DIPLOMATURA EN CIENCIA DE DATOS

Análisis exploratorio

Análisis exploratorio

Temario a tratar:

- Qué es
- Cuando se realiza
- Qué busca
- Tipos de datos
    - Numéricos
    - Categóricos
        - Ordenables
        - No ordenables
- Análisis individual
    - Numéricos
    - Categóricos
- Análisis conjunto
    - Numéricos vs numéricos
    - Categóricos vs categóricos
    - Numéricos vs categóricos
- Reducción de la dimensionalidad
    - Por que?
    - Cómo hacerlo mal?
    - Cómo hacerlo bien?
- Datos faltantes
    - Cómo hacerlo mal? (aunque sea a veces)
        - Eliminar filas
        - Eliminar columnas
        - Reemplazar por datos aleatorios que sigan la distribución de los presentes
    - Cómo hacerlo bien? (snif)

¿Qué es el análisis exploratorio?

El análisis exploratorio es la primera toma de contacto que realiza un científico de datos con el paquete de datos que se propone utilizar para construir un modelo.

Tiene el propósito de familiarizarnos con el contenido de los datos, evaluar sus posibilidades y limitaciones.

Es peligroso construir modelos sin haber realizado el análisis exploratorio porque vamos a estar menos alerta para detectar errores que invaliden el trabajo realizado.

¿Cuándo se realiza el análisis exploratorio?
En principio es la primera actividad que realizamos al recibir los datos.
Si hay cambios en los datos puede ser necesario volver a visitar el análisis exploratorio.

¿Qué busca el análisis exploratorio?
El análisis exploratorio busca saber los tipos de los datos que tenemos disponibles, sus características generales y hacerse una idea de la calidad de los datos.
De esta manera tomamos conciencia de las posibilidades de los datos y de las limitaciones que pudieran tener.

Tipos de datos:
Los datos pueden ser, en principio, numéricos o categóricos.
Los datos numéricos pueden ser tanto números enteros como con decimales.
Los datos categóricos pueden subdividirse en ordenables y no ordenables.

Una serie de categorías es ordenable cuando existe un criterio que permite acomodarlas de forma que las que están vecinas representan circunstancias parecidas mientras que las apartadas representan circunstancias bien distintas.

Por ejemplo, el nivel socioeconómico A, B, C1, C2, C3, D1, D2, E es un dato categórico ordenable pues A es más parecido a B que a C1.
Otro ejemplo sería el nivel de estudios máximo alcanzado primario, secundario, universitario, postgrado.

La localidad es un ejemplo de un dato categórico no ordenable.

Hay algoritmos, como los árboles de decisión que permiten utilizar tanto datos numéricos como categóricos. Sin embargo no es siempre el caso. Las redes neuronales o vecinos cercanos, por ejemplo sólo admiten datos numéricos.

El desafío es entonces, para no perder poder predictivo, ver cómo convertir esos datos categóricos en numéricos.

Si se trata de datos ordenables no es un problema.
Si se trata de datos no ordenables se complica.
Existe la práctica extendida de ordenar por orden alfabético los datos categóricos no ordenables y asignar un número entero creciente a cada dato dentro de la categoría.
Por ejemplo

Amarillo:	1
Rojo:	2
Verde:	3

Esto es algo peligroso de hacer ya que naturalmente tendería a confundir al algoritmo.
Hay que tener presente que nuestro modelo predictivo necesita, para tener sentido real, transformar datos parecidos en resultados parecidos.

Si, al reemplazar los datos categóricos por números generados de esta manera, introducimos esa información en un algoritmo nos vamos a encontrar que pueden tener números parecidos realidades muy distintas y números muy distintos realidades que son parecidas.

Consideremos el caso de la localidad que es un dato categórico no ordenable y que puede ser de gran importancia para muchos problemas de negocios:

San Antonio de Areco	1
San Antonio de los cobres	2
San Antonio Oeste	3

Son localidades muy distintas en cuanto a su tamaño, clima, actividad económica. Que se les asignen números cercanos porque los nombres empiezan con San Antonio no guía al algoritmo de ninguna manera y esconde que, por ejemplo, San Antonio de Areco puede ser más parecida a Pergamino que va a tener un número muy lejano.

Para no perder el poder predictivo encerrado en los datos categóricos podemos reemplazar el dato categórico por datos numéricos que representen lo que importa de la localidad para nuestro problema.

Por ejemplo puede ser la población, el PBI x cápita, la temperatura promedio, etc. No estamos obligados a reemplazar la categoría no ordenable por un solo número, pueden ser varios.

