Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Metros Cuadrados Mínimos Lineales

Introducción

El objetivo de este trabajo práctico es el desarrollo y evaluación de una herramienta de predicción de características de inmuebles.

Se quiere desarrollar un algoritmo de clasificación supervisado el cual se deberá *entrenar* con una base de avisos de ventas de inmuebles con precios **conocidos**, que luego nos servirá para aproximar el precio de avisos de inmuebles no presentes en la base de datos de entrenamiento.

Desarrollo

En este trabajo vamos a utilizar la técnica de Cuadrados Mínimos Lineales para aproximar una característica de los inmuebles, en función de varias otras características conocidas. Al tener avisos inmobiliarios con un conjunto fijado de características (ej: Metros cuadrados, Precio, Cantidad de baños) podemos tomar una de estas y buscar una manera de explicarla mediante una función simplificada del resto de las características.

Para este trabajo contamos con un set de datos de avisos inmobiliarios de México. A continuación, listamos algunas de las características que tendremos disponibles en nuestro set de datos.

- id: Número identificador del aviso sin otra semántica particular.
- titulo: Contiene el título del aviso inmobiliario.
- descripción: Contiene una descripción en texto plano del aviso.
- tipodepropiedad: Variable categórica para definir el tipo (Casa, Apartamento, etc).
- Ubicación: varias características autoexplicativas (ciudad, provincia, etc) sobre la ubicación del inmueble.
- Características numéricas: varias carácteristicas cuantificables (baños, piscina, escuelascercanas, etc).

Como objetivo principal del trabajo, usaremos el precio como la variable que queremos aproximar. Sin embargo, podríamos utilizar los mismos métodos para intentar explicar diversas relaciones entre las variables y responder algunas preguntas como:

- ¿Podemos aproximar la cantidad de baños según la antiguedad y el tamaño en metros cuadrados del inmueble?
- ¿Podemos explicar la cantidad de piscinas según la latitud del inmueble?
- ¿Podemos aproximar la antiguedad de un inmueble según su ciudad y zona?

Variables numéricas vs categóricas

El conjunto de datos a utilizar contiene variables numéricas y variables categóricas. Las variables numéricas son aquellas que toman algún valor númerico, ya sea entero o real. Las variables categóricas son aquellas en las que el valor corresponde a una categoría dentro de un conjunto posible. Si bien las variables categóricas pueden codificarse como números enteros (Categoría 1, 2, etc), es importante notar que no por eso pasan a ser variables numéricas.

Las variables numéricas tienen un concepto de distancia asociado. Si tomamos una propiedad que tiene 1 baño, entonces el hecho de tener 2 baños es un concepto más cercano que tener 5. Las variables categóricas no tienen este concepto de distancia asociado. La categoría 1 es tan distinta de la categoría 2 como de la 5. De hecho, que se llamen categoría 1, 2 y 5 es una mera cuestión de representabilidad que podría ser intercambiable por otras maneras.

En el método de cuadrados mínimos lineales trabajaremos siempre con variables numéricas en donde el concepto de distancia es más que importante. Las variables categóricas no serán utilizadas en el método de cuadrados mínimos, pero eso no significa que no pueden ser utilizadas en el algortimo final. Veremos más adelante como podremos utilizar dichar variables.

RMSE vs RMSLE

Una vez hecha la aproximación de la variable a estimar, podremos poner a prueba nuestro algoritmo con un conjunto de validación con muestras que no hayamos utilizado durante el entrenamiento. Si, por ejemplo, nuestro objetivo fue estimar el precio de los inmuebles, entonces necesitaremos una métrica para ver que tanto difieren los precios estimados de los reales.

Una primera opción posible es utilizar el *Root Mean Squared Error* (RMSE).

Dado un modelo $\hat{f}$ y una observación $(x_{(i)}, y_{(i)})$, definimos $\hat{y}_{(i)} = \hat{f}(x_{(i)})$ y $e_{(i)} = y_{(i)} - \hat{y}_{(i)}$. Con estas definiciones, podemos calcular el RMSE del modelo $\hat{f}$ como:

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

Si bien esta métrica es ampliamente utilizada, puede no resultar ideal en algunos casos. Por su definición, la métrica RMSE pesa más las muestras con valores altos que las muestras con valores bajos. Si mejoramos la aproximación de una muestra alta en un 10 %, tiene mucho más influencia en la métrica RMSE que lograr el mismo objetivo en una muestra con valor bajo. Es por esta característica que en algunas aplicaciones tiene sentido considerar como métrica alternativa a el *Root Mean Squared Log Error* (RMSLE). La métrica RMSLE se define como:

$$RMSLE(\hat{f}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(y_{(i)}+1) - log(\hat{y}_{(i)}+1))^2}$$

La métrica RMSLE tiene la propiedad de pesar de la misma manera la mejora porcentual sobre cualquiera de las muestras sin importar su valor absoluto. Sin embargo, su definición introduce algunas otras propiedas no deseadas.

