Aprendizaje de máquinas: Sesgos y Retroalimentación circular

Alvaro J. Riascos Villegas
Universidad de los Andes y Quantil

30 de junio de 2020

Contenido

Outline

Introducción

- El aprendizaje de maquinas se utiliza de forma generalizada para resolver problemas de predicción en muchas circunstancias de la vida humana: otorgamiento de credito, predicción de crimen, alertas de desempeño académico, transacciones atipicas en el sector financiero, etc.

Introducción

- El aprendizaje de maquinas se utiliza de forma generalizada para resolver problemas de predicción en muchas circunstancias de la vida humana: otorgamiento de credito, predicción de crimen, alertas de desempeño académico, transacciones atipicas en el sector financiero, etc.
- En todos estos casos surgen problemas normativos:
  - La predicción esta correlacionada con el instrumento de recolección de información (input del modelo). Este fenómeno se conoce como retroalimentación circular o *feeback loop*: genera potenciales sesgos y resultados disciminatorios en una población.
  - Los modelos de ML infieren información de variables no observadas que normativamente no deben ser utilizadas en los modelos. Este fenómeno se llama triangulación: potencialmente genera resultados discriminatorios.

Otros problemas

- Sub reporte: los datos son una muestra sesgada de la realidad.
- Equilibrio: las predicciones deconocen que pueden existir interacciones estratégicas.
- Privacidad: con suficiente información es posible identificar las identidades.

Outline

Problema

- Basdo en To predict and serve? http://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2016.00960.x

Problema

- Basdo en To predict and serve? `http://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2016.00960.x`
- Qué tan sesgado son los datos recoletados por la policia?.

Poblaciones sintéticas y verdadera distribución

- Solución: Combinar imformación demográfica de Oakland, California con datos de la encuesta 2011: National Survey on Drug Use and Health (NSDUH).
- De esta forma se obtiene información sintética de alta resolución para Oakland, que estima el uso ilegal de drogas que después se puede comparar con los datos de la policía.

Poblaciones sintéticas y verdadera distribución

- A synthetic population is a demographically accurate individual-level representation of a real population – in this case, the residents of the city of Oakland.

Poblaciones sintéticas y verdadera distribución

- A synthetic population is a demographically accurate individual-level representation of a real population – in this case, the residents of the city of Oakland.
- Individuals in the synthetic population are labeled with their **sex, household income, age, race, and the geo-coordinates of their home**.

Poblaciones sintéticas y verdadera distribución

- A synthetic population is a demographically accurate individual-level representation of a real population – in this case, the residents of the city of Oakland.
- Individuals in the synthetic population are labeled with their **sex, household income, age, race, and the geo-coordinates of their home**.
- These characteristics are assigned so that the demographic characteristics in the synthetic population match data from the US Census at the highest geographic resolution possible.

Poblaciones sintéticas y verdadera distribución

- Se estima un modelo con los datos de la encuesta NSDUH para predecir la probabilidad de consumo de drogas en el último mes basado en características demográficas (i.e. sex, household income, age, and race).

Poblaciones sintéticas y verdadera distribución

- Se estima un modelo con los datos de la encuesta NSDUH para predecir la probabilidad de consumo de drogas en el último mes basado en características demográficas (i.e. sex, household income, age, and race).
- Aplicar este modelo en la población sintética para obtener una estimación de la probabilidad de uso de drogas para cada individuo de la población sintética.

Poblaciones sintéticas y verdadera distribución

- Se estima un modelo con los datos de la encuesta NSDUH para predecir la probabilidad de consumo de drogas en el último mes basado en características demográficas (i.e. sex, household income, age, and race).
- Aplicar este modelo en la población sintética para obtener una estimación de la probabilidad de uso de drogas para cada individuo de la población sintética.
- These estimates are based on the assumption that the relationship between drug use and demographic characteristics is the same at the national level as it is in Oakland.

Synthetic population and estimation of ground truth

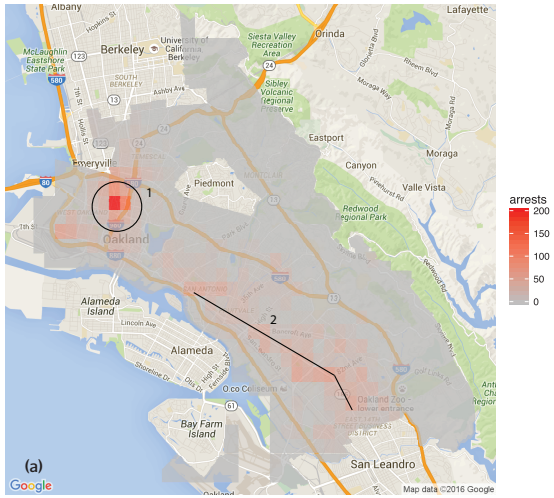


FIGURE 1 (a) Number of drug arrests made by Oakland police department, 2010. (1) West Oakland, (2) International Boulevard. (b) Estimated number of drug users, based on 2011 National Survey on Drug Use and Health

