Aprendizaje Automático - Modelos Aditivos Generalizados (GAM)

Máster en Data Science para Finanzas

Contacto:

- Gaspar Cólogan Barajas

  Correo: gaspar.cologan@cunef.edu

- Jose Manuel de Castro Beristaín

  Correo: josemanuel.decastro@cunef.edu

1. ¿Qué son los modelos GAM?

Para poder explicar correctamente un Modelo Aditivo Generalizado, conocido como Generalized Additive Model (GAM), es necesario definir primero qué es un modelo lineal.

Un modelo lineal es un enfoque estadístico que asume una relación lineal entre las variables predictoras y la variable respuesta. Se busca ajustar una línea recta que represente la mejor aproximación a la relación subyacente. Los coeficientes de esta línea indican la magnitud y dirección de la influencia de cada variable predictora sobre la variable de respuesta. Los modelos lineales son ampliamente utilizados debido a su simplicidad y facilidad de interpretación, pero pueden no capturar relaciones no lineales en los datos.

Un Modelo Aditivo Generalizado representa una extensión de los modelos lineales tradicionales. A diferencia de los modelos lineales, los GAM poseen una característica distintiva: la capacidad de capturar relaciones no lineales entre las variables predictoras y la variable de respuesta.

En lugar de limitarse a la estructura de una simple suma ponderada de variables predictoras, los GAM introducen flexibilidad al asumir que la relación puede modelarse como la suma de funciones arbitrarias de cada característica. Esta flexibilidad es esencial para abordar patrones complejos y no lineales presentes en los datos.

En el contexto de los GAM, los coeficientes beta, es decir, los parámetros que se utilizan en los modelos de regresión lineal para describir la relación entre las variables predictoras y la variable de respuesta, son reemplazados por funciones flexibles. Estas funciones, conocidas como splines, desempeñan un papel crucial al permitir la modelización de relaciones no lineales para cada característica. Los splines son funciones matemáticas complejas que suavizan la relación entre las variables, lo que resulta en un modelo altamente flexible que conserva parte de la interpretabilidad de una regresión lineal.

2. Ventajas y limitaciones de los GAM:

La aplicación de los Modelos Aditivos Generalizados tiene algunas ventajas y desventajas. A continuación comentaremos brevemente cada una de ellas.

Ventajas de los GAM:

- Flexibilidad:

  Los Modelos Aditivos Generalizados ofrecen flexibilidad a la hora de modelarr relaciones complejas entre predictores y variables de respuesta, adaptándose a una amplia gama de patrones de datos.

- Interpretabilidad:

  Proporcionan información valiosa sobre la dirección y la importancia de los efectos de los predictores, lo que permite a los investigadores comprender cómo contribuyen los predictores a la variable de respuesta.

- Manejo de múltiples tipos de predictores:

  Los GAM pueden manejar una mezcla de predictores continuos, categóricos y ordinales dentro de un marco unificado.

- Selección automática de variables:

  La incorporación de técnicas de regularización en los GAM ayuda a evitar el sobreajuste y mejorar la generalización.

- Visualización:

  Las funciones suaves de los GAM pueden representarse visualmente, lo que ayuda a interpretar el modelo.

Desventajas de los GAM:

- Complejidad en la interpretación:

  La interpretación de los resultados de los Modelos Aditivos Generalizados puede suponer un reto debido a la complejidad de las funciones suaves, lo que requiere experiencia en modelización estadística.

- Subjetividad en la selección de modelos:

  La elección de grados de libertad y suavidad puede depender del juicio del investigador, introduciendo sesgos o incertidumbre.

- Sensibilidad a Parámetros de Suavizado:

  La elección de parámetros de suavizado puede afectar los resultados del modelo, requiriendo ajustes cuidadosos.

- Tratamiento Limitado de Datos Faltantes:

  Los GAM manejan los datos faltantes a través del análisis de casos completos, lo que puede resultar en tamaños de muestra reducidos y sesgos potenciales.

