DEPARTAMENTO
DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Metros Cuadrados Mínimos Lineales

Trabajo Práctico 3

19 de julio, 2020 — Métodos Numéricos

Grupo 4

Integrante	LU	Correo electrónico
López Menardi, Justo	374/17	juslopezm@gmail.com
Strobl, Matías	645/18	matias.strobl@gmail.com
Wehner, Tomás	67/17	tomi.wehner10@gmail.com
Yulita, Federico	351/17	fyulita@dc.uba.ar

Instancia	Docente	Nota
Primera entrega		
Segunda entrega		

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2610 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (++54 +11) 4576-3300
https://exactas.uba.ar

Resumen ??

Palabras Clave: ??

1. Introducción

En el presente trabajo se desarrollará y evaluará una herramienta de predicción de características de inmuebles. Para ello, se implementará un algoritmo de clasificación supervisado que será entrenado con una base de avisos de ventas de inmuebles con precios conocidos, que luego servirá para aproximar el precio de avisos de inmuebles no presentes en la base de datos de entrenamiento.

La motivación principal de este trabajo gira en torno al análisis y la experimentacion de resolver problemas de regresion. Las regresiones pueden ser de suma importnacia en casos donde se busque encontrar la manera de explicar datos que en principio se desconozca la familia de funciones que los responden.

2. Desarrollo

Se cuenta con un set de datos de avisos inmobiliarios de México. Se utilizará la técnica de Cuadrados Mínimos Lineales para aproximar una determinada característica en función de otras conocidas usando al precio como la variable que se quiere aproximar.

Se trabajará con variables numéricas en donde es importante el concepto de distancia asociado.

Una vez hecha la aproximación de la variable a estimar, se pondrá a prueba el algoritmo con un conjunto de validación con muestras que no hayan sido usadas durante el entrenamiento.

Una opción es usar la métrica RMSE (Root Mean Squared Error). Dado un modelo $\hat{f}$ y una observación (x_i, y_i), se define $\hat{y}_i = \hat{f}(x_i)$ y $\epsilon_i = y_i - \hat{y}_i$, se tiene:

$$\mathrm{RMSE}(\hat{\mathrm{f}}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\epsilon_i^2} \tag{1}$$

El problema que esta métrica tiene es que las muestras con valores altos pesarán más que aquellas con valores bajos. Por ello, en algunas ocasiones tiene mas sentido considerar a la métrica RMSLE (Root Mean Squared Log Error) definida como:

$$\mathrm{RMSLE}(\hat{\mathrm{f}}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\ln(y_i+1) - \ln(\hat{y}_i+1)\right)^2} \tag{2}$$

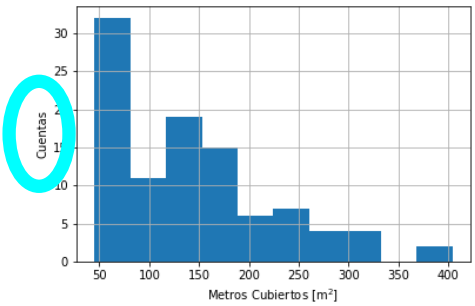
Esta última métrica tiene la propiedad de pesar de la misma manera la mejora porcentual sobre cualquiera de las muestras sin importar su valor absoluto.

En el trabajo presente se experimentará con ambas métricas para analizar el impacto en el algoritmo dependiendo de la característica con la que se esté trabajando.

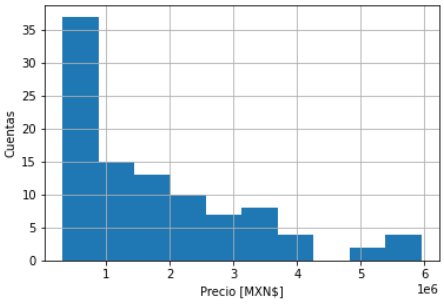
Uno de los problemas que se puede ver es que el método de Cuadrados Mínimos Lineales es un algoritmo que intenta explicar todos los datos con una sola función y el conjunto de datos a utilizar es bastante extenso y heterogéneo, por lo que será muy difícil conseguir una buena aproximación mediante cuadrados mínimos de todos los datos.

Para resolver este problema, se recurre a la segmentacion de datos. Con esto se logra enfocar el problema en un espacio mas controlable, osea, en el que menos variables esten en juego. Por ende, reducimos la complejidad, y al aplicar CML podemos predecir con mejor precision.

Otro de los procesos con los que se experimentara es el de feature engineering, que consiste en producir caracteristicas nuevas a raiz de los datos existentes. Esto se aplica para poder profundizar el estudio y a su vez intentar encontrar patrones o correlaciones que con los meros datos crudos uno no descubriria.