Figura 1: XXX

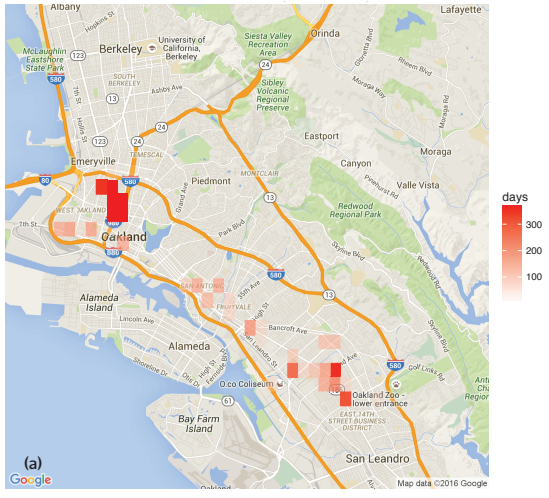
Consecuencias policiamiento predictivo

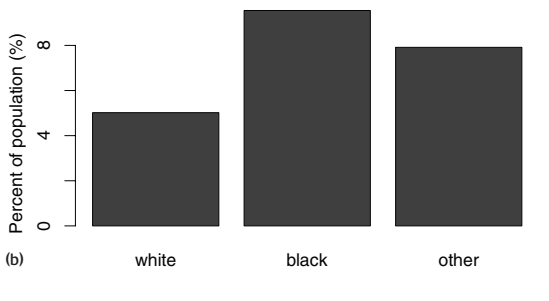
- Se aplica el modelo de PredPol a Oakland con el fin de predecir el uso de drogas ilícitas.

Consecuencias policiamiento predictivo

- Se aplica el modelo de PredPol a Oakland con el fin de predecir el uso de drogas ilícitas.
- Los creadores de PredPol describen el modelo como: only three data points in making predictions: past type of crime, place of crime and time of crime. It uses no personal information about individuals or groups of individuals, eliminating any personal liberties and profiling concerns.

Consecuencias policiamiento predictivo





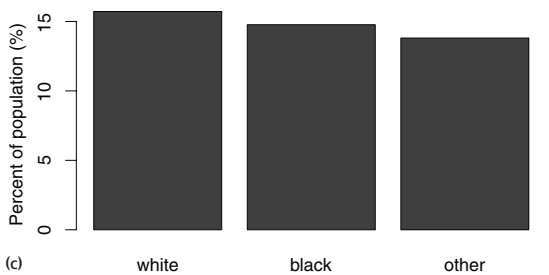


FIGURE 2 (a) Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland police data. (b) Targeted policing for drug crimes, by race. (c) Estimated drug use by race

Sesgo y retroalimentación

- The results so far rely on one implicit assumption: that the presence of additional policing in a location does not change the number of crimes that are discovered in that location.

Sesgo y retroalimentación

- The results so far rely on one implicit assumption: that the presence of additional policing in a location does not change the number of crimes that are discovered in that location.
- But what if police officers have incentives to increase their productivity as a result of either internal or external demands?

Sesgo y retroalimentación

- The results so far rely on one implicit assumption: that the presence of additional policing in a location does not change the number of crimes that are discovered in that location.
- But what if police officers have incentives to increase their productivity as a result of either internal or external demands?
- We can investigate the consequences of this scenario through simulation. For each day of 2011, we assign targeted policing according to the PredPol algorithm. In each location where targeted policing is sent, we increase the number of crimes observed by 20

Sesgo y retroalimentación

- These additional simulated crimes then become part of the data set that is fed into PredPol on subsequent days and are factored into future forecasts.
- We study this phenomenon by considering the ratio of the predicted daily crime rate for targeted locations to that for non-targeted locations. This is shown in the next Figure,

