Interpolación

Carlos Malanche

8 de marzo de 2018

Ok, ahora contamos con un arreglo de datos al que ya le hemos hecho estadística y ya nos hemos encargado de entenderlo y limpiarlo un poco. Una de las primeras cosas que nos gustaría hacer es adaptarle un modelo para poder describir con precisión lo que hace

1. Definiciones

Una serie de parejas de datos estará definida como

$$D = \{x_i, y_i\}_{i=1}^{n} \tag{1}$$

Por simplicidad, tanto x_i como y_i serán escalares (pronto usaremos vectores).

2. Interpolación

El primer paso para hacer una predicción de información es la interpolación (la cual viene con sus desventajas). Un ejemplo muy lindo de interpolación son los polinomios de lagrange.

Dada una serie de parejas de datos D, vamos a definir como su polinomio interpolante de Lagrange

$$L(x) = \sum_{i=1}^{n} \ell_i(x) y_i \tag{2}$$

aquel de grado n cuyos componentes $\ell_i(x)$ están dados por

$$\ell_i(x) = \prod_{j \neq i}^{n} \frac{x_j - x}{x_j - x_i} \tag{3}$$

Note rápidamente que

$$\ell_i(x_j) = \begin{cases} 1, & \text{si } i = j \\ 0, & \text{o.c.} \end{cases} \tag{4}$$

El resultado es una función que interpola nuestra información de manera perfecta.

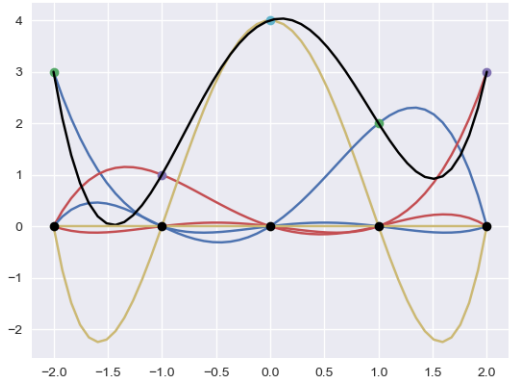


Figura 1: Polinoio de Lagrange con 5 puntos

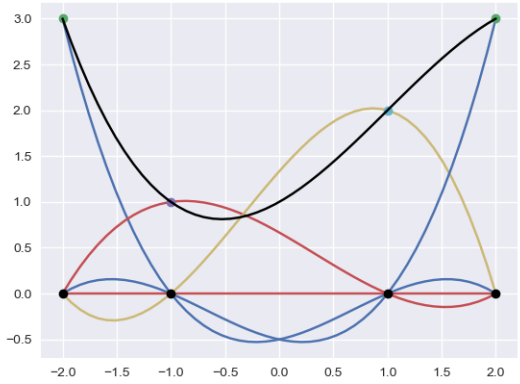
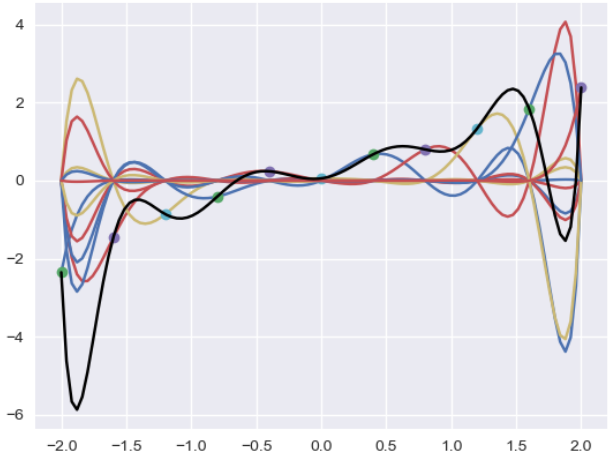


Figura 2: Polinomio de Lagrange con 4 puntos

Este resultado es demasiado inestable. Basta con retirar uno de los puntos para que los dos modelos no sean parecidos ni si quiera. A este fenómeno le llamaremos *overfitting* o sobreajuste.

El problema principal con una interpolación es que nos obligamos a tener tantos *grados de libertad* como datos tenemos. Esto convierte a nuestro modelo en uno muy complejo, lo hace dependiente de la información que le pasamos y lo hace poco flexible ante datos que aún o conocemos.



Noten que los puntos que estamos interpolando están **casi** en una línea, pero como el polinomio es de un grado muy alto comienza a dispararse cerca de las orillas.

2.1. No todo es malo

Tal y como lo he planteado, una interpolación siempre tiene desventajas. Los polinomios de Lagrange son probablemente el método de interpolación más sencillo, por lo cual no es muy recomendable usarlo pero ha aportado mucho a este campo. Una *interpolación* mucho, mucho más compleja son las series y transformada de Fourier, las cuales permiten pasar de una serie de datos equiespaciados a un conjunto de frecuencias que definen una función continua con la cual podemos hacer predicciones de valores que no se midieron en un principio.

