Visualización

Este trabajo trata sobre visualización de datos, presente en todo proceso KDD. Aquí se presenta el desarrollo del trabajo realizado para la asignatura Machine Learning Engineering, del Máster en Ingeniería del Software: Cloud, Datos y Gestión TI de la Universidad de Sevilla, curso académico 2022-23.

Tabla de contenido

Autor

Este trabajo y este *notebook* han sido desarrollados por **Mario Ruano Fernández** (mruano@us.es).

Introducción

En este trabajo se pretende realizar un estudio del estado del arte de la visualización de datos, marcando como objetivos.

- Describir los diferentes tipos de formas gráficas que se utilizan para la visualización de datos.
- Establecer una serie de criterios para el correcto uso de la visualización.
- Desarrollo de ejemplos y casos de uso.

Para dar cumplimiento a estos objetivos, en la sección Elementos para la visualización se especifican algunas características importantes que deben tenerse en cuenta en toda representación gráfica, de modo que pueda obtenerse, como resultado, una buena visualización de datos. En la sección Criterio de uso se incluye una descripción de los diferentes tipos de visualizaciones que se pueden encontrar según en qué situaciones dentro de un análisis exploratorio de datos. Por último, en la sección Caso práctico, se pondrán en práctica los aspectos teóricos estudiados, realizándose ejemplos con la librería Bokeh.

Definición

La visualización de los datos puede definirse como una forma gráfica y visual de representar información y datos a través de tablas, gráficos, mapas y otros elementos visuales (Tableau, s.f.). Gracias a la visualización, las personas pueden acceder a la información de los datos de manera más sencilla y comprensible (Sahay, 2016).

Frente a grandes cantidades y conjuntos de datos, la visualización ayuda a sintetizar y detectar de manera efectiva características que los datos esconden, relaciones entre variables, patrones, tendencias, etc. Por esta razón, la visualización se utiliza en la gran mayoría de fases o etapas de cualquier proceso KDD. Desde la selección de datos hasta la comunicación del conocimiento extraído de los mismo, pasando por el preprocesamiento, el análisis y la aplicación de técnicas de Machine Learning y Data Mining.

La visualización de datos es un proceso, en esencia, simple: tomamos los valores de los datos y los convertimos sistemática y lógicamente en elementos visuales que conforman una representación gráfica o visual (Wilke, 2019).

Sin embargo, el campo de la visualización es extremadamente amplio. Aquí convergen una serie de disciplinas, como el diseño gráfico, el arte, la ciencia cognitiva, las matemáticas, la geometría, etc.

Siguiendo a Sahay (2016), encontramos dos categorías dentro de la visualización de datos: la exploración y la explicación. Estas atienden a momentos distintos y existen técnicas y herramientas que se ajustan mejor a cada una de ellas.

La exploración de datos consiste en la búsqueda del significado y el sentido de los datos cuando realmente no se sabe qué aportan, frente a la explicación, cuando sí se tiene conocimiento sobre los datos y se trata de contar y transmitir el conocimiento de estos. Aun así, se suelen dar casos híbridos, ya que estas categorías no se ven representadas de manera exacta siempre (Sahay, 2016).

En definitiva, en todo caso será determinante el contexto en el que se realiza la visualización y con qué enfoque se desarrolla. Por lo general, en este trabajo, se trabaja en la línea de la visualización aplicada al campo del Machine Learning.

Elementos para la visualización

Tal y como define Wilke (2019) en su libro Fundamentals of Data Visualization, la visualización de datos comprende una parte de arte y otra parte de ciencia. Se debe buscar un equilibrio entre ambas partes para cumplir con su propósito de uso: comunicar una información correcta, comprensible, sin ambigüedades, útil y de interés. En definitiva, la visualización de datos no es más que una forma de comunicación visual que presenta los datos de forma gráfica (Sahay, 2016).

Por esta razón, y debido a esa necesidad de equilibrar las creaciones gráficas, aquí se identifican una serie de elementos y aspectos a tener en cuenta a la hora de generar visualizaciones. Es un amplio campo de trabajo y todos estos elementos pueden ser estudiados con gran profundidad y detalle. En este *notebook* solo se van a identificar y se indicarán algunas buenas prácticas, especialmente en aquellos elementos que pueden ser manipulados a través de las librerías de visualización, en general, y de los que permite personalizar la librería Bokeh, en particular.

En definitiva, cuando se pretende generar gráficos a partir de datos, se está tratando de desarrollar un procedimiento de mapeo, mediante el cual se pasa de una serie de datos en bruto, normalmente numéricos, a una representación estética y visual de los mismos. Aquí entran en juego aspectos como la posición, las formas, el tamaño, el color, etc.

Es, en ese procedimiento de mapeo, donde se detectan los elementos que darán identidad visual a la realidad intrínseca de los propios datos que representan, y que se enuncian a continuación.

Ejes

En este trabajo se hace referencia a gráficos en 2D, generalmente aquellos que utilizan un sistema de coordenadas cartesianas para sus ejes.

Respecto a los ejes hay que tener en cuenta un par de aspectos:

- La ratio de aspecto del eje horizontal frente al vertical. Esto está relacionado con el ancho y alto de la gráfica que se visualiza. Afecta a cómo se visualizan los resultados, aunque sean diferentes versiones correctas.
- La escala numérica de los ejes. Pueden seguir una escala lineal, logarítmica o de raíz cuadrada, entre otras. Esto afectará a la forma de visualizar los datos y de dar una correcta interpretación de los ellos.