(a) Metros cubiertos de la muestra de viviendas tomada.



(b) Precios de la muestra de viviendas tomada.

Figura 1: Histogramas de los metros cubiertos y del precio de las viviendas de la muestra tomada.

3. Resultados y Discusión

3.1. Caso 1 - Metros Cubiertos vs. Precio

En este caso de experimentación se obtuvo una relación lineal entre los metros cubiertos y el precio de las viviendas. Debido a que las viviendas más grandes suelen ser más caras se esperaba hallar una relación positiva entre ambas variables.

Se tomó una muestra al azar de 100 viviendas del conjunto de entrenamiento de datos. En la **Figura 1** se encuentran histogramas de los metros cubiertos y los precios de las viviendas de la muestra. Notemos que la gran mayoría de las viviendas son de hasta 150 m^2 y salen menos de 1.5 millones de pesos. Esperamos entonces que la densidad de puntos para bajos precios y para viviendas chicas sea mucho mayor que para altos precios o viviendas grandes. Luego, con el algoritmo implementado se obtuvo un ajuste lineal de los datos y se lo usó para formar el gráfico de la **Figura 2**. Notemos que el ajuste es adecuado para la muestra y se obtuvo una relación positiva entre los datos como se esperaba. Además, la densidad de puntos para los distintos rangos de valores corresponde con lo que vimos en los histogramas. Sin embargo, debido a que los datos no están determinados por este único factor la cantidad de outliers es significante, especialmente para viviendas más grandes.

Se usó K-Fold Cross Validation en el conjunto de datos de entrenamiento para verificar la precisión del ajuste obtenido con un método robusto. Se usaron las fórmulas (1) y (2) para calcular ambos errores de los ajustes en los datos de la muestra. Se obtuvieron los valores RMSE $= (1{,}4317 \pm 0{,}0001) \times 10^6$ MXN\$ y RMSLE $= (0{,}590 \pm 0{,}005)$ con $K = 10$. Notemos que el RMSE es mucho mayor que el RMSLE. Esto es razonable, ya que el RMSE es del orden del precio de las viviendas (del millón de pesos) debido a los outliers. Sin embargo, el RMSLE no se ve tan afectado por outliers ya que es esencialmente el logaritmo del error relativo del ajuste; no del error absoluto.

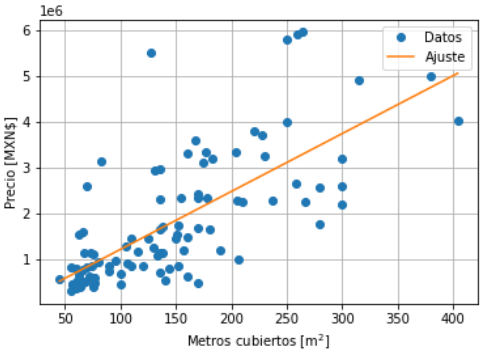


Figura 2: Precio de las viviendas en función de su tamaño en metros cubiertos.

3.2. Caso 2 - Latitud y Longitud vs. Precio

En este caso de experimentación se obtuvo una relación lineal entre la longitud y latitud y el precio de las viviendas. Esto nos permitiría hallar una dirección cardinal en la que el precio de las viviendas de México incrementa.

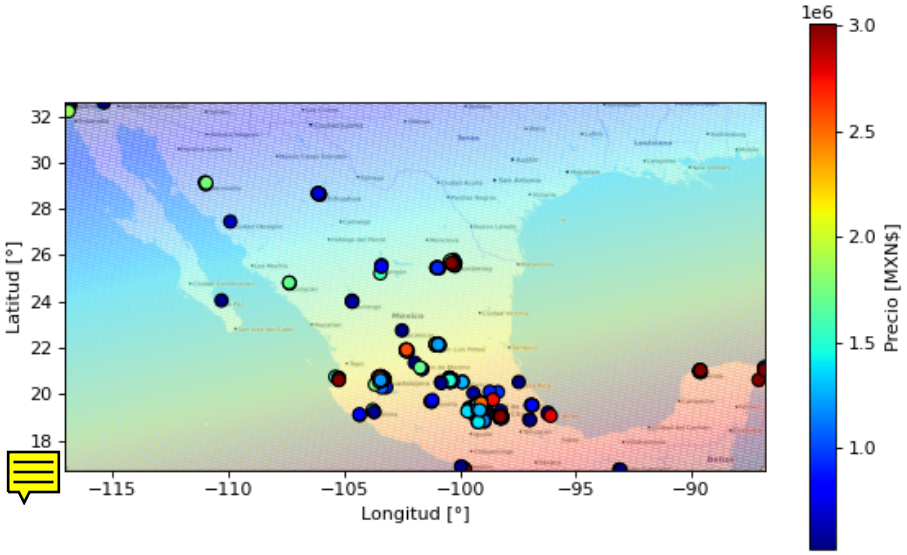


Figura 3: Precio de viviendas en México por ubicación.