La aplicación de esta interpolación es un tanto distinta, pero es un ejemplo en donde se asume que cada medición no contiene ruido, se asume que es una medición perfecta.

3. Una función más compleja

La interpolación es una buena solución cuando tenemos certeza de los datos con los que contamos. Ahora vamos a suponer que los datos que medimos fueron generados por una función $f(x)$ **la cual es imposible de encontrar**, y asumiremos también que en el proceso de medición hubo un poco de ruido involucrado, con lo que cualquier medición se podría describir como:

$$y = f(x) + \epsilon \tag{5}$$

donde ϵ es una variable aleatoria de distribución de gauss, centrada en cero y con varianza σ^2. Tan rápido como hemos optado por este modelo, es claro que *no tiene sentido buscar una interpolación* pues la probabilidad de que $y = f(x)$ es cero. Si no buscamos una interpolación, buscamos entonces una función estimadora $\hat{f}(x)$ que aproxime *suficientemente bien* nuestros datos...

Antes de intentar arreglar un problema, hay que entender bien el problema y definir un marco para solucionarlo. En este caso, hay varias maneras de *medir* qué tan buenas son nuestras funciones estimadoras, lo haremos por medio de funciones conocidas como *funciones de costo*, que por lo regular serán denotadas por una J. Como casi todo en este campo, no hay una regla única para optar por una función de costo, pero la gran mayoría de los métodos de aprendizaje estadístico utilizan una función de costo cuadrática (en busca de obtener un problema de optimización

convexo).

Definimos así pues, la función de pérdidas cuadrática como:

$$J(\hat{\underset{\rightarrow}{f}}) = \frac{1}{2N}\sum_{i=1}^{n} \|\hat{\underset{\rightarrow}{f}}(\underset{\rightarrow}{x}_i) - \underset{\rightarrow}{y}_i\|^2 \tag{6}$$

donde en un abuso de notación he convertido a $\hat{\underset{\rightarrow}{f}}$ en el parámetro de la función de costo.

3.1. Pequeña nota sobre la interpolación

La interpolación es un modelo de muchos parámetros, supongamos pues $\hat{\underset{\rightarrow}{f}} \coloneqq L(x)$. Resulta pues que en este modelo la varianza es muy alta respecto a la parcialidad (o el sesgo). Para ver esto, estimemos el error cuadrático generado por un dato que no está contenido en la interpolación $\{x_0, y_0\}$.

Trivialmente, $\mathrm{Var}[y_0] = \sigma^2$ (por el comportamiento determinístico de f). Haciendo un cálculo pequeño, se tiene que de la definición de la varianza de una variable aleatoria X

$$\mathrm{Var}[X] = E[X^2] - E[X]^2 \tag{7}$$

Para acortar un poquito la notación, $\hat{f}_0 \coloneqq \hat{f}(x_0)$ (lo mismo sin el gorrito).

$$\begin{aligned}
E[(y_0 - \hat{f}_0)^2] &= E[y_0^2 + \hat{f}_0^2 - 2y_0\hat{f}_0] \\
&= E[y_0^2] + E[\hat{f}_0^2] - 2E[y_0\hat{f}_0] \\
&= \mathrm{Var}[y_0] + E[y_0]^2 + \mathrm{Var}[\hat{f}_0] + E[\hat{f}_0]^2 - 2E[y_0]E[\hat{f}_0] \\
&= \mathrm{Var}[y_0] + f_0^2 + \mathrm{Var}[\hat{f}_0] + E[\hat{f}_0]^2 - 2f_0E[\hat{f}_0] \\
&= \mathrm{Var}[y_0] + \mathrm{Var}[\hat{f}_0] + (f_0^2 - 2f_0E[\hat{f}_0] + E[\hat{f}_0]^2) \\
&= \sigma^2 + \underbrace{\mathrm{Var}[\hat{f}_0]}_{\text{Varianza}} + \underbrace{(f_0 - E[\hat{f}_0])^2}_{\text{Sesgo}^2}
\end{aligned}$$

A lo mejor es un poco complicado de entenderlo así nadamás de verlo, pero imaginemos que los métodos de predicción son elegidos al azar (o más bien, sus parámetros son elegidos al azar). Resulta que el error esperado está compuesto por una parte irreducible, que es la varianza de la misma medición, y dos términos que representan la varianza de la función de predicción (se interpreta como qué tan diferentes son las funciones entre sí dadas la elección aleatoria de sus parámetros) y su sesgo (se interpreta como qué tan lejos va a estar la media de la evaluación de x_0 en $\hat{f}$ del valor sin ruido real de la medición $f(x_0)$).

Por el momento tendrán que tomar mi palabra, pero estas dos tienen una correlación negativa: si la varianza es alta, el sesgo es bajo y al revés. El método de interpolación polinómica es de varianza muy elevada, pues cada poliniomio es muy diferente con sólo modificar un punto.