Métodos de Validación de Modelos de Predicción de Criminología

Juan Felipe Cerón Uribe

4 de abril del 2017

A continuación presentamos una revisión de la literatura disponible al respecto de la validación de modelos de predicción de criminología. La dificultad principal que esto conlleva es que la literatura de aprendizaje no supervisado está enfocada en problemas de estimación de una distribución, mientras que el nuestro es un problema de estimación de un proceso puntual.

1. Análisis residual

De manera análoga al análisis residual en regresión lineal, una estrategia común en la literatura de predicción de procesos puntuales es examinar la diferencia entre los valores predichos y los de la muestra, donde la intensidad condicional en un punto juega el papel de la predicción [1]. Sean B una región espacio-temporal, N_B el número de eventos observados en la región y $\hat{\lambda}$ una función de intensidad acomodada, el residuo en la región se define como:

$$R(B) = N_B - \int_B \hat{\lambda}(x)dx \tag{1}$$

Algunas ideas de análisis usando este estadístico:

- Comparar la bondad de ajuste de dos modelos (se prefiere el de menor residuo).
- Se espera que R(B) sea cercano a cero.
- Graficar R(B) como función del número de iteraciones de un algoritmo y verificar que se achica.

Además de una buena derivación teórica, las secciones 11 y 12 del artículo *Residual analysis for spatial point processes*[1] describen más gráficas para el análisis de estos residuos.

1.1. Residuos en subregiones

Los autores de [5] sugieren medir el error medio de la siguiente forma: Para un periodo temporal de prueba t dividimos la región de interés en C subregiones

cuadradas. Sea y_{tc} el número de eventos observados en la región definida por t y c:

$$RMSE(t) = \sqrt{\frac{1}{C}\sum_{c=1}^{C}(y_{tc} - \hat{y}_{tc})^2} \tag{2}$$

Si se observa antisimetría en los residuos podemos reducir la variabilidad de la estimación mediante el residuo de Anscombe [4]:

$$ANSC(t) = \sqrt{\frac{1}{C}\sum_{c=1}^{C}\left(\frac{(3/2)(y_{tc}^{2/3} - \hat{y}_{tc}^{2/3})}{\hat{y}_{tc}^{1/6}}\right)^2}$$

El cual normaliza aproximadamente los residuos.

2. Residuos en predicciones

El análisis residual descrito mide el error de acomodación dentro de la muestra. Este se extiende naturalmente a una medida del error predictivo del modelo cuando la dimensión temporal de la región B no se tuvo en cuenta para la construcción del modelo. A manera de ejemplo, si se tiene un mes de datos se puede acomodar el modelo con las primeras tres semanas y analizar el residuo $R(B)$ donde B corresponde a la última semana del mes.

3. Verosimilitud

Sean $\{p_i\}_{i\leq n}$ realizaciones de un proceso puntual con intensidad condicional $\lambda(p)$. La probabilidad observar estas realizaciones (conocida como verosimilitud) está dada por:

$$L = \prod_{i=1}^{n}\lambda(p_i)$$

Si estamos comparando dos modelos, preferimos el de mayor verosimilitud. Una métrica análoga, en la cuál buscamos minimizar la información (según su definición en teoría de la información) en la muestra es la verosimilitud logarítmica:

$$l = \sum_{i=1}^{n}\log\lambda(p_i)$$

4. Análisis de hotspots

Las siguientes métricas evalúan la capacidad de un modelo de identificar **hotspots**; subregiones espaciales con una alta densidad en una ventana temporal w_t.

