Métricas, datos y calibración inteligente

L. A. Núñez
Escuela de Física, Facultad de Ciencias,

18 de agosto de 2020

1. El problema

Estamos viviendo una época de desarrollo explosivo de sensores que pueblan y generan datos en todas las facetas de nuestra cotidianidad. Estos sensores de bajo costo forman parte de dispositivos de la llamada revolución de la *Internet de las cosas, IoT*. Muchas veces estos sensores no son lo suficientemente precisos y deben ser calibrados con un patron de referencia (Para un ejemplo de este tipo de calibraciones inteligente pueden consultar [ZPK+18]). Este ejercicio busca mostrar que esa calibración está íntimamente ligada a la idea de métrica (pueden consultar [Col20]).

El problema está en cuantificar cuál es el error de medición del sensor de bajo costo y, como calibrarlo para que podamos establecer nuevas lecturas que sean mas precisas.

En el directorio del pie de página[1] podrán encontrar los datos de referencia y los de las estaciones *IoT*. El archivo *Datos Estaciones AMB* contiene las medidas de referencia de concentración de material particulado $PM_{2,5}$[2], vale decir: concentración de partículas en suspensión de dimensiones $\leq 2{,}5\mu$m. Los archivos etiquetados por *mediciones..* contiene los registro de las estaciones de bajo costo.

2. Una posible estrategia para calcular la distancia

Para comenzar es importante estimar la distancia entre las medidas de las estaciones de referencia y de las de bajo costo. Para ellos utilizamos la distancia euclidea entre las dos mediciones.

$$\mathcal{D}(\mathbb{D}_i, \hat{\mathbb{D}}_{\hat{i}}) = \sqrt{\sum_{i,\hat{i}} \left(\mathbb{D}_i - \hat{\mathbb{D}}_{\hat{i}}\right)^2} \tag{1}$$

[1] https://www.dropbox.com/sh/97lqlzsac7qpykz/AAAeA0t1PC_5eRlBCvC5f1eSa?dl=0
[2] https://blissair.com/what-is-pm-2-5.htm

Donde hemos definido como $\mathbb{D}_i = \{(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)\}$ al conjunto de datos de referencia y como $\hat{\mathbb{D}}_{\hat{i}} = \{(\hat{x}_1, \hat{y}_1), (\hat{x}_2, \hat{y}_2) \cdots (\hat{x}_m, \hat{y}_m)\}$ al conjunto de datos a calibrar.

Claramente, estamos identificando por $f(x_i) = y_i$ como el valor de la variable dependiente (en este caso la concentración de material particulado) medido por la estación patrón y las mediciones de la estación a calibrar por $\hat{f}(\hat{x}_i) = \hat{y}_i$. Adicionalmente, note que las dimensiones de los dos conjuntos de datos ($i = 1, 2, \cdots n$ e $\hat{i} = 1, 2, \cdots m$) son distintas y, que hemos denotamos por $x_i, \hat{x}_i$ las variables independientes (en este caso el tiempo) "mas cercanas" que intervienen en el cálculo de la distancia (1).

Una posible estrategia para identificar los datos "mas cercanos" es utilizar el criterio de el promedio móvil[3] y comparar los promedios locales de ambos conjuntos $f(\xi_j)$ y $\hat{f}(\xi_j)$, calculados para una ventana común $a_j \leq x_i, \hat{x}_i \leq b_j$. Donde una posible elección de ξ_j puede ser $\xi_j = a_j + (b_j - a_j)/2$, con j indica el número de ventanas a definir en el rango de variación de los datos.

Definir el ancho de la ventana para calcular los promedios locales es un arte y requiere de experimentación para balancear el tiempo de cálculo con la precisión lograda. Entonces, **calcule la distancia euclidea $\mathcal{D}(\mathbb{D}_i, \hat{\mathbb{D}}_{\hat{i}})$ para varios valores de la ventana móvil** y determine el mejor de los valores para la ventana.

3. Una posible estrategia para calibrar las mediciones

Si graficamos los puntos $\left(\hat{f}(\xi_j), f(\xi_j)\right)$, y hacemos un ajuste de mínimos cuadrados podremos determinar un modelo de ajuste lineal, $f(\xi_j) = \alpha \hat{f}(\xi_j)$. Claramente si ambas estaciones midieran lo mismo, tendríamos $\alpha = 1$.

Determine el alcance de validez del modelo lineal. Esto es, defina una tolerancia[4] y encuentre el alcance en $\hat{x}_i$ para la validez su modelo lineal.

Otra estrategia posible es dividir los conjuntos de datos por la mitad, $\mathbb{D}_i = \left\{(x_1, y_1), (x_2, y_2) \cdots (x_{\approx n/2}, y_{\approx n/2})\right\}$ y $\hat{\mathbb{D}}_{\hat{i}} = \left\{(\hat{x}_1, \hat{y}_1), (\hat{x}_2, \hat{y}_2) \cdots (\hat{x}_{\approx n/2}, \hat{y}_{\approx n/2})\right\}$, implementar el modelo lineal y comparar la predicción del modelo con las próximas mediciones de referencia $\mathbb{D}_{\approx n/2 \to n} = \left\{(x_{\approx n/2+1}, y_{\approx n/2+1}), (x_{\approx n/2+2}, y_{\approx n/2+2}) \cdots (x_n, y_n)\right\}$.

Determine cual el alcance para realizar predicciones dentro de la tolerancia.

El rango de los valores para generar el modelo (en este caso lineal) y el alcance de su predicción es un arte que debe definirse para cada conjunto de datos. Arriba se propuso utilizar la mitad del conjunto de datos para modelar, pero eso en general no es necesario. **Determine entonces, el mínimo conjunto de datos para generar el modelo y cuál será su máximo alcance para una tolerancia dada.**

[3] https://en.wikipedia.org/wiki/Moving_average

[4] Error que está dispuesto a aceptar

Referencias

[Col20] N. Colombo. Multiple metric learning for structured data. *arXiv preprint arXiv:2002.05747*, 2020.

[ZPK+18] N. Zimmerman, A. A. Presto, S. P. N. Kumar, J. Gu, A. Hauryliuk, E.S. Robinson, A.L. Robinson, and R. Subramanian. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmospheric Measurement Techniques*, 11(1):291–313, 2018.