de clases, el balanceo de las clases, la elección de las características y la prevención del sobreajuste.
7. **Evaluación del rendimiento:** Las métricas comunes para evaluar el rendimiento de los modelos de clasificación multiclase incluyen la precisión multiclase, la matriz de confusión, el puntaje F1 macro/micro ponderado, entre otros.
8. **Optimización del modelo:** Al igual que con la clasificación binaria, los modelos de clasificación multiclase pueden ser optimizados a través de técnicas como la validación cruzada, la selección de características y el ajuste de hiperparámetros.