Se tomó una muestra al azar de 500 viviendas del conjunto de entrenamiento de datos. Luego, con el algoritmo implementado se ajustaron los datos de la muestra y se obtuvo una dirección en la que los precios incrementan. Esta información se refleja en la **Figura 3**. Notemos que los puntos de la figura superpuestos en la imagen del mapa de México reflejan las viviendas de la muestra siendo el color el precio de cada vivienda, mientras que el gradiente de colores refleja el ajuste obtenido. Notemos que el ajuste muestra que las viviendas más caras están en la costa sur del atlántico. Esto es razonable, ya que esta zona es la del caribe de México con gran atractivo turista. ok!

Al igual que en el caso anterior, se usó K-Fold Cross Validation para calcular el RMSE y RMSLE. Se obtuvo RMSE $= (1{,}7816 \pm 0{,}0002) \times 10^6$ MXN\$ y RMSLE $= (0{,}820 \pm 0{,}004)$ con $K = 5$. Notamos que, al igual que en el caso anterior, RMSE es mucho mayor que RMSLE y esto refleja la variación y la presencia de outliers en la muestra.

3.3. Caso 3 - Precios de Viviendas en Cancún

En este caso de experimentación se obtuvo una relación cuadrática entre la longitud y latitud y el precio de la viviendas pero solo para la zona céntrica y costera de la ciudad de Cancún. Es decir, ajustamos el precio de esta zona a las variables $long$, lat, $long^2$, lat^2 y $long \cdot lat$; donde $long$ es la longitud y lat la latitud. Esto nos permitió no solo encontrar una dirección cardinal en donde los precios son más altos sino que también nos perimitió dividir esta zona con mayor precisión ya que al tener estas variables adicionales podemos dividir zonas con curvas cuadráticas. Elegimos esta ciudad ya que nos pareció interesante ver cómo cambian los precios a medida que las viviendas se alejan de la costa. Esperamos hallar que las viviendas más cercanas a la costa sean las más costosas de la ciudad.

ok!

Se tomó una muestra al azar de 200 viviendas y restringimos las viviendas a los rangos $long \in [-86{,}92, -86{,}74]$ y $lat \in [21{,}05, 21{,}20]$. Luego, se ajustaron los datos de la muestra y se obtuvo el ajuste mostrado en la **Figura 4**. Notemos que el ajuste es adecuado a los datos de la muestra y que además la curvatura hallada corresponde con la curvatura natural de la costa de Cancún que separa adecuadamente por zonas los precios de las viviendas. Es decir, las viviendas azules se encuentran en el relleno azul del ajuste, la banda de viviendas verdes y amarillas se encuentran en la banda verde y amarilla de relleno y las viviendas más costosas se encuentran en la costa donde la banda de relleno es roja. Esto se corresponde con lo que esperábamos hallar. Esta división precisa entre las viviendas es algo que no se obtuvo en el caso anterior ya que solo habíamos obtenido una dirección cardinal donde los precios tienden a ser mayores.

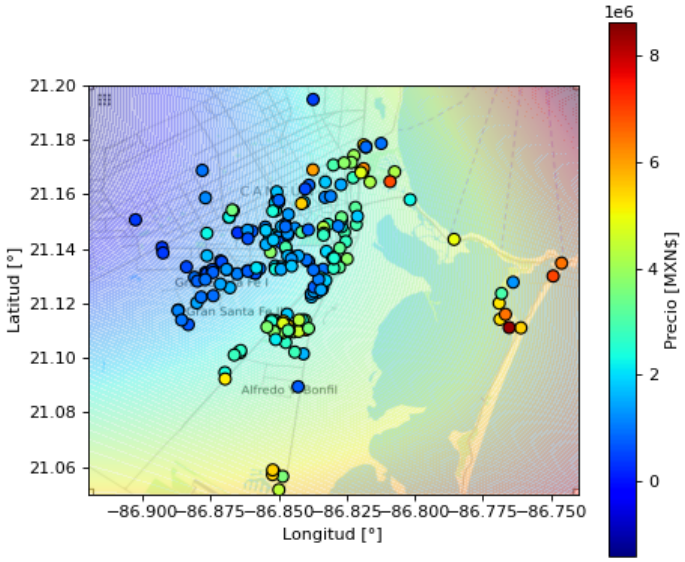


Figura 4: Precio de viviendas en Cancún por ubicación.

