Análisis de Datos

Carlos Malanche

15 de febrero de 2018

1. Qué es el análisis de datos

Directamente sacado de Wikipedia, el análisis de datos consiste en tomar datos recabados de análisis estadísticos o de experimentos para inspeccionarlos, transformarlos, limpiarlos y eventualmente utilizarlos para generar conclusiones y modelos, y posiblemente comenzar a realizar predicciones del sistema del que se han obtenido los datos.

Últimamente, se ha puesto de moda el término *data science*, que engloba más etapas del proceso descrito anteriormente. La ciencia de datos incluye la manera de recabar los mismos, pues no siempre se encuentran disponibles (en el ejemplo de un experimento, donde tenemos un conjunto de sensores y nuestro proceso de captura está determinado). Al proceso de recaudar información que **ya existe** pero no se encuentra **orddenada** se le llama *minería de datos*. Este paso no existe en un experimento científico.

1.1. Un ejemplo tonto

La *minería de datos* es un proceso complicado pues se trata de darle más utilidad a la información ya existente. Un ejemplo podría ser el siguiente:

Imagine que se han realizado múltiples mediciones para determinar si el calentamiento global en verdad está ocurriendo como una consecuencia de las emisiones de ciertos gases. Por cuestiones históricas, cada laboratorio ha decidido tomar mediciones de la concentración en el aire de ciertos gases. Digamos pues, algunos laboratorios han medido Metano (CH_4) y Ozono (O_3), otros decidieron medir vapor de agua y óxido de nitrógeno N_2O y la gran mayoría se ha dedicado a medir dióxido de carbono (CO_2). Pues bien, es posible que por condiciones atmosféricas relacionadas a la región, estas medidas de manera aislada no tienen mucho valor. Sin embargo, un científico puede darse a la tarea de recolectar la información de todos los laboratorios que han hecho estas medidas, localizarlas geográficamente y temporalmente (hay registros históricos de estos gases desde hace años, y la meta es observar una tendencia de crecimiento tanto en las concentraciones de dichos gases como en la temperatura promedio de varias regiones del planeta) para comenzar a crear un modelo mucho más completo, en donde puede incluír información de otra índole (mediciones de velocidad del aire para estimar la difusión de ciertos gases, incluso información un tanto menos directa o no considerada en los estudios originales como la actividad solar y la densidad de algas en varios puntos del oceano que capturan el dióxido de carbono). Esta tarea **es** de los científicos y no es una tarea fácil. El problema de estos datos es que son **demasiados** y que se encuentran escritos para humanos, no *para máquinas*; es decir, la información no está en el formato que un programa fácilmente podría entender.

Esto ni si quiera es un problema que se pueda solucionar estableciendo un estandar para la descripción de la información en la investigación científica. Sería *irrealista* esperar que todos sigan un formato, además, la información tiene una representación óptima dependiendo el tipo de datos que se manejan y a qué público va dirigido. Incluso peor, la información puede no ser legítima (como ejemplo, es sabido que las grandes refresqueras daban dinero a varios laboratorios para publicar investigación que desmintiera los efectos negativos del azúcar en el cuerpo humano, para que su negocio pudiera crecer a costa de la salud mundial) y es la tarea de los científicos desmentir dichos estudios.

Antes de comenzar a dar el curso, se me preguntó si era pertinente con respecto al programa de estudios de la carrera de física (que le pertenece a *ciencias de la computación*, o que de *física nada hay*, que *los actuarios son los que hacen estadística*, o que al fin de cuentas son *matemáticas*). *Sí, sí es pertinente*, pues está en todos los que estamos en el área de la investigación formar un pensamiento crítico y tener nosotros mismos la capacidad de verificar que los datos que nos proporciona un estudio ajeno están bien manejados, ya sean estudios sobre comportamientos so-

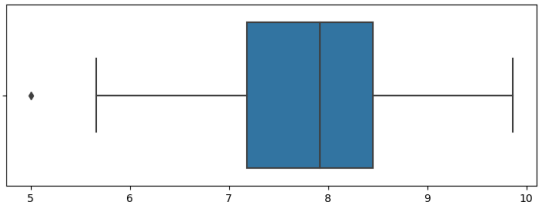

Figura 1: Gráfica de caja para los datos de la serie de promedios.

ciales, descubrimiento de partículas o especies nuevas, o lo que sea. Ya me puse muy poético y me estoy desviando...

