Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Ajustando suavemente

Introducción

El objetivo de este trabajo práctico es el desarrollo y evaluación de un algoritmo de regresión multivariado. En varias aplicaciones se utilizan algoritmos de regresión que no intentan explicar el conjunto de datos entero con una misma función, si no que van explicando pequeños subconjuntos del dataset progresivamente. El criterio para crear los diferentes subconjuntos suele tener que ver con algún tipo de definición de cercanía.

Un ejemplo clásico de aplicación para este tipo de regresiones son las series de tiempo. Estas son conjuntos de datos en donde tenemos uno o más valores medidos en función de un eje temporal que va progresando. En las series de tiempo, un tipo de algoritmo muy utilizado son los de la categoría *moving average* que no intenta explicar todo el conjunto de datos a la vez sino que va tomando diferentes ventanas de tiempo consecutivas y va ajustando la regresión paulatinamente.

En este trabajo vamos a tratar con el algoritmo *loess (locally estimated scatterplot smoothing)*. Este algoritmo se conoce coloquialmente como *regresión local* dado que utiliza una forma similar a cuadrados mínimos para explicar la localidad de cada uno de los puntos del conjunto de datos.

El algoritmo que vamos a desarrollar se encuentra introducido en [1] en su forma univariada. En este trabajo no vamos a trabajar directamente con esta publicación, sino que buscaremos replicar el trabajo en donde la regresión se presenta de manera multivariada.

La publicación que replicaremos es [2]. Dicho trabajo consiste de varias secciones. En la sección 1 y 2 se hace una introducción al tema y al método propiamente dicho. En la sección 3 se muestra una primera explicación con el objetivo de analizar el comportamiento del regresor. En la sección 4 se denotan algunas características matemáticas del método. En la sección 5 se utiliza el método sobre otro dataset y se utilizan las nociones matemáticas anteriormente presentadas. Entre las secciones 6 y 9 se presentan nuevas formas de visualizar algunas características del método y se muestran aplicaciones particulares para ejemplificar los conceptos presentados. Por último, el trabajo concluye con la sección 10 de discusión.

Para poder desarrollar el algoritmo pedido se hace especial hincapié en las secciones 1 y 2 en dónde se explica las nociones básicas del mismo. El algoritmo en sus pasos centrales utiliza una forma de cuadrados mínimos lineales que no exactamente la vista en la materia, si no que se trata de Cuadrados Mínimos Lineales Pesados (o Ponderados). Es una variante muy directa al problema visto en la materia en donde se utiliza un término más en las ecuaciones normales para poder dar diferente importancia a los diferentes puntos. Un desarrollo del algoritmo Loess, junto con el detalle en la fórmula matemática de la variación de Cuadrados Mínimos se puede encontrar en [3].

Enunciado

Se pide replicar parcialmente el trabajo [2]. En el informe se debe explicar los pasos necesarios para implementar el algoritmo propuesto, junto con las dificultades encontradas

durante su implementación.

Luego, se debe realizar una implementación computacional del método utilizando Python con las diferentes librerías que fuimos utilizando en el laboratorio como Numpy. La implementación deberá utilizarse para replicar resultados del trabajo original, así como para realizar nueva experimentación.

Una vez comprendido e implementado el método, se debe realizar experimentación para corroborar y analizar el comportamiento tanto cuantitativa como cualitativamente.

Experimentación

La experimentación de este trabajo es deliberamente más abierta que la de los trabajos anteriores. Quedará a criterio del grupo definir los ejes y métricas con los que experimentar.

Como guía de la experimentación mínima, se detallan los siguientes puntos obligatorios:

- Replicar la sección 5 de la publicación. Para esto se debe utilizar el conjunto de datos sobre cuestiones climáticas provistas con el enunciado.
- Experimentación con diversa data sintética creada por el grupo para analizar diferentes situaciones ante las que se puede presentar el método.
- Experimentar variando diferentes parámetros del algoritmo cómo:
  - Ajuste lineal o cuadrático de las variables independientes.
  - q, el tamaño del vecindario.
  - f, la proporción de datos en el vecindario.
  - Función de distancia (al menos una alternativa a la tricúbica).

Además de la experimentación, se debe desarrollar una sección que explique la utilización de las diferentes herramientas de visualización que se utilizan en el paper mencionado. Se debe investigar y explicar cada uno de los siguientes:

- *qqplot*
- *residual vs fitted*
- *component-residual plot*

Fecha de entrega

- Formato Electrónico: Jueves 23 de Junio de 2022, hasta las 23:59 hs, enviando el trabajo (informe + código) a la dirección `metnum.lab@gmail.com`. El subject del email debe comenzar con el texto `[TP3]` seguido de la lista de apellidos de los integrantes del grupo y `[Grupo N]` con el número de grupo correspondiente. Adjuntar el fichero y no linkear a carpetas compartidas. Incluir `[Grupo N]` con el número de grupo en los archivos que vayan adjuntos al correo.

Importante: El horario es estricto. Los correos recibidos después de la hora indicada no serán considerados.

Referencias

[1] Robust Locally Weighted Regression and Smoothing Scatterplots, William S. Cleveland. 1979.

[2] Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, Cleveland, William S. and Devlin, Susan J. 1988.

[3] `https://towardsdatascience.com/loess-373d43b03564`