Regresión Funcional

Autores: Devouassoux Julián, Moon Joseph, Tschopp Joaquín

10 de Noviembre 2024

Abstract

El presente estudio explora la aplicación de la *Regresión Funcional* en la predicción de valores escalares a partir de datos funcionales, específicamente utilizando datos meteorológicos diarios del dataset *CanadianWeather* en el paquete `fda` de `R`. El objetivo principal es predecir la precipitación acumulada anual en distintas localidades de Canadá a partir de las curvas de temperatura media diaria, evaluando diversos enfoques de representación funcional, tales como bases spline (con y sin penalización) y Regresión sobre Componentes Principales Funcionales, conocida en inglés como Functional Principal Component Regression (FPCR). Se llevará a cabo una comparación entre estos modelos de regresión funcional y un modelo de regresión lineal simple como línea base para analizar la efectividad de cada enfoque. La precisión de las predicciones será evaluada mediante métricas de error como el Error Cuadrático Medio (RMSE) y el Error Absoluto Medio (MAE) y comparadas con las obtenidas con un modelo lineal.

1 Descripción General del Problema

El objetivo general de esta investigación es explorar el uso de métodos de *Regresión Funcional* en problemas donde las variables predictoras son de naturaleza funcional, para predecir valores escalares. Para esto, se utilizarán diferentes metodologías de transformación de los datos en funciones, como representaciones mediante bases spline (con y sin penalización) y componentes principales.

Además, se realizará una comparación en la predicción del escalar utilizando regresión lineal simple, para evaluar las ventajas y limitaciones de los modelos de regresión funcional frente a técnicas tradicionales cuando se trabaja con datos funcionales.

El estudio también considerará diferentes conjuntos de datos utilizados en cada tipo de regresión, con el objetivo de obtener una comparación entre ellos y analizar cómo las características específicas de cada dataset influyen en los resultados obtenidos.

Por otro lado la predicción de lluvias es crucial para la agricultura y puede ayudar a prevenir importantes pérdidas financieras, con un gran impacto económico (Gupta, Mall, & Janarthanan., 2022). Por eso puede resultar útil utilizar una variedad de modelos estadísticos o de aprendizaje automático, como regresión de cuantiles (Wasko & Sharma, 2014) o redes neuronales convolucionales (Pan, Hsu, AghaKouchak, & Sorooshian, 2019) para poder modelar las variables que influyen en las precipitaciones y poder predecir la magnitud de las mismas. Por lo tanto para presentar estos temas, se realizarán pruebas en el dataset *CanadianWeather*

del paquete `fda` en `R`. Este dataset contiene datos meteorológicos diarios de varias estaciones en Canadá, lo cual es adecuado para aplicar y comparar los diferentes enfoques de regresión funcional.

2 Objetivos

Esta investigación se centra en el análisis de la *Regresión Funcional* y su aplicación para un problema predictivo. A continuación, se detallan los objetivos específicos que guiarán este estudio:

- Investigar los principios teóricos de la regresión funcional y su aplicabilidad en problemas de predicción escalar.
- Explorar distintos enfoques de regresión funcional, tales como el uso de bases funcionales sin penalización y con penalización, y FPCR, evaluando cuál se adapta mejor al modelado de la temperatura diaria en series temporales.
- Evaluar las implicancias y dificultades de la regresión funcional aplicada al modelado y predicción de la acumulación anual de la precipitación por localidad.
- Interpretar los resultados obtenidos de los modelos, evaluando la precisión de las predicciones.
- Presentar los análisis y conclusiones mediante videos desarrollados por los tres participantes del proyecto.

3 Estado del Arte

La *Regresión Funcional* es una extensión de los métodos de regresión clásicos que permite manejar datos funcionales, es decir, funciones observadas a lo largo de un dominio continuo como el tiempo o el espacio. Este campo ha cobrado importancia debido al incremento en la capacidad de recolectar y almacenar datos de alta dimensionalidad y naturaleza funcional.

Diversos autores han contribuido al desarrollo de la regresión funcional. Ramsay y Silverman proporcionan los fundamentos teóricos y prácticos del análisis de datos funcionales, con un enfoque en cómo representar, suavizar y analizar funciones derivadas de observaciones en dominios continuos. En su trabajo, se destacan herramientas esenciales para convertir datos discretos en representaciones funcionales mediante métodos de suavizado y técnicas como el uso de bases funcionales (por ejemplo, splines y funciones de Fourier), facilitando la representación continua de los datos (J. Ramsay & Silverman, 2005). Por otro lado, Ferraty y Vieu exploran métodos no paramétricos en regresión funcional, enfatizando la flexibilidad de estos modelos para capturar relaciones complejas sin asumir una forma funcional específica (Ferraty & Vieu, 2006) .

En el contexto de predicción escalar a partir de variables funcionales se presentan técnicas de regresión lineal funcional que extienden los modelos de regresión lineal clásica al ámbito funcional (Cuevas, Febrero, & Fraiman, 2002). Para la predicción de funciones a partir de variables numéricas o funcionales, se desarrollan métodos de

regresión funcional con componentes principales funcionales, permitiendo reducir la dimensionalidad y capturar la variabilidad principal en los datos funcionales (Yao, Müller, & Wang, 2005).

Estos trabajos establecen las bases para el análisis y aplicación de modelos de regresión funcional en diversas áreas, proporcionando un marco teórico y metodológico que será fundamental para el desarrollo de esta investigación.