En fin, al aceptar la idea de que la información va a estar escrito para humanos, también se acepta el reto de desarrollar sistemas que sean capaces de recolectar la información de manera automatizada sin importar el formato en el que se encuentre.

Por cuestiones de tiempo, nosotros nos vamos a saltar el paso mencionado anteriormente, y vamos a partir de tener la información disponible que podría estar corrupta. A partir de ahora, todo lo que diga se aplica a experimentos en la ciencia también.

2. Series de datos

Vamos a definir como serie de datos una secuencia finita $S = \{s_i\}_{i=1}^{n}$ de n elementos que pertenecen a $\mathbb{R}$. En este caso particular, vamos a suponer que estos valores numéricos están asociados al resultado de la mediciones $\{X_1, X_2, \dots X_n\}$ de una variable aleatoria $X : \Omega \to \mathbb{R}$ (con X definida sobre el espacio de probabilidad (Ω, F, P)).

Como la serie no necesariamente tiene que estar ordenada, llamaré $\dot{S}$ a la versión de esta serie cuyos elementos están acomodados en orden ascendente.

2.1. Mínimo, máximo y rango

Los elementos $\dot{s}_1$ y $\dot{s}_n$ son el mínimo y el máximo de la serie respectivamente. El rango está definido como la diferencia de estos dos elementos, $\dot{s}_n - \dot{s}_1$.

2.2. Cuantiles

Los cuantiles son puntos tomados a intervalos regulares de una serie de datos para repartir la información en bloques. Uno de los más usados son los llamados *cuartiles*, los cuales distribuyen la información en 4 bloques (3 cuantiles; 0.25, 0.5 y 0.75).

En general, no hay una regla establecida para determinar el valor de estos puntos (pues el punto puede o no ser parte de la serie de datos, dependiendo si hay un número par o impar de elementos), más bien hay convenciones. Nótese que el segundo cuartil es lo mismo que la media.

Por ejemplo, supongamos la siguiente serie de datos:

Estudiante	Promedio	Estudiante	Promedio	Estudiante	Promedio	Estudiante	Promedio
1	8.41	11	8.10	21	7.24	31	9.57
2	9.10	12	6.46	22	7.12	32	7.95
3	8.30	13	5.94	23	7.14	33	7.79
4	8.70	14	7.65	24	7.98	34	8.96
5	8.58	15	9.16	25	9.48	35	7.19
6	7.49	16	7.88	26	6.10	36	5.95
7	7.42	17	7.58	27	6.94	37	7.69
8	8.25	18	5.66	28	7.19	38	8.22
9	9.86	19	8.53	29	9.27	39	6.55
10	8.02	20	8.43	30	8.22	40	5.00

La manera de determinar la media en este curso es la siguiente:

- Si el número de datos n es impar, tomar el elemento de $\dot{S}$ con índice $r = (n-1)/2 + 1$
- Si el número de datos n es par, tomar el promedio de los elementos de $\hat{S}$ con índices $r_1 = n/2$ y $r_2 = n/2 + 1$

Para determinar los cuartiles 1 y 3, se toman las respectivas mitades de los datos marcadas por la media. Si el número de datos es impar, **se incluye la media en ambos grupos**. Para la tabla anterior, los tres cuartiles son:

Cuartil 1	7.17
Cuartil 2 (media)	7.91
Cuartil 3	8.45

Nótese que esto es sólo una convención, y no hay problema con que utilicen otra simpre y cuando la justifique y la mencione en su trabajo.

