Nuevas Tendencias

Minería de Datos - Grado en Matemáticas

DSLAB

2025-11-13

Ética en el Aprendizaje Automático

- La toma de decisiones automatizada basada en datos plantea desafíos éticos y sociales que deben ser abordados de manera responsable

Ética

- Disciplina filosófica que estudia el bien y el mal y sus relaciones con la moral y el comportamiento humano.
- Conjunto de costumbres y normas que dirigen o valoran el comportamiento humano en una comunidad.
- La ética en el Aprendizaje Automático se centra en garantizar que los sistemas de IA tomen decisiones justas, imparciales y éticas.

Aprendizaje Automático Explicable

El aprendizaje máquina explicable (**XML**, "Explainable Machine Learning" en inglés) se refiere a la capacidad de los modelos de ML para proporcionar explicaciones claras y comprensibles de sus decisiones.

Métodos basados en reglas

- **Reglas de decisión**: Estos modelos generan reglas lógicas que explican el razonamiento detrás de las predicciones del modelo.
- **Árboles de decisión**: Los árboles muestran la secuencia de decisiones tomadas por el modelo en cada nodo, lo que facilita la interpretación.

Métodos basados en características

- **Importancia de características**: Calcula la importancia de cada característica en el modelo, lo que permite identificar las variables más influyentes en las predicciones.
- **Análisis de efectos parciales**: Evalúa el impacto de una sola característica en las predicciones, manteniendo las demás constantes.

Métodos basados en ejemplos

- **Prototipos**: Encuentra ejemplos representativos o prototipos de datos que explican cómo el modelo toma decisiones.
- **Casos de prueba**: Genera instancias que muestran cómo el modelo reacciona a diferentes escenarios.

Métodos de aproximación

- **Modelos locales interpretables**: Crea modelos más simples (lineales, regresiones, etc.) en regiones locales del espacio de características para comprender decisiones en áreas específicas.
- **Regresiones lineales localmente ponderadas (LWLR)**: Asigna pesos a las instancias cercanas para ajustar una regresión lineal local.

Métodos basados en atención

- **Atención y atención saliente**: Modelos basados en atención destacan características o regiones de interés que influyen en las predicciones.
- **Redes neuronales con atención**: Las redes neuronales con mecanismos de atención permiten entender qué partes de la entrada son relevantes para la salida.

Métodos basados en métricas

- **Métricas de proximidad**: Evalúan la similitud entre entradas y cómo se relacionan con las predicciones.
- **SHAP (SHapley Additive exPlanations)**: Utiliza conceptos de teoría de juegos para asignar valores de importancia a las características.

Métodos basados en resumen

- **Reglas de decisión generadas por el modelo**: El modelo crea reglas que resumen su comportamiento.
- **Análisis de componentes**: Reduce la dimensión de los datos para visualizar y resumir características significativas.

Métodos de muestreo y generación de datos

- **Perturbación de datos**: Se modifican las características de entrada para entender cómo afectan a las predicciones.
- **Muestreo de datos contrapuestos**: Se generan ejemplos que muestran cómo las predicciones cambiarían si los datos fueran diferentes.

Métodos de comparación

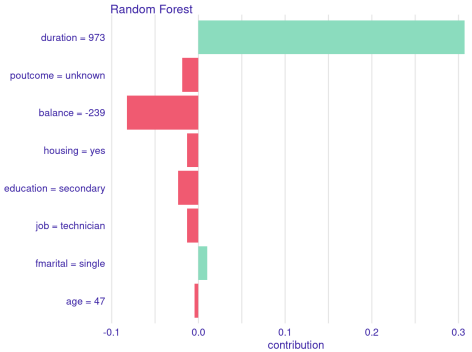
- **Modelos interpretables frente a modelos de caja negra**: Compara modelos interpretables con modelos complejos en términos de rendimiento y capacidad de explicación.

Herramientas de visualización

- **Gráficos interactivos**: Visualizaciones que muestran cómo las características afectan a las predicciones.
- **Heatmaps y perfiles de importancia**: Muestran la importancia de las características en un formato visual.

SHAP

SHAP se basa en el principio de que cada característica o atributo de entrada de un modelo contribuye de alguna manera a la predicción final. SHAP cuantifica esta contribución para cada característica, permitiendo una interpretación más profunda de cómo el modelo llega a sus conclusiones.



Contrafácticos

- Los contrafácticos son preguntas o declaraciones que plantean "¿Qué habría ocurrido si...?" con el propósito de analizar cómo un modelo habría respondido si las condiciones o las entradas hubieran sido diferentes.
- Esta técnica se utiliza para obtener información sobre cómo un modelo realiza sus predicciones y para explicar su razonamiento.

Deriva conceptual

La deriva conceptual es un fenómeno crítico que se refiere a la evolución de los datos y la cambiante naturaleza de las relaciones entre las variables a lo largo del tiempo.

Cambios

- en la **distribución de los datos**
- en la **importancia relativa de las características**
- en las **relaciones entre variables**
- **nuevos patrones** que antes no estaban presentes.

Deriva conceptual

Para abordar la deriva conceptual, es necesario implementar técnicas de adaptación de modelos que permitan a los algoritmos de ML ajustarse a los cambios en los datos.

- **reentrenamiento periódico** de los modelos
- **monitorización constante de la calidad** del modelo
- **identificación de los momentos** en que se produce una deriva conceptual significativa.

Referencias

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering*, *31*(12), 2346-2363.