- Exigencias Computacionales:

  Pueden ser intensivos computacionalmente, especialmente para grandes conjuntos de datos o modelos complejos, demandando tiempo y recursos considerables.

- Limitación a Regresión y Clasificación:

  Son más adecuados para tareas de regresión y clasificación y pueden no ser idóneos para tareas más complejas, como el reconocimiento de imágenes.

3. ¿Cuáles son sus usos?

Los Modelos Aditivos Generalizados son versátiles y pueden utilizarse para diversos fines en la modelización estadística. Estas son algunas aplicaciones comunes de los GAM: Análisis de regresión:

1. Análisis de regresión:

   La regresión lineal tradicional asume una relación lineal entre los predictores y la variable de respuesta. Los GAM amplían este supuesto al permitir relaciones no lineales, lo que los hace adecuados para situaciones en las que la relación es compleja y no puede representarse adecuadamente mediante una línea recta.

2. Análisis de series temporales:

   Los datos de series temporales suelen mostrar patrones no lineales. Los Modelos Aditivos Generalizados pueden captar tendencias no lineales, estacionalidad y otras dinámicas temporales de los datos. Esto los hace valiosos para predecir valores futuros basándose en patrones históricos.

3. Suavizado y ajuste de curvas:

   Los GAM destacan en el suavizado de datos ruidosos y en el ajuste de curvas para captar las tendencias subyacentes. Esto es especialmente útil cuando se trata de datos que pueden tener fluctuaciones o irregularidades, y el objetivo es discernir los patrones subyacentes.

4. Clasificación:

   Aunque la regresión logística se utiliza habitualmente para la clasificación binaria, los GAM también pueden adaptarse a tareas de clasificación. Pueden modelar relaciones no lineales entre los predictores y la probabilidad de pertenecer a una clase específica, lo que proporciona flexibilidad a la hora de capturar límites de decisión complejos.

5. Análisis espacial:

   En el análisis espacial, los GAM pueden aplicarse para modelizar relaciones entre variables espaciales. Esto resulta útil cuando se estudian fenómenos que presentan dependencias espaciales o cuando se investiga cómo contribuyen los factores espaciales a la variación de la variable de respuesta.

6. Modelización ecológica y medioambiental:

   Los GAM se emplean en estudios ecológicos y medioambientales para modelizar las relaciones, a menudo intrincadas, entre los factores medioambientales y las respuestas ecológicas. Las no linealidades y las interacciones entre variables pueden ser cruciales para comprender los sistemas ecológicos, y los GAM son muy adecuados para captar estas complejidades.

7. Epidemiología:

   En epidemiología, los GAM pueden utilizarse para modelizar la relación entre las variables de exposición (por ejemplo, factores ambientales, elecciones de estilo de vida) y los resultados sanitarios. Permiten la detección de efectos no lineales e interacciones, proporcionando una representación más realista de las relaciones en los datos epidemiológicos.

8. Selección de variables:

   Los GAM realizan de forma natural la selección de variables incorporando sólo las variables relevantes a través de las funciones de suavizado. Esto puede ser ventajoso en escenarios con datos de alta dimensionalidad, ayudando a centrarse en los predictores más influyentes.

9. Predicción y pronóstico:

   Los GAM pueden utilizarse para realizar predicciones basadas en datos históricos, especialmente cuando existen patrones no lineales en los datos. Tanto si se trata de predecir valores futuros en una serie temporal como de pronosticar resultados en otros ámbitos, los GAM ofrecen flexibilidad para captar relaciones complejas.

En resumen, la flexibilidad de los GAM los hace aplicables en una amplia gama de campos en los que las relaciones entre variables no son lineales o en los que se necesita una modelización más compleja para captar los matices de los datos. La elección de utilizar un GAM debe guiarse por las características específicas del conjunto de datos y los objetivos del análisis.