2.3. Momentos de una serie

Cuando se tiene una función de variable real $f(x)$, es posible definir los momentos de esta función alrededor del punto c de la siguiente manera:

$$\mu_k \coloneqq \int_{-\infty}^{\infty} (x-c)^k f(x) dx \tag{1}$$

A partir de $k = 2$, estos momentos son llamados centrales si $c = \mu \coloneqq \mu_1$. Si $f(x)$ es una función de densidad material, entonces los primeros tres momentos incluyendo el 0 son la masa total de un sistema, su centro de masa y su momento de inercia al rededor del punto c. Si $f(x)$ representa la función de distribución de probabilidad de una variable aleatoria X, entonces

$$\mu_k = E[(x-c)^k] \tag{2}$$

y el primer momento ($k = 1$) es el valor esperado, el segundo momento central es la varianza. Se definen como momentos normalizados, a partir de $k = 3$, las cantidades μ_k/σ^k. De utilidad son el tercer momento normalizado, conocido como asimetría (positivo si hay más elementos a la derecha del promedio), y el cuarto momento normalizado, conocido como Kurtosis (da información sobre las colas de la distribución respecto al resto de los datos).

En general, no vamos a conocer la distribución de probabilidad que originó nuestra serie de datos, y en muchas ocasiones haremos estadística inferencial en lugar de descriptiva, es decir; No tendremos todos los datos de la población disponibles, más bien un mustreo en representación de toda la población. Cuando es este el caso, se debe usar un estimador para determinar los parámetros estadísticos de una población. El estimador del k-momento de una población al rededor de un punto c basado en una serie S se define como

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} (s_i - c)^k \tag{3}$$

Para la población total, se obtiene que la varianza es:

$$\sigma^2 \coloneqq \mu_2 = \frac{1}{N} \sum_{i=1}^{N} x^2 - \mu^2 \tag{4}$$

2.3.1. Estimadores imparciales

Debería quedar claro que los momentos, son variables aleatorias en sí (dependen de una serie de valores tomados de manera independiente, al azar). Por ello, lo que estamos obteniendo es una estimación del parámetro (momento) de la población total. Se le llama estimador imparcial a un estimador cuyo valor esperado es el parámetro que se intenta estimar. Esto suena tonto, y que debiera cumplirse por sí solo, pero no es así: Tomemos el ejemplo de la varianza.

El valor esperado del estimador de la varianza es:

$$E[\hat{\sigma}^2] = E[\frac{1}{n}\sum_{i=1}^{n}(s_i - \frac{1}{n}\sum_{j=1}^{n}s_j)^2] \tag{5}$$

$$= E[\frac{1}{n}\sum_{i=1}^{n}(s_i^2 - \frac{2s_i}{n}\sum_{j=1}^{n}s_j + \frac{1}{n^2}\sum_{j=1}^{n}s_j\sum_{k=1}^{n}s_k)] \tag{6}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\frac{n-2}{n}E[s_i^2] - \frac{2}{n}\sum_{j\neq i}E[s_i s_j] + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k\neq j}^{n}E[s_j s_k] + \frac{1}{n^2}\sum_{j=1}^{n}E[s_j^2]) \tag{7}$$

$$= \frac{n-1}{n}\sigma^2 \tag{8}$$

El valor esperado del estimador de la varianza difiere del de la población total por un factor de $(n-1)/n$. Esto quiere decir que si hubieramos multiplicado nuestro estimador por $n/(n-1)$ habríamos obtenido un **estimador imparcial**, pues su valor esperado es el del parámetro de la población total que se intenta estimar. A dicho factor se le conoce como correción de Bessel.

Es valido mencionar que para n grande, este factor casi no hace diferencia alguna. Lo importante del análisis es saber que, al construír un estimador de un parámetro de una población se debe buscar que su esperanza coincida con el valor a estimar para la población total, pues ese factor pudo haber sido de mucha mayor relevancia. (aunque en este caso, ha habido un sacrificio en el error promedio cuadrático, pues cada medición va a estar más lejana al valor esperado).

2.3.2. Momentos mixtos: covarianza

Una de las herramientas más útiles para hacer análisis de datos o alimentar métodos de machine learning es deshacernos de información que no aporta nada nuevo (o en su caso, detectar relaciones entre series de datos). Este puede ser el caso cuando dos series de datos fluctúan manteniendo cierta relación.