4 Metodología

4.1 Dataset: fuente, revisión y estructura

Se utilizará el conjunto de datos *CanadianWeather*, disponible en el paquete `fda` de `R`. Este dataset contiene datos diarios de temperaturas medias y precipitaciones en diferentes estaciones meteorológicas de Canadá, ideales para aplicar técnicas de regresión funcional debido a su naturaleza temporal y continua (J. O. Ramsay, Wickham, Graves, & Hooker, n.d.).

Tras una revisión y visualización inicial de los datos para asegurar su calidad, se confirmó que el dataset está completo y bien estructurado, por lo que no se requieren transformaciones ni preprocesamientos adicionales. Las series temporales de temperatura y precipitación se considerarán directamente para su representación funcional.

La estructura del dataframe que se utilizará para entrenar los modelos es la siguiente:

- **Estación meteorológica**: Identificador de cada localidad.
- **Función de temperatura anual**: Curva de temperatura media diaria a lo largo del año, representada como una función continua para cada estación (utilizada en los modelos de regresión funcional).
- **Temperatura media anual**: Valor escalar que representa el promedio de la temperatura media diaria anual para cada estación (utilizada en el modelo de regresión lineal simple).
- **Total de precipitación anual**: Valor escalar que representa la suma total de la precipitación anual en cada estación (variable objetivo en ambos modelos).

Al incluir tanto la función de temperatura anual como la temperatura media anual en el mismo dataframe, se facilita la comparación entre los modelos de regresión funcional y la regresión lineal simple. Los modelos de regresión funcional aprovecharán la información detallada de las curvas de temperatura, mientras que el modelo de regresión lineal simple utilizará únicamente el promedio anual de temperatura. Esta configuración permitirá evaluar cómo el uso de la información funcional completa afecta la capacidad predictiva en comparación con un enfoque más simplificado.

4.2 Modelos

Para el análisis de los datos, se implementará y probará un modelo de regresión funcional en el entorno de `R`, utilizando paquetes especializados como `fda`, `fds` y

`refund`. El enfoque se centrará en la transformación de la variable de temperatura en el tiempo, utilizando tres métodos(Kokoszka & Reimherr, 2017):

- **Representación mediante bases sin penalización**: Utilizando bases funcionales, como splines o funciones de Fourier, para representar las series temporales sin aplicar penalizaciones. Este método captura la estructura de los datos mediante una combinación de funciones base.
- **Representación mediante bases con penalización**: Similar al método anterior, pero incorporando una penalización para suavizar la función estimada y evitar el sobreajuste. La penalización controla la complejidad de la función ajustada.
- **Regresión sobre Componentes Principales Funcionales**: utilizando como base la base de componentes principales permite reducir el problema a uno de regresión múltiple. Mediante FPCR se reduce la dimensión original infinita al espacio de 'p' componentes principales.

4.3 Comentarios adicionales

Para evaluar y comparar el desempeño de los modelos, se utilizarán métricas como el Error Cuadrático Medio (RMSE) y el Error Absoluto Medio (MAE), adecuadas para variables objetivo escalares para cuantificar la precisión de las predicciones realizadas.

Se realizarán procedimientos de entrenamiento y prueba mediante una división del conjunto de datos en conjuntos de entrenamiento y prueba (*train-test split*) por estación/localidad. Es decir, se utilizarán algunas estaciones para ajustar los modelos (entrenamiento) y otras para evaluar su desempeño (prueba), lo que permitirá analizar la capacidad de generalización de los modelos a nuevas localidades.

Además de los modelos mencionados, se planteará el modelo de regresión lineal para su comparación. Se evaluará cómo las distintas representaciones y metodologías impactan la precisión de las predicciones, identificando cuál ofrece mejores resultados según las métricas seleccionadas.

References

Cuevas, A., Febrero, M., & Fraiman, R. (2002). Linear functional regression: the case of fixed design and functional response. *Canadian Journal of Statistics*, *30*(2), 285–300.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis: Theory and practice*. Springer.

Gupta, A., Mall, H. K., & Janarthanan., S. (2022). *Rainfall prediction using machine learning.*

Kokoszka, P., & Reimherr, M. (2017). *Introduction to functional data analysis* (1st ed.). Taylor Francis Group.

Pan, B., Hsu, K., AghaKouchak, A., & Sorooshian, S. (2019). Improving precipitation estimation using convolutional neural network. *Water Resources Research*, *55*(3), 2301-2321. Retrieved from

https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR024090 doi: https://doi.org/10.1029/2018WR024090

Ramsay, J., & Silverman, B. (2005). *Functional data analysis* (2nd ed.). Springer.

Ramsay, J. O., Wickham, H., Graves, S., & Hooker, G. (n.d.). *fda: Functional data analysis.* R package version 6.2.0. Retrieved from https://www.rdocumentation.org/packages/fda/versions/6.2.0/topics/CanadianWeather (Accessed: 2024-11-10)

Wasko, C., & Sharma, A. (2014). Quantile regression for investigating scaling of extreme precipitation with temperature. *Water Resources Research*, *50*(4), 3608-3614. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013WR015194 doi: https://doi.org/10.1002/2013 WR015194

Yao, F., Müller, H., & Wang, J. (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, *33*(6), 2873–2903.