Para una ventana de tiempo w_t dividimos la región considerada, de área A, en subregiones cuyo tamaño depende de la granularidad del análisis deseada. Hay varias alternativas para definir qué hace a una subregión un hotspot. Siguiendo a [3] definimos un **grado de cobertura** $\frac{a}{A}$ donde a es el área cubierta por hotspots. Posteriormente elegimos las subregiones de más alta densidad/intensidad

como hotspots hasta cumplir con el grado de cobertura. Esta puede definirse siguiendo límites prácticos como la máxima cobertura que puede lograr la policía de una ciudad en la duración de w_t. Evaluamos las siguientes métricas:

4.1. Precisión

Mide la proporción de eventos capturados por hotspots en la ventana temporal w_t. Sea N_t el número total de eventos en w_t y n_t el número de eventos capturados por algún hotspot. Consideramos las siguientes métricas:

$$hit\ rate = \frac{n_t}{N_t} \tag{3}$$

esta proporción solo es comparable entre modelos con el mismo grado de cobertura. Alternativamente medimos el *Predictied Accuracy Index*:

$$PAI = \frac{n_t}{N_t} / \frac{a}{A} \tag{4}$$

el cuál puede interpretarse como la proporción de densidad de eventos en los hotspots sobre densidad en la región.

Una vez evaluadas estas métricas en las predicciones de dos modelos en ventanas de tiempo $w_t : t \in 1, 2, .., m\}$ obtenemos una sucesión de diferencias entre los hit rates (podría usarse PAI) $\{d_t = \frac{n_{1t}}{N_t} - \frac{n_{2t}}{N_t}\}$. Podemos utilizar el promedio de esta sucesión para decidir sobre los modelos o comparar las sucesiones originales analíticamente. Alternativamente, si asumimos que la diferencia entre el poder predictivo de los modelos es independiente de la intensidad de crimen en la ventana de tiempo los $\{d_t\}_{t \leq m}$ son una muestra iid y podemos probar paramétricamente si la variable es significativamente diferente de 0.

4.2. Compacidad

Define el nivel de agrupamiento de los hotspots generados por un modelo predictivo. Esta es una medida práctica de la dificultad de patrullar los hotspots identificados. Una medida clásica es la razón *área/perímetro*, sin embargo esta depende de la escala de la región, lo cuál la hace incomparable entre regiones. Los autores de [3] proponen medir el *Clumpiness Index*. Este mide la desviación proporcional de adyacencias hotspot-hotspot (comunes) en la región de la esperada bajo una distribución aleatoria. Definimos:

- g_{ij} := Número de adyacencias entre una subregión de la clase i y una de la clase j ($g_{ij} = g_{ji}$). La clase 1 representa a los hotspots, la clase 2 los no hotspots.
- p_i := Perímetro de la figura formada por la clase i cuando esta está maximalmente agrupada.
- $G_1 := \frac{g_{11}}{g_{11}+g_{12}-p_1}$
- P_1 :=Proporción de la región ocupada por la clase i.

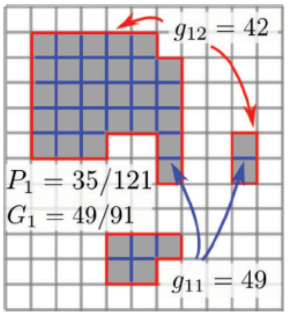


Figura 1: Cálculo del CI en [3]

Veamos un ejemplo de estas definiciones:

$$CI = \begin{cases} \frac{G_1 - P_1}{P_1} \text{ si } P_1 > 0{,}5, \\ \frac{G_1 - P_1}{1 - P_1} \end{cases} \tag{5}$$

$CI = -1$ cuando los hotspots están maximalmente desagregados, y $CI = 1$ cuando están maximalmente agrupados. Se define por casos para que lo anterior se cumpla:

- Si $P_1 \leq 0{,}5$, $G_1 = 0$ cuando hay desagregación maximal, pues los hotspots se pueden esparcir de modo que no queden hotspots adyacentes. $G_1 = 1$ bajo agrupamiento maximal.
- Si $P_1 > 0{,}5$, $G_1 = 2P_i - 1$ bajo desagregación y $G_1 \to 1$ bajo agrupamiento maximal.