Sesgo y retroalimentación

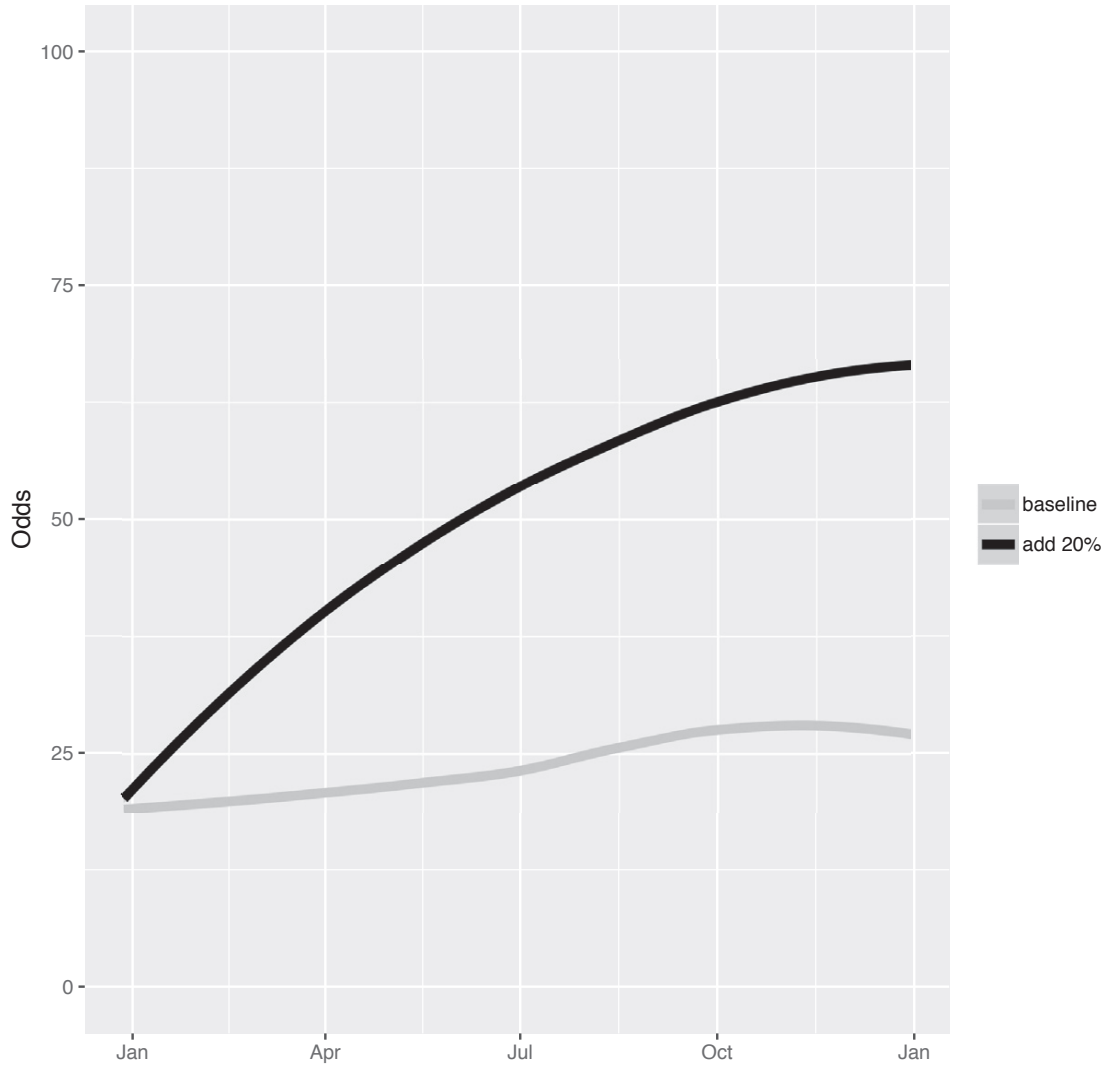


FIGURE 3 Predicted odds of crime in locations targeted by PredPol algorithm, relative to non-targeted locations. 'Baseline' is original Oakland police data. 'Add 20%' simulates the effect of additional crimes being observed in targeted locations

Outline

Modelo de policiamiento predictivo

- Runaway Feedback Loops in Predictive Policing: `https://arxiv.org/abs/1706.09847`

Modelo de policiamiento predictivo

- Runaway Feedback Loops in Predictive Policing: `https://arxiv.org/abs/1706.09847`

Definition

Dados históricos de incidentes de crimen en un conjunto de regiones, decidir como asignar la fuerza policial a algunas areas con el fin de detectar incidentes de crimen.

- Vamos a utilizar la teoría de urnas para mostrar por qué algunos métodos populares estan destinados a fracasar y como mitigar los problemas identificados.