4. Ejemplos:

```
# pip install pygam
```

```
# pip install --upgrade numpy
```

Vamos a desarrollar un modelos que nos permita anticipar qué clientes **nos van a realizar una reclamación como empresa de seguros**. Para ello creamos ficticiamente una base de datos de 10.000 instancias:

- **Edad**: distribución normal con 40 años de media y 10 de desviación típica.
- **Historial de reclamaciones**: cada cliente, en el tiempo que le dura la póliza, reclama de media unas dos veces.
- **Tipo de póliza**: atribuímos dos tipos: una poliza estándar (0), y una póliza premium (1). La estándar la elige un 30% de los clientes y la premium un 70%.
- **Duración de la póliza**: utilizamos también una distribución normal de media 2 y desviazión típica uno.
- **Indemnizaciones**: Creamos una distribución normal con media de 1.000 euros y desviación típica de 500.

```
import numpy as np
from pygam import LogisticGAM
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

# Crear datos simulados
np.random.seed(42)

# Variables predictoras
age = np.random.normal(40, 10, 10000)
claims_history = np.random.poisson(2, 10000)
policy_type = np.random.choice([0, 1], size=10000, p=[0.3, 0.7])
policy_duration = np.random.normal(2, 1, 10000)
coverage_amount = np.random.normal(1000, 500, 10000)
```

A partir de la función de probabilidades, atribuímos 1 a aquellos clientes que estimamos que sí reclamarán al seguro y 0 a los que no realizarán la reclamación. El umbral se sitúa en el 0.5.

```
# Variable de respuesta
claim_prob = 1 / (1 + np.exp(-(0.1 * age + 0.2 * claims_history + 0.5 * policy_type - 0.3 * policy_duration + 0.1 * coverage_amount + np

claims = np.random.binomial(1, claim_prob)
```

Dividimos en train y test, siendo X el conjunto de edad, historial de reclamaciones, tipo de póliza, duración de la póliza e indemnizaciones. Como y, tenemos la variable claims, cuyo resultado viene determinado por la probabilidad y se le otorga un valor entero de 1 o 0 según la probabilidad que haya de fraude, con un umbral de 0,5.

La parte test tendrá un tamaño de 0,2.

```
# Dividir los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(
    np.column_stack((age, claims_history, policy_type, policy_duration, coverage_amount)),
    claims,
    test_size=0.2,
    random_state=42
)
```

Construimos el modelo, al que llamamos gam_model. Nos exige, para ejecutar el modelo, que le incluyamos el parámetro lambda, que controla la regularización del modelo. Hemos considerado correcto poner 0.1 porque si lambda fuera más bajo, tendría una mayor sensibilidad y llegar a sobreajustarse. Sin embargo, si fuera más alto, podría llegar a omitir detalles relevantes e incluso patrones a la hora de entrenar.

```
# Construir y ajustar un modelo GAM
# gam_model = LogisticGAM().fit(X_train, y_train)
gam_model = LogisticGAM(lam=0.1).fit(X_train, y_train)
```

```
/usr/local/lib/python3.10/dist-packages/pygam/links.py:151: RuntimeWarning: divide by zero encountered in divide
  return dist.levels / (mu * (dist.levels - mu))
```

```
/usr/local/lib/python3.10/dist-packages/pygam/pygam.py:629: RuntimeWarning: invalid value encountered in multiply
  self.link.gradient(mu, self.distribution) ** 2
```

Hacemos la predicción del modelo

```
y_pred = gam_model.predict(X_test)
```

Evaluamos el modelo con parte de las métricas aprendidas para el proyecto anterior

```
from sklearn.metrics import f1_score, precision_score, recall_score

accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy}')
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

print("Precision:", precision)
print("Recall:", recall)
print("F1 Score:", f1)
```

```
Accuracy: 0.998
Precision: 0.9979654120040692
Recall: 1.0
F1 Score: 0.9989816700610997
```

```
y_pred
```

```
array([ True,  True,  True, ...,  True, False,  True])
```