La covarianza es un momento mixto que detecta esta relación (de manera lineal), y se define como:

$$\sigma(x,y) = E[(x - E[x])(y - E[y])] \tag{9}$$

el cual tiene su propio estimador

$$\hat{\sigma}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{10}$$

donde ya he incluído la correción de Bessel. Como observación trivial, $\sigma_{xx} = \sigma^2$

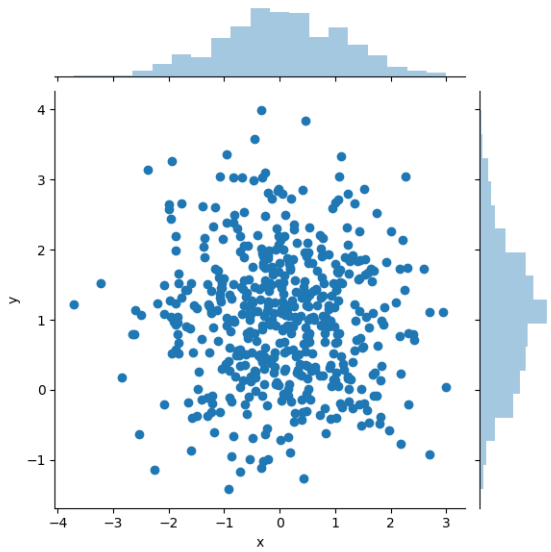

Figura 2: $\sigma_{xy} = 0{,}0$

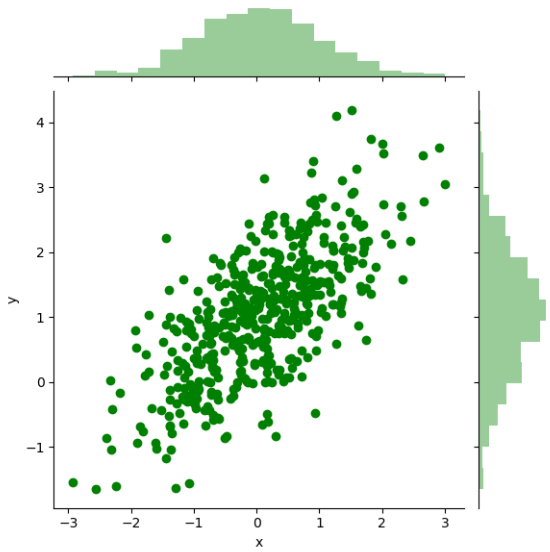

Figura 3: $\sigma_{xy} = 0{,}8$

Las figuras 2 y 3 son los resultados de una simulación de un proceso gaussiano multivariado de 500 mediciones. El primer proceso tiene ($\vec{\mu} = (0,1)^T, \boldsymbol{\sigma} = ((1,0)^T, (0,1)^T)$), mientras el segundo tiene como parámetros ($\vec{\mu} = (0,1)^T, \boldsymbol{\sigma} = ((1,0{,}8)^T, (0{,}8,1)^T)$). Los histogramas de cada dimensión son indistinguibles, pero es claro que hay una correlación en el segundo gráfico.

2.3.3. Matriz de covarianza

La matriz de covarianza $\boldsymbol{\sigma}$ (o bien, en un futuro *función de covarianza*) es una matriz cuyas entradas están dadas por:

$$\boldsymbol{\sigma}_{x_i x_j} = E[(x_i - E[x_i])(x_j - E[x_j])] \tag{11}$$

Donde cada x_i es lavariable aleatoria asociada a una serie de datos. La matriz contiene en la diagonal las varianzas de cada variable. Como nota, nosotros trabajaremos en su mayoría con la estimación de la matriz de covarianza:

$$\hat{\boldsymbol{\sigma}}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{12}$$

Para cualquier par de variables x e y.

2.3.4. Coeficiente de correlación

Una vez obtenida la estimación de la matriz de covarianza, se puede definir una cantidad conocida como *coeficiente de correlación*:

$$r_{xy} = \frac{\hat{\boldsymbol{\sigma}}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y} \tag{13}$$

Donde el denominador está compuesto de las desviaciones estándar de las variables x e y respectivamente.