Al igual que en los casos anteriores, se usó K-Fold Cross Validation para calcular el RMSE y RMSLE. Se obtuvo RMSE = $(1{,}7814 \pm 0{,}0001) \times 10^6$ MXN$ y RMSLE = $(0{,}821 \pm 0{,}004)$ con $K = 5$.

3.4. Caso 4 - Metros cubiertos

Se obtuvo estimador para los metros cubiertos de un inmueble. Para ello se realizó una regresión lineal en función de la cantidad de habitaciones , baños y garage.

Observando una muestra al azar de 100 viviendas para una mayor simplicidad en los graficos se obtuvo la figura **Figura** 5 En la cual se puede observar a la cantidad de metros cubiertos en función de la cantidad de ambientes (suma del número de habitaciones, baños y garages) que contenga el inmueble.

Veamos que si bien existe una varianza entre inmuebes con la misma cantidad de ambientes se puede observar una proporcionalidad directa entre ambas variables, a más ambientes más metros cubiertos.

Se utilizo K-Fold cross validation con K=5 sobre el conjunto de entrenamiento para aproximar el error generado por este ajuste de forma más robusta. Se utliizó las ecuaciones (1) y (2) para obtener asi los valores de RMSE y RMSLE respectivamente.Obteniendo como resutados a un RMSE = $(59{,}66 \pm 0{,}67)$ y un RMLSE = $(0{,}3869 \pm 0{,}0051)$

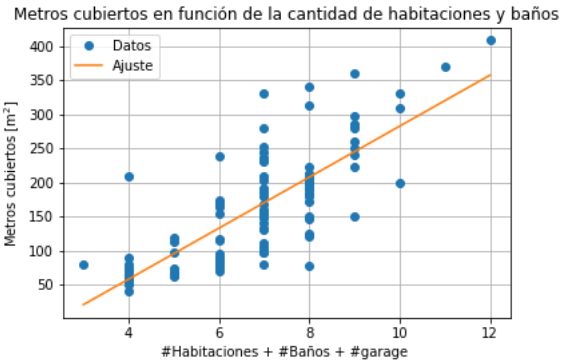


Figura 5: Metros Cubiertos

Continuando con la experimentación, se decidió observar otra carecterista extra. Se observó la cantidad de metros cubiertos en función de los Metros totales y ambientes. Y realizando una resion lineal sobre estas varaibles. Utilizando un sample de datos de 100 inmuebles se obtuvo la **Figura** 6. Del mismo modo se utilizo K-fold cross validation para obtener los valores de RMSE = $(48{,}83 \pm 0{,}44)$ y RMLSE = $(0{,}3032 \pm 0{,}0041)$

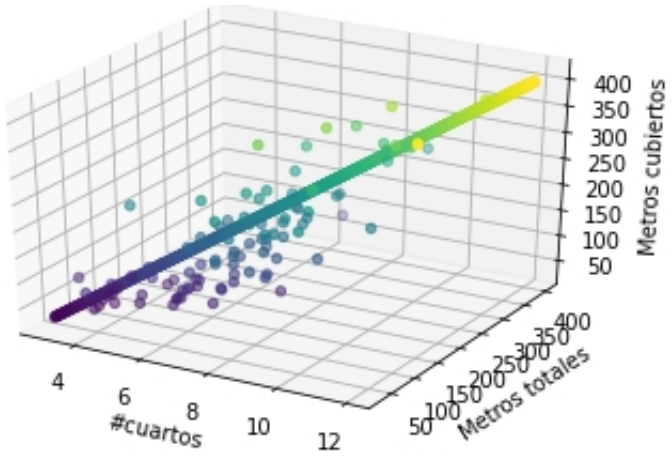


Figura 6: Metros Cubiertos en función de la cantidad de ambientes y metros totales

Luego agregó se observaron cometarios positivos encontrados en las descripciones de los cometarios Se creo una nueva nueva caracteristica para cada inmueble llamada / cometarios positivos.La cual refleja la cantidad de palabras positivas elejidas de una lista creada por nosotros que se encontró en su descripción , por ejemplo si contiene la palabra *luminoso* o *espacioso*. Utilizando a esta nueva caracteristica como una variable más al obtener el ajuste mediante regresión lineal se obtuvo un RMSE = $(48{,}76 \pm 0{,}45)$ y un RMLSE = $(0{,}3030 \pm 0{,}0034)$

4. Conclusiones

Se observó que la presición del ajuste mejoraba en un muy pequeño porcentaje al analizar las descripciones de los inmuebles.En el caso de los metros cubiertos se obsevó una reducción del error de un RMSE 0.25 %. Se cree que esto puede ser debido a la subjetividad de los vendedores lo cua les permite escribir descripciones positivas en inmuebles que no las merecen para así poder vender mejor. Provocando que la mayoria de estos tengan los mismos tipos de cometarios positivos.