¿Cómo estamos seguros de cuales convienen?

Probando, para lo cual vamos a necesitar una forma de medir cuan bueno es cada modelo predictivo, tema que estaremos abordando más adelante.

Análisis individual:
Es el análisis que hacemos sobre cada dato por separado.
Este análisis depende de si se trata de datos numéricos o categóricos.

Para los datos numéricos vamos a mirar los extremos, mínimo y máximo, promedio, mediana, cuartiles y modas.

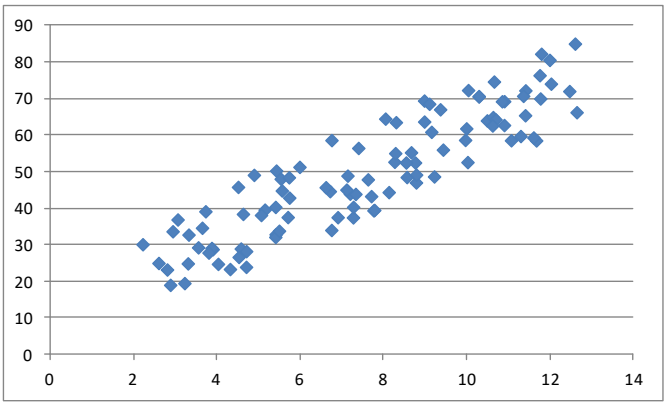
Toda esa información la vamos a resumir en un histograma que nos va a dar una visión clara de cómo se distribuyen los datos.

Para los datos categóricos vamos a mirar las frecuencias absolutas y relativas. Para obtener una visión de estos datos podemos realizar gráficos de torta.

Análisis conjunto:
Es el análisis que hacemos sobre cada par de datos.
Se dan tres situaciones:

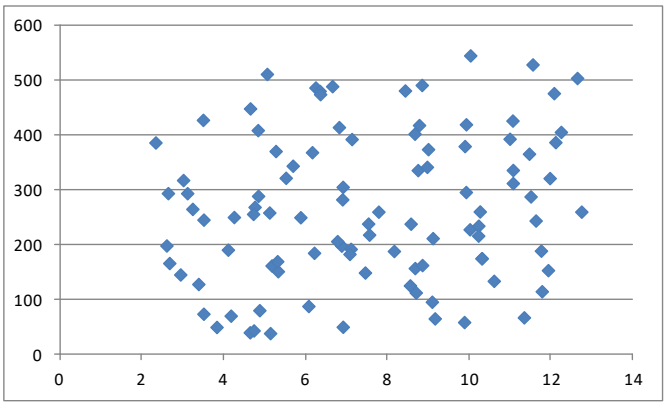
- Dos datos numéricos
- Dos datos categóricos
- Un dato numérico y un dato categórico

Entre dos datos numéricos vamos a explorar la existencia de correlaciones. Esto implica ver si la variación de un dato se ve acompañada en muchos casos por la variación del otro:



En este caso parece haber correlación entre el dato representado en el eje X y el representado en el eje Y.

Sin embargo la relación podría parecerse a:



En este caso no hay correlación alguna.

La forma de reflejar la correlación en un número se consigue con el coeficiente de correlación:

$$C = \frac{\sum_{i=1}^{100}(x_i - x_{medio}) * (y_i - y_{medio})}{\sqrt{\sum_{i=1}^{100}(x_i - x_{medio})^2} * \sqrt{\sum_{i=1}^{100}(y_i - y_{medio})^2}}$$

Solo para que vayan calibrando el ojo para el primer gráfico da 0.93 y para el segundo 0.13

Cuando tenemos dos datos categóricos corresponde analizar la independencia entre ambos. Dos datos A y B son independientes si y sólo si:

$$P(A) * P(B) = P(A \cap B)$$

Un ejemplo es lo que pasa con las cartas españolas que usamos habitualmente para jugar al truco. Son 40 cartas, divididas en 4 palos (oro, copas, espadas, bastos) con 10 cartas cada uno.

La probabilidad del as de espadas es 1 en 40: 1/40 = 0.025
La probabilidad de espadas es 10 en 40 = 0.25
La probabilidad de un as es 4 en 40 = 0.1
Y se cumple exactamente que:

0.25 * 0.1 = 0.025

¿Qué significa esto?
Esto significa que el saber que una carta es de espadas no nos aporta información alguna sobre si es un as.