Modelo Espacio - Temporal: especificación

El modelo de predicción de crimen se basa en tres supuestos:

Modelo Espacio - Temporal: especificación

El modelo de predicción de crimen se basa en tres supuestos:

1. Los crímenes se concentran en lugares específicos de la ciudad:

Modelo Espacio - Temporal: especificación

El modelo de predicción de crimen se basa en tres supuestos:

1. Los crímenes se concentran en lugares específicos de la ciudad:
   - Cerca a estaciones de transmilenio.

Modelo Espacio - Temporal: especificación

El modelo de predicción de crimen se basa en tres supuestos:

1. Los crímenes se concentran en lugares específicos de la ciudad:
   - Cerca a estaciones de transmilenio.
   - Cerca a zonas de entretenimiento (bares, discotecas, etc.).

Modelo Espacio - Temporal: especificación

El modelo de predicción de crimen se basa en tres supuestos:

1. Los crímenes se concentran en lugares específicos de la ciudad:
   - Cerca a estaciones de transmilenio.
   - Cerca a zonas de entretenimiento (bares, discotecas, etc.).
2. Ocurren más crímenes en ciertas horas:

Modelo Espacio - Temporal: especificación

El modelo de predicción de crimen se basa en tres supuestos:

1. Los crímenes se concentran en lugares específicos de la ciudad:
   - Cerca a estaciones de transmilenio.
   - Cerca a zonas de entretenimiento (bares, discotecas, etc.).

2. Ocurren más crímenes en ciertas horas:
   - Mayor criminalidad a las 10:00pm que a las 11:00am.

Modelo Espacio - Temporal: especificación

El modelo de predicción de crimen se basa en tres supuestos:

1. Los crímenes se concentran en lugares específicos de la ciudad:
   - Cerca a estaciones de transmilenio.
   - Cerca a zonas de entretenimiento (bares, discotecas, etc.).
2. Ocurren más crímenes en ciertas horas:
   - Mayor criminalidad a las 10:00pm que a las 11:00am.
3. El crimen se propaga como si fuera una enfermedad (en espacio y tiempo):

Modelo Espacio - Temporal: especificación

El modelo de predicción de crimen se basa en tres supuestos:

1. Los crímenes se concentran en lugares específicos de la ciudad:
   - Cerca a estaciones de transmilenio.
   - Cerca a zonas de entretenimiento (bares, discotecas, etc.).
2. Ocurren más crímenes en ciertas horas:
   - Mayor criminalidad a las 10:00pm que a las 11:00am.
3. El crimen se propaga como si fuera una enfermedad (en espacio y tiempo):
   - Cerca espacialmente de donde se han presentado crímenes anteriormente, es más probable que ocurran nuevos crímenes.
   - Cerca temporalmente es también más probable.

Modelo Espacio - Temporal: especificación

De esta manera, un crimen puede ocurrir por dos razones:

Modelo Espacio - Temporal: especificación

De esta manera, un crimen puede ocurrir por dos razones:

- Ser el primero de una nueva oleada de crímenes, dados su ubicación y un momento de la semana.

Modelo Espacio - Temporal: especificación

De esta manera, un crimen puede ocurrir por dos razones:

- Ser el primero de una nueva oleada de crímenes, dados su ubicación y un momento de la semana.
- Como réplica de un crimen anterior cercano.

Modelo Espacio - Temporal: especificación

Con estas consideraciones estimamos la siguiente función de intensidad:

$$\lambda(x, y, t)$$

Modelo Espacio - Temporal: especificación

Con estas consideraciones estimamos la siguiente función de intensidad:

$$\lambda(x, y, t) = \underbrace{\mu(x, y)}_{\substack{\text{patrones} \\ \text{espaciales}}}$$

Modelo Espacio - Temporal: especificación

Con estas consideraciones estimamos la siguiente función de intensidad:

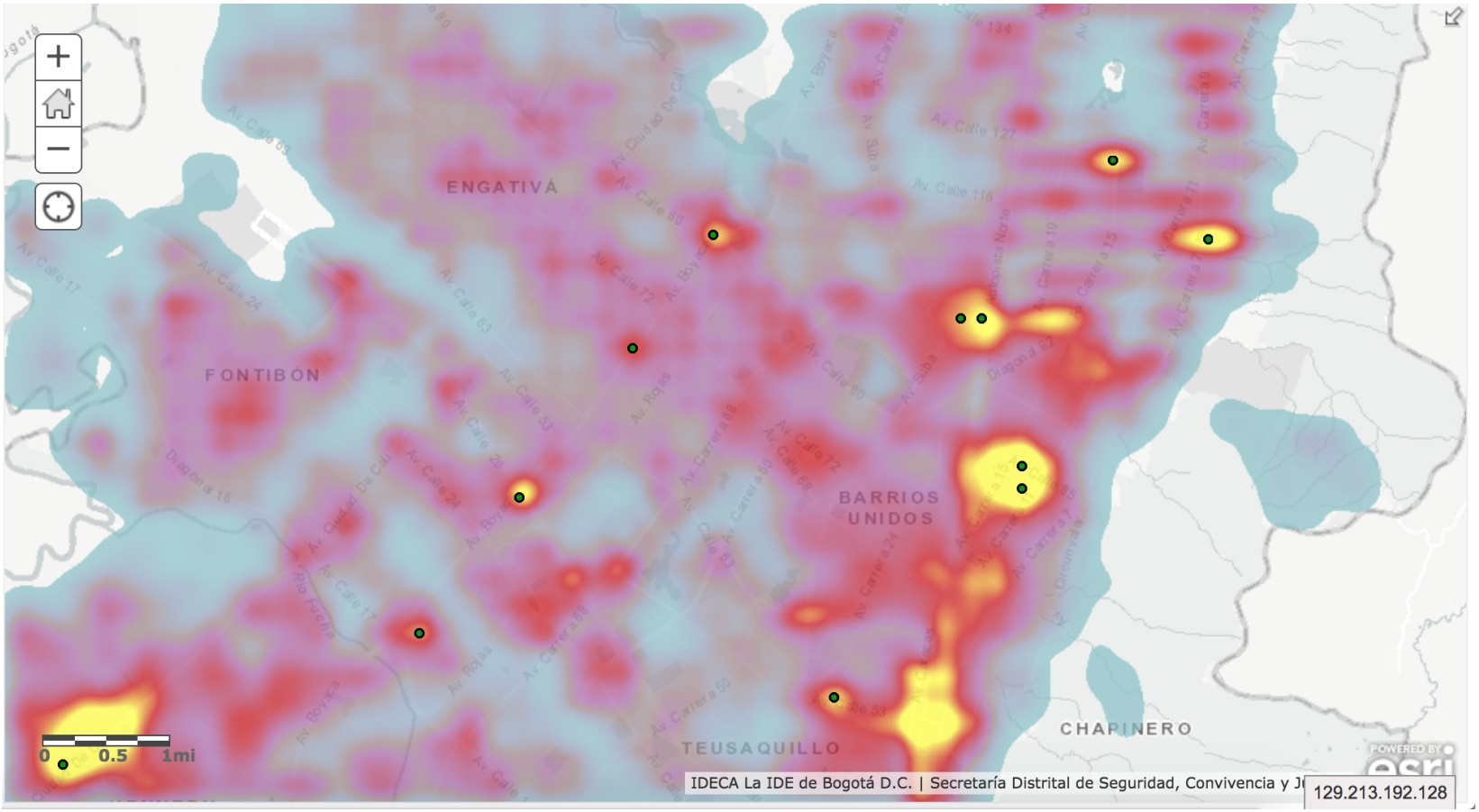
$$\lambda(x, y, t) = \underbrace{\mu(x, y)}_{\substack{\text{patrones} \\ \text{espaciales}}} * \underbrace{\nu(t)}_{\substack{\text{patrones} \\ \text{temporales}}}$$

Modelo Espacio - Temporal: especificación

Con estas consideraciones estimamos la siguiente función de intensidad:

$$\lambda(x,y,t) = \underbrace{\mu(x,y)}_{\substack{\text{patrones} \\ \text{espaciales}}} * \underbrace{\nu(t)}_{\substack{\text{patrones} \\ \text{temporales}}} + \sum_{\substack{\text{crímenes} \\ \text{anteriores}}} \underbrace{g(x-x_k, y-y_k, t-t_k)}_{\substack{\text{patrones de} \\ \text{réplicas}}}$$

Modelo Espacio - Temporal: resultados y visualización



Modelo de urnas

- Vamos a suponer que solo hay dos regiones: A y B.

Modelo de urnas

- Vamos a suponer que solo hay dos regiones: A y B.
- Cada día un policia es enviada a una de las dos regiones.

Modelo de urnas

- Vamos a suponer que solo hay dos regiones: A y B.
- Cada día un policia es enviada a una de las dos regiones.
- Puede descubrir (y registrar) un crimen con cierta probabilidad (la verdadera probabilidad de incidencia del crimen).
- Otros registros pueden ser reportados por terceros.

Modelo de urnas

- Vamos a suponer que solo hay dos regiones: A y B.
- Cada día un policia es enviada a una de las dos regiones.
- Puede descubrir (y registrar) un crimen con cierta probabilidad (la verdadera probabilidad de incidencia del crimen).
- Otros registros pueden ser reportados por terceros.

Definition

Decimos que el policiamiento es efectivo cuando una región con una proporción λ del crimen recibe una proporción λ del policiamiento.

- La verdadera proporción en A y B es: λ_A, λ_B.

