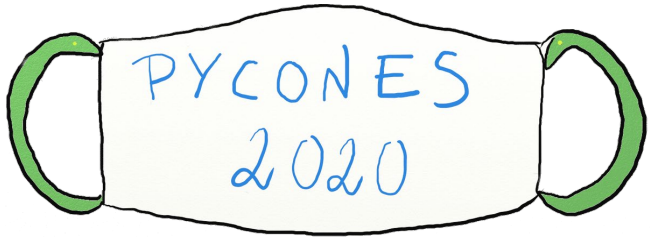


Interpretabilidad de Modelos de Machine Learning con Python

Antonio Soto

Director Unidad Data & AI Verne Tech

@antoniosql

asoto@solidq.com

SolidQ | VERNE TECHNOLOGY GROUP

Agenda

- ¿Por qué necesitamos interpretar los modelos?
- ¿Qué técnicas tenemos para interpretar?
- Algunos Frameworks Python para interpretabilidad
- Demostración

La Importancia del ML Interpretable

- ¿Cómo sé que puedo confiar en el modelo?
- ¿Cómo toma sus decisiones?
- Balance entre rendimiento e interpretabilidad
- Interpretaciones
  - Global ¿Cómo hace las predicciones? ¿Cómo influyen los subconjuntos de datos?
  - Local ¿Por qué ha tomado una decisión en concreto para un caso en concreto?

SolidQ | VERNE TECHNOLOGY GROUP

Motivaciones

- Legales
- Éticas
  - Sesgo
- De Negocio

Beneficios

- Dar confiabilidad en los resultados.
- Ayudar en la Depuración.
- Informar a la Ingeniería de Características (Feature Engineer).
- Detectar necesidad de recoger nuevas muestras.
- Ayudar a una persona en la toma de decisiones.
- Mayor seguridad/robustez en el modelo obtenido.

Las Explicaciones

Una explicación relaciona los valores de las características de una instancia, con sus predicciones de un modo entendible por las personas.

- Propiedades del Método de Explicación
  - Poder de expresividad
  - Translucidez
  - Portabilidad
  - Complejidad algorítmica

¿Qué técnicas Tenemos?

Técnicas Tradicionales

- Técnicas EDA
  - Reducción de Dimensionalidad (PCA)
  - ¿Qué debemos de “buscar”?
    - **Balanceo**
    - **Escalado**
- Métricas de Evaluación de rendimiento de nuestro modelo
  - Aprendizaje Supervisado – Clasificación
    - Matriz de Confusión
    - Accuracy, Precision, Recall, F1-Score
    - ROC y AUOC score
  - Aprendizaje Supervisado – Regresión
    - Coeficiente de determinación (R-square)
    - RMSE (root mean-square error)
    - MAE (Mean absolute error)

Limitaciones Técnicas Tradicionales

- Medimos el rendimiento
- ¿Cumple con parámetros de rendimiento válidos?
- Con estas técnicas, ¿cómo aseguramos?
  - JUSTICIA
  - RESPONSABILIDAD
  - TRANSPARENCIA
- Balance entre Rendimiento e Interpretabilidad

Técnicas Adicionales

1. Uso de Modelos interpretables
2. Importancia de Características
3. Plots de dependencia parcial (PDP)
4. Modelos Subrrogados Globales

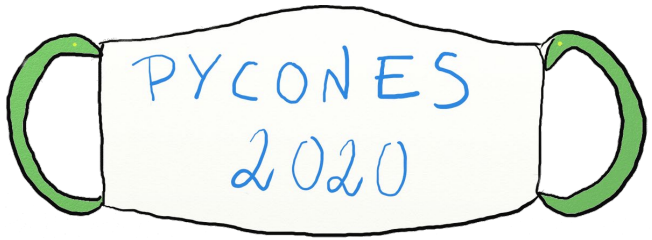
Algunos Frameworks de Interpretabilidad

LIME

- Explicaciones para una predicción individual de un modelo de caja negra
- Concepto
  - Nos olvidamos de datos de entrenamiento
  - Tenemos una caja negra a la que podemos pasarle las instancias que queramos las veces que queramos para ver que predicciones genera
  - Objetivo: Entender porque hace cada una de esas predicciones
  - LIME prueba que ocurre con las predicciones cuando se varían los datos de entrada
  - Genera un conjunto de datos compuesto de muestras permutadas y su correspondiente predicción por parte del modelo de Caja Negra
  - Entrena un modelo interpretable con ese conjunto de datos

SHAP

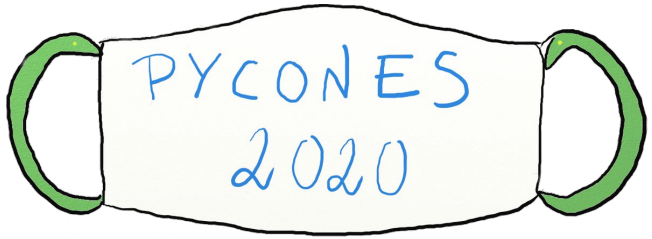
- Aplicación de Teoría de Juegos
- Intenta explicar una predicción asumiendo que cada característica es un “jugador” en un juego donde la predicción es el “premio”
  - Nos indica como distribuir el “premio” entre los “jugadores”
  - Juego→ Tarea de predicción para una instancia
  - Ganancia→ predicción real menos la predicción promedio para todas las instancias
  - Jugadores→ Valores de las características de la instancia
- Valor de Shapley
  - Contribución marginal promedio de un valor de característica entre todas las posibles relaciones



Demostración

El Framework definitivo: Interpret

SolidQ | VERNE TECHNOLOGY GROUP



Interpretabilidad de Modelos de Machine Learning con Python

Antonio Soto

Director Unidad Data & AI Verne Tech

@antoniosql

asoto@solidq.com

SolidQ | VERNE TECHNOLOGY GROUP