Por supuesto que también se cumple que saber que una carta es un as no nos dice nada sobre su palo.

Si el mazo de cartas estuviera trampeado de manera que los cuatro ases fueran de espadas las probabilidades serían:

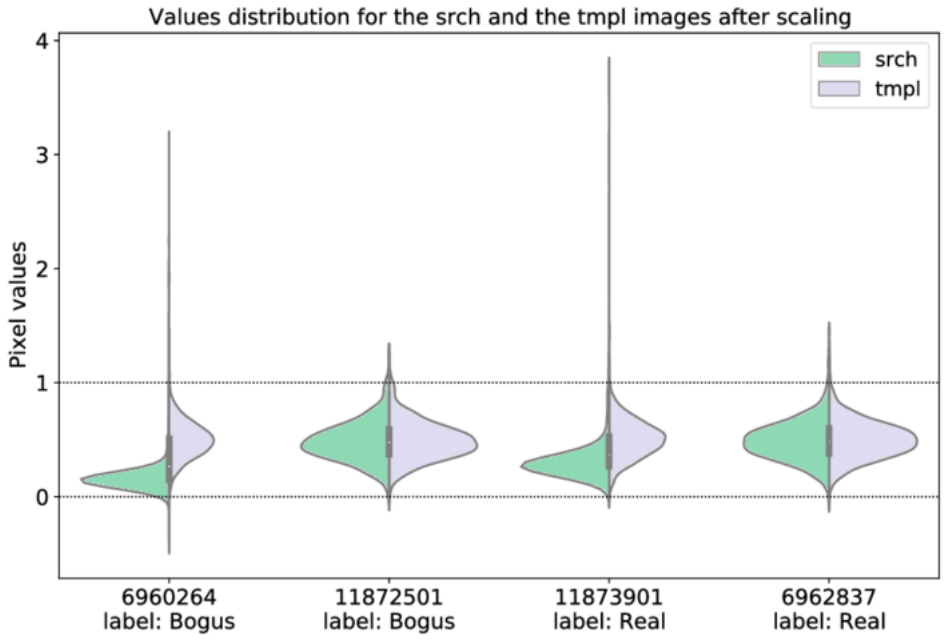
P(as) = 4/40
P(espadas) = 13/40
P(as de espadas) 4/40

En este caso no se cumpliría la igualdad pues

$$4/40 \neq 13/40 * 4/40$$

En este mazo con trampa las categorías espadas y as no son independientes. Significa que saber que una carta es de espadas aporta información sobre si puede ser un as y saber que una carta es un as nos asegura que es de espadas.

Si estamos analizando un dato numérico y un dato categórico tenemos que ver en qué medida el dato categórico influye en la distribución del dato numérico:



Reducción de la dimensionalidad

El esfuerzo computacional que se debe hacer para crear un modelo depende de dos factores, la cantidad de datos (columnas) y la cantidad de casos (filas).

En una época en la que el poder de cómputo disponible era escaso y caro se hacían esfuerzos para optimizar su uso.

Una de las prácticas en boga era ver si dos atributos estaban fuertemente correlacionados. Luego se procedía a eliminar uno de ellos ya que se estimaba que contenían la misma información.

En algunas circunstancias esto puede llevar a pérdida irreparable de información.

Esto NO debe ser confundido con la técnica de análisis de componentes principales que si produce resultados confiables.

En principio no hay razón para andar eliminando datos.

Gestión de los datos faltantes:

Es muy común que en nuestro conjunto de datos nos enfrentemos a datos faltantes.

En R los vamos a ver como NA y en SQL los vamos a identificar como NULL.

Muchos algoritmos son completamente incapaces de manejar estos datos.

Desgraciadamente no hay formas automáticas de resolver con seguridad el problema de los datos faltantes.

Eliminar las filas que posean datos faltantes puede introducirnos un sesgo que sabotee el resultado que busca nuestro modelo.

Eliminar las columnas que posean datos faltantes puede llevarnos a perder poder predictivo.

Una técnica más sofisticada es reemplazar los datos faltantes por datos aleatorios tal que sigan la misma distribución de probabilidad de los datos presentes en esa columna. Está técnica, sin embargo, en algunas circunstancias, también genera resultados falsos.

Para abordar el tema de los datos faltantes se debe tener en cuenta el origen de los datos y tratar de averiguar las razones de esa falta. A veces esto ayuda a reponer razonablemente esos datos.

Como siempre tenemos que probar distintas hipótesis y ver cuál de ellas produce el mejor modelo.