Hipótesis

- El planificador utiliza los estadísticos actuales del crimen en cada región para decidir como asignar su fuerza policial.

Hipótesis

- El planificador utiliza los estadísticos actuales del crimen en cada región para decidir como asignar su fuerza policial.
- Contexto: La única información que se registra sobre el crimen es el conteo.

Hipótesis

- El planificador utiliza los estadísticos actuales del crimen en cada región para decidir como asignar su fuerza policial.
- Contexto: La única información que se registra sobre el crimen es el conteo.
- La verdadera distribución del crimen es revelada por los crimenes descubiertos que tienen un peso w_d y los reportados que tiene un peso w_r: $w_d + w_r = 1$ (i.e., proporción de registros que viene de una u otra fuente).

Hipótesis

- El planificador utiliza los estadísticos actuales del crimen en cada región para decidir como asignar su fuerza policial.
- Contexto: La única información que se registra sobre el crimen es el conteo.
- La verdadera distribución del crimen es revelada por los crimenes descubiertos que tienen un peso w_d y los reportados que tiene un peso w_r: $w_d + w_r = 1$ (i.e., proporción de registros que viene de una u otra fuente).
- Sean d_A la tasa de descubrimiento de la policia en A y r_A la tasa de reportes en A.

Urna de Polya generalizada

- Vamos a considerar una matriz de reemplazo de la siguiente forma:

$$M = \begin{bmatrix} & \text{Adición R} & \text{Adición B} \\ \text{Muestra R} & a & b \\ \text{Muestra B} & c & d \end{bmatrix}$$

Si se muestrea una bola R, se reemplaza y se añaden a bolas R y b bolas B. Una urna de Polya estándar es $a = d, b = c = 0$.

Proceso

1. En una urna se cuenta con n_r, n_b bolas rojas y negras, el estimativo actual del crimen en la regiones A y B.
2. Sacar bola de la urna para decidir qué región visitar.
3. Visitar región y lanzar una moneda sesgada que cae cara con probabilidad la verdadera probabilidad del crimen en esa región (i.e., λ_A o λ_B).
4. Aplicar actulización usando urna generalizada de Polya.
5. Repetir.

La esperanza es que, asintóticamente, la proporción de bolas converja a las verdaderas probabilidad del crimen en cada región. Por ejemplo que en la región A:

$$\frac{n_r(t)}{n_r(t) + n_b(t)} \to \frac{\lambda_A}{\lambda_A + \lambda_B}$$

Solo crimenes descubiertos - Distribución uniforme del crimen

- Los incidentes son solo descubiertos. Si no hay presencia policial en una region no hay reportes: $w_d = 1$.
- Si un policia va a cualquiera de la regiones en ambas el crimen se descubre con probabilidad $\lambda = \lambda_A = \lambda_B$

Solo crimenes descubiertos - Distribución uniforme del crimen

- Los incidentes son solo descubiertos. Si no hay presencia policial en una region no hay reportes: $w_d = 1$.
- Si un policia va a cualquiera de la regiones en ambas el crimen se descubre con probabilidad $\lambda = \lambda_A = \lambda_B$

Theorem (Renlund)

Supongamos que aplicamos el anterior proceso con una urna de Polya estándar ($a = c = \lambda$). Entonces la distribución de probabiliad asintótica de una bola roja es Beta(n_r, n_b)

Solo crimenes descubiertos - Distribución uniforme del crimen

- Los incidentes son solo descubiertos. Si no hay presencia policial en una region no hay reportes: $w_d = 1$.
- Si un policia va a cualquiera de la regiones en ambas el crimen se descubre con probabilidad $\lambda = \lambda_A = \lambda_B$

Theorem (Renlund)

Supongamos que aplicamos el anterior proceso con una urna de Polya estándar ($a = c = \lambda$). Entonces la distribución de probabiliad asintótica de una bola roja es Beta(n_r, n_b)

- Obśervese que no se aprende nada.

Solo crimenes descubiertos - Distribución uniforme del crimen

- Los incidentes son solo descubiertos. Si no hay presencia policial en una region no hay reportes: $w_d = 1$.
- Si un policia va a cualquiera de la regiones en ambas el crimen se descubre con probabilidad $\lambda = \lambda_A = \lambda_B$

Theorem (Renlund)

Supongamos que aplicamos el anterior proceso con una urna de Polya estándar ($a = c = \lambda$). Entonces la distribución de probabiliad asintótica de una bola roja es Beta(n_r, n_b)

- Obśervese que no se aprende nada.
- En ausencia de un modelo para asignar la policia, simplemente visitando con la misma probabilidad cada región, se descubriría la verdadera distribución del crimen.