En el trabajo se debe experimentar con estas métricas, y otras posibles, para ver como impactan en el algoritmo según la característica que estemos intentando explicar.

Segmentación

El método de cuadrados mínimos lineales es un algoritmo ambicioso en el sentido de que intenta explicar todos los datos con una única función. Muchas veces el conjunto de datos disponible es lo suficientemente heterogeneo como para que sea muy dificil conseguir una buena aproximación mediante cuadrados mínimos de todos los datos. En estos casos una opción viable para atacar el problema es la de segmentar el conjunto de datos. Es posible que una única función no pueda explicar bien al conjunto de todos los inmuebles pero ¿qué pasa si primero segmentamos los inmuebles por la provincia a la que pertenecen?, ¿Qué sucede si armamos un modelo para los inmuebles *baratos*, uno para los inmuebles *intermedios* y uno para los inmuebles *caros*?

Hay muchas manerar de segmentar el conjunto de datos, y es posible que las variables categóricas puedan tomar un rol fundamental en este paso. No es necesario producir un único modelo de cuadrados mínimos lineales que explique todo el conjunto de datos, sino que el algoritmo final puede estar compuesto de diferentes aproximaciones.

Feature engineering

Las características de un conjunto de datos se suelen denominar *features* en inglés. El proceso de *feature engineering* consiste en producir nuevas caracteríStcas para utilizar en los métodos de aproximación. Estas nuevas características pueden provenir de fuentes internas del conjunto de datos así como de fuentes externas.

Algunas posibilidades para la generación de nuevas características son:

- Combinación de características. Más allá del peso que puedan tener las características por sí solas en una aproximación de cuadrados mínimos, es posible que la combinación de características existentes sea útil para lograr una mejor aproximación. Ejemplo: Un inmueble puede tener la característica de *ser copado para el verano* sí y solo sí el inmueble tiene pileta o se encuentra al norte del paralelo 26.
- Extracción de información de los campos de texto. El conjunto de datos tiene algunos campos que son párrafos en texto plano. Es posible crear nuevas características en base a la información que contienen dichos campos. Ejemplo: ¿Es importante que la descripción del inmueble diga *luminoso* para explicar el precio?
- Características de fuentes externas. Otra posible opción para generar nuevas características se basa en sumar información externa, uniendo el conjunto de datos disponible con otros conjuntos de datos que se puedan encontrar. Ejemplo: No todas las provincias de

un país se comportan de la misma manera, es posible que sumar las características *Población*, *Producto Bruto Interno* e *Índice de Precios del Consumidor* a nuestro conjunto de datos sea beneficioso para mejorar nuestra aproximación de precios.

Enunciado

Se pide implementar un programa en `C++` que lea los datos de entrenamiento correspondientes a distintos inmuebles y que, utilizando los métodos y métricas descriptos en la sección anterior, dado una nuevo inmueble pueda determinar el valor de la característica que se quiere explicar.

Experimentación

La experimentación de este trabajo es deliberamente más abierta que la de los trabajos anteriores. Quedará a criterio del grupo definir los ejes y métricas con los que experimentar.

Para guiar un poco la experimentación, se detallan los siguientes lineamientos:

- Realizar modelos para explicar al menos dos características del conjunto de datos. Una de las características elegidas debe ser necesariamente el precio del inmueble.
- Utilizar diversas métricas para evaluar las aproximaciones. Minimamente deben utilizar el RMSE y el RMSLE en algunos de los análisis.
- Proponer al menos dos segmentaciones para mejorar las aproximaciones.
- Utilizar al menos dos técnicas de *feature engineering* para enriquecer el conjunto de datos de entrenamiento.
- Realizar los experimentos utilizando *cross-validation* para evitar conseguir aproximaciones que funcionen bien en el conjunto de entrenamiento pero que luego no generalicen bien sobre otros inmuebles.

En todos los casos es **obligatorio** fundamentar los experimentos planteados, proveer los archivos e información necesaria para replicarlos, presentar los resultados de forma conveniente y clara, y analizar los mismos con el nivel de detalle apropiado. En caso de ser necesario, es posible también generar instancias artificiales con el fin de ejemplificar y mostrar un comportamiento determinado.

Fecha de entrega

- Formato Electrónico: Lunes 7 de Diciembre de 2020, hasta las 23:59 hs, enviando el trabajo (informe + código) a la dirección `metnum.lab@gmail.com`. El subject del email debe comenzar con el texto `[TP3]` seguido de la lista de apellidos de los integrantes del grupo.

Importante: El horario es estricto. Los correos recibidos después de la hora indicada no serán considerados.