4.3. Variabilidad dinámica

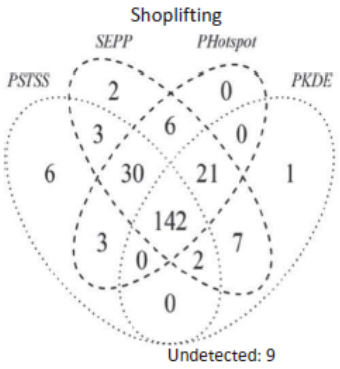
Mide la variación en la distribución espacial de los hotspots entre ventanas de tiempo consecutivas. Considere los mapas predictivos generados para dos ventanas de tiempo consecutivas w_t y w_{t+1}. Tenemos r (repitentes) hotspots que aparecen en los dos mapas, e (emergentes) que aparecen en w_{t+1} pero no en w_t y d (desaparecidos) que solo aparecen en w_t. Definimos el *Dynamic Variability Index* en t como:

$$DVI = \frac{E}{E + R} \tag{6}$$

tomamos la medida para cada par de ventanas temporales consecutivas.

4.4. Complementareidad

Método analítico que busca detectar el grado en el cuál varios modelos predictivos detectan los mismos eventos de una base de datos. Los autores de [3] utilizan un diagrama de Venn como el siguiente:



5. Simulación de un proceso ETAS

Esta metodología asume que los eventos de criminología tienen un comportamiento similar al modelo ETAS para validar la convergencia correcta de un algoritmo de estimación. En este sentido ofrece condiciones necesarias, mas no suficientes, para concluir la bondad de un modelo ETAS ajustado (**no es muy relevante para otros tipos de modelos**). Recordamos que la metodología de Mohler pretende acomodar los eventos al modelo puntual:

$$\lambda(t,x,y) = \nu(t)\mu(x,y) + \sum_{k:t_k<t} g(t-t_k, x-x_k, y-y_k) \tag{7}$$

En el mismo artículo (Self-Exciting Point Process Modeling of Crime) propone una metodología de validación basada en la simulación de eventos según la superposición de un proceso de Poisson arbitrario $\nu\mu$ y un proceso definido por g desencadenado por cada evento ocurrido en la simulación. Estas simulaciones pueden implementarse aprovechando el hecho de que en un proceso de Poisson el tiempo hasta el siguiente evento tiene una distribución exponencial [2].

Luego de ejecutar el modelo de aprendizaje sobre los datos analiza el error L2 $\|P_n - P_{n-1}\|$ en las n-ésimas iteraciones. Se espera que este decrezca y converja a 0. También se compara esta matriz en alguna iteración (con n alto) con la matriz teórica de probabilidades de causalidad que proviene de la simulación. Finalmente compara el número de eventos de transfondo determinado por el algoritmo con el observado en la simulación, además de su convergencia tras varias iteraciones.

Teorema del re-escalamiento temporal

Una prueba de bondad de ajuste se deriva del siguiente teorema:

Teorema del re-escalamiento temporal: Sean $u_1, u_2, ..., u_n < T$ realizaciones de un proceso puntual con intensidad condicional λ. Defina la transformación:

$$A(u_k) = \int_0^{u_k} \lambda(u|H_u)du \tag{8}$$

Los $A(u_k)$'s son un proceso de Poisson con parámetro 1.

Recordamos que los tiempos τ_k entre eventos en un proceso de Poisson son variables aleatorias exponenciales independientes. Haciendo la transformación $z_k = 1 - exp(-\tau_k)$ los z_k son uniformes en el intervalo $(0, 1)$. Dado que todas las transformaciones propuestas son uno a uno, la acomodación de los z_k a una variable uniforme evalúa directamente el supuesto del teorema; que λ describe el proceso original. Esto último puede probarse con una prueba paramétrica Kolmogorov-Smirnov o analíticamente mediante una gráfica Q-Q.

Referencias

[1] J. Moller M. Hazelton A. Baddeley R. Turner. "Residual analysis for spatial point processes". En: *Journal of the Royal Statistical Society* (2005).

[2] Patrick McQuighan. "Simulating the Poisson Process". En: *VIGRE: University of Chicago* (2010).

[3] "Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions - a crime case study, author=M. Adepeju, G. Rosser and T. Cheng, journal=International Journal of Geographical Information Science, volume=30, number=11, pages=2133-2154, year=2016". En: ().

[4] J.A. Nelder P. McCullagh. *Generalized Linear Models*. Chapman y Hall, 1989.

[5] "Predicting Melbourne ambulance demand using kernel warping, author=Z. Zhou, D. Matteson, publisher=Cornell University, year=2015". En: ().