Solo crimenes descubiertos - Distribución no uniforme del crimen

- Los incidentes son solo descubiertos. Si no hay presencia policial en una region no hay reportes: $w_d = 1$.
- $\lambda_A \neq \lambda_B$.
- El modelo de urna de Polya es ahora:

$$M = \begin{bmatrix} & \text{Adición R} & \text{Adición N} \\ \text{Muestra R} & \lambda_A & 0 \\ \text{Muestra N} & 0 & \lambda_B \end{bmatrix}$$

Solo crimenes descubiertos - Distribución no uniforme del crimen

Theorem (Renlund)

En una urna de Polya con matriz de adición como la anterior, la distribución asíntotica de las bolas rojas es 1 si $\lambda_A > \lambda_B$.

- Obśervese que el proceso tiende a hacer una representación completamente errada de la verdadera distribución de los eventos en las regiones.

Distribución arbitraria del crimen

- El modelo de urna de Polya es ahora:

$$M = \begin{bmatrix} & \text{Adición R} & \text{Adición N} \\ \text{Muestra R} & w_d d_A + w_r r_A & w_r r_B \\ \text{Muestra N} & w_r r_A & w_d d_B + w_r r_B \end{bmatrix}$$

- En este caso el resultado principal es que la única forma de que se descubra la verdadera distribución es ignorando los eventos descubiertos $w_d = 0$ (i.e., sin retroalimentación) o que $d_A \approx d_B$.

Conclusión

- El modelo de urnas sugiere que la presencia de retroalimentación impide descubrir la verdadera distribución.
- Peor aún, puede llevar a resultados completamente errados.

Corrigiendo el sesgo: Solo crimenes descubiertos

- En el modelo de Polya:

$$M = \begin{bmatrix} & \text{Adición R} & \text{Adición B} \\ \text{Muestra R} & \lambda_A & 0 \\ \text{Muestra B} & 0 & \lambda_B \end{bmatrix}$$

λ_A representa la probabilidad de adicionar una bola roja condicional a que se elgió inicialmente en el proceso un bola roja.

Corrigiendo el sesgo

- La probabildad no condicional de adicionar una bola roja es:

$$\lambda_A \frac{n_R}{n_R + n_B}$$

Corrigiendo el sesgo

- La probabildad no condicional de adicionar una bola roja es:

$$\lambda_A \frac{n_R}{n_R + n_B}$$

- Luego despues de un paso de usar el algoritmo la fracción de bolar rojas que se añaden es:

$$\frac{\lambda_A n_R}{\lambda_A n_R + \lambda_B n_B}$$

para anular el sesgo seria suficiente que esta probabilidad fuera: $\frac{\lambda_A}{\lambda_A + \lambda_B}$

Corrigiendo el sesgo

- La probabildad no condicional de adicionar una bola roja es:

$$\lambda_A \frac{n_R}{n_R + n_B}$$

- Luego despues de un paso de usar el algoritmo la fracción de bolar rojas que se añaden es:

$$\frac{\lambda_A n_R}{\lambda_A n_R + \lambda_B n_B}$$

  para anular el sesgo seria suficiente que esta probabilidad fuera: $\frac{\lambda_A}{\lambda_A + \lambda_B}$
- Esto sugiere que una forma de solucionar el problema es antes de adicionar una bola, sacar de nuevo una bola de la urna y adicionar solo si es de un color distinto. En este caso la probabilidad no condicional es:

$$\lambda_A \frac{n_R}{n_R + n_B} \frac{n_B}{n_R + n_B}$$

Corrigiendo el sesgo

- En regiones donde la probabilidadde enviar a un policía es el 95 % es de esperar que se descubran muchos crimenes. Entonces solamente regitrarlo el 5 % de las veces.
- Este procedimiento es una versión del estimador de Thompson-Horvitz en teoría de muestreo.

Simulación urnas con datos realistas

- Verdadera distribución datos consumo drogas. Modelo de predicción proporciones históricas.

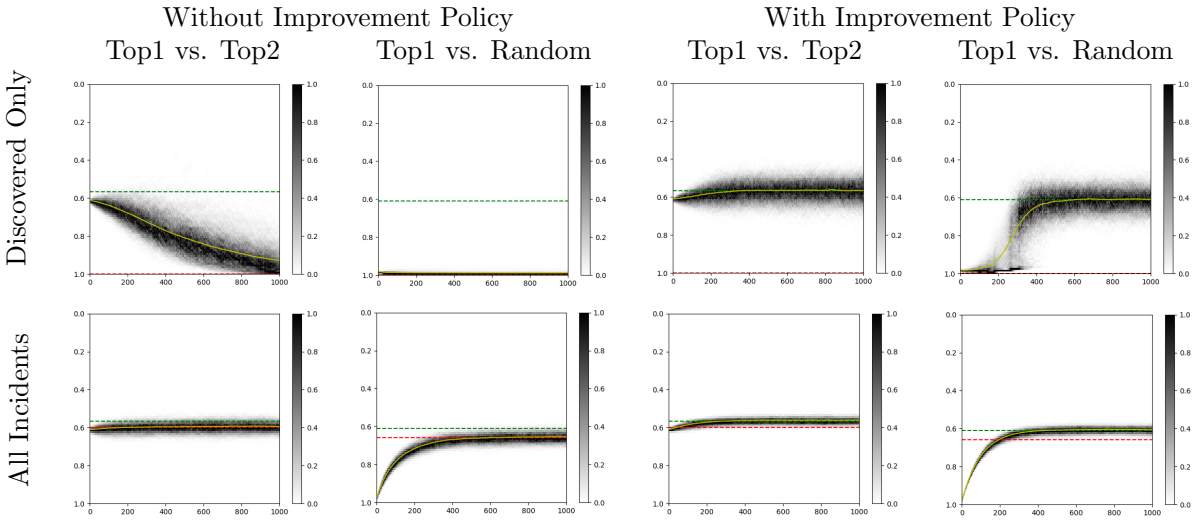


Figure 1: The distribution over 1000 days versus percentage of balls from region $Top1$ in the urn over 1000 runs. A police force deployed based on the underlying crime rates would send 56.7% of the force to $Top1$ instead of $Top2$ and 61.0% of the force to $Top1$ instead of *Random* (the green line shown). Top row (discovered incidents only): both charts (left) converge to sending 100% of the force to $Top1$, while with the improvement policy (right) the charts appear to converge to the correct crime rates. Bottom row (all incidents, equally weighted): both charts (left) converge to the incorrect rate (red line), while with the improvement policy (right) the charts appear to converge correctly to the true crime rates.

Simulación PredPol

- Verdadera distribución datos consumo drogas. Modelo de predicción PredPol.

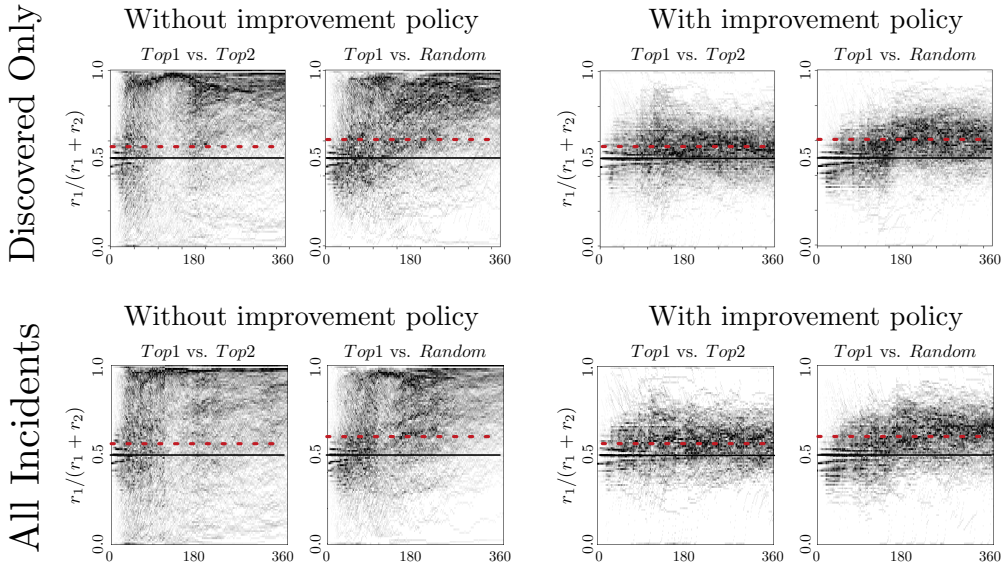


Figure 2: PredPol's relative deployment to region *Top*1 versus *Top*2 or *Random*. Along the top row, we use the model which only accounts for discovered incidents (those based on police having been deployed to an area). Along the bottom row, we use the model which accounts for both discovered and reported incidents. Left: PredPol operating as usual. Right: discovered incident entries modified using our improvement policy. Police deployment based on underlying crime rates would send 56.7% of the force to *Top*1 instead of *Top*2 and 61.0% of the force to *Top*1 instead of *Random*. These correct crime rates (indicated with a dashed red line) appear to be what PredPol converges to under the improvement policy.