Es recomendable hacer uso de la escala logarítmica en aquellos casos en los que se quieren representar valores en los que hay una gran diferencia en cuanto a magnitudes (Wilke, 2019).

Otro tipo de sistema de representación es el sistema de coordenadas polares, con eje curvo. Es óptimo hacer uso de este tipo de ejes cuando se quieren representar datos de naturaleza periódica.

Color

El color es uno de los principales elementos visuales que captan la atención en un gráfico. Por tanto, si se hace una correcta elección y aplicación del mismo puede ser de gran utilidad para comunicar la información. De manera general, el color cubre los siguientes requisitos (Wilke, 2019):

- Distinción de elementos discretos o grupos: mediante el uso de escalas cualitativas de color, el objetivo es poder distinguir de forma clara entre distintos conjuntos, evitando que algún color destaque por encima de otros o que den sensación de ordenación. Por ejemplo, esto es muy importante a la hora de representar clusters.
- Representación de valores: generalmente se busca representar una escala de valores o un rango, por lo que se aplica una escala secuencial de color. Esta debe presentar un rango uniforme perceptible, que puede bien puede ser monocromático o multicromático, pero lo importante es que en todo momento se pueda comprender ya apreciar con facilidad la diferencia de magnitud entre los valores de los datos.
- Resaltar algún valor o elemento de los datos: con la intención de llamar la atención o resaltar un aspecto concreto, usar un color que destaca sobre el resto hace resaltar una idea. Para conseguir este efecto, es muy importante

que el resto de los colores que acompañen al color de resaltado no distraigan. Una buena técnica es dejar sin color o utilizar un tono gris para todos los elementos, frente a un color para el elemento a resaltar, aumentando el efecto.

Con esto, en todo caso se debe tratar de hacer una elección de color correcta, evitando algunos problemas que pueden derivar de un mal uso del color:

- Codificar (colorear) demasiada información de forma irrelevante, no solo no llegando a aportar nada, sino haciendo absolutamente compleja la tarea de distinguir elementos. Cuando se quiere diferenciar más de ocho elementos categóricos, se recomienda usar etiquetas textuales, siempre y cuando no saturen tampoco.
- Utilizar demasiados colores muy saturados o intensos dificulta la visión y provoca una lectura estresante del gráfico. Estos colores pueden tener cabida, pero de forma mínima y para resaltar algún aspecto.
- Seleccionar escalas de color que hacen compleja la comprensión para aquellas personas tienen deficiencias cognitivas de visión. Se recomienda utilizar algún simulador que simula estas patologías para testear los colores elegidos y certificar que funcionan en todos los casos.
- Elegir escalas de colores que no permiten diferenciar las magnitudes de los valores, qué es mayor o qué es menor. Esto suele pasar si se utilizan escalas como la escala arcoíris, en la que hay más de un color y donde hay tonalidades monótonas.

Títulos, subtítulos y etiquetas

Por lo general, en la visualización de datos los títulos, subtítulos y etiquetas se encargan de ofrecer contexto.

Todas las gráficas necesitan tener un título, el que marca el significado y el asunto principal de la figura. Además, si se genera más de una gráfica en el mismo contexto, el título debe seguir el mismo criterio, esto es: puede ir tanto integrado, normalmente en la cabecera de la gráfica, como al pie, en lo que se considera el subtítulo. Sea de una forma u otra, siempre se debe seguir el mismo criterio para todas las gráficas.

En cuanto a los subtítulos, pueden apartar un mayor detalle de contexto, aclarando alguno de los aspectos que, a simple vista, no queden totalmente claros a través del resto de elementos. No debe ser redundante, sino complementar al gráfico.

En cuanto a las etiquetas, se pueden localizar en distintos lugares de una gráfica. Aparecen como los identificadores de los valores de los ejes, en las leyendas o como denominadores de los ejes. Para este último caso, es altamente recomendable indicar la unidad de medida junto a la etiqueta, si el eje represente una medición. Si de lo contrario el eje representa valores categóricos, simplemente se aplica una etiqueta que dé título a dicho eje.

Por último, en relación con esto, no hay que despreciar los tamaños. De nada sirve disponer de todos estos elementos textuales si tienen un tamaño difícil de leer. Se debe aplicar un tamaño apto para la lectura, sin necesidad de realizar zoom sobre la gráfica.

Tipos de gráficos

Por último, en este apartado se enuncian y definen las principales formas gráficas y los tipos de gráficos más utilizados dentro del contexto de trabajo del análisis exploratorio de datos en Machine Learning (Sahay, 2016). Esta tipología será la referencia de partida para establecer el criterio de uso, según qué situaciones y contextos, y la base para el caso práctico que se implementa en este trabajo con la librería Bokeh.

Tablas

Aunque el concepto y la definición de visualización de datos se acerca más al hecho de representación mediante elementos gráficos como puntos, líneas, colores, etc., las tablas también son una forma más de visualización. Una de las formas más primitivas de visualizar datos y quizás una de las menos intuitivas, a menos que se cuente con una buena descripción de contexto, que pueda ofrecer información sobre cómo leer e interpretar los datos.

Table 3.10 Tally for Product 2 Rating

Tally for Discrete Variables: Product 2 Rating				
Rating	Count	Percent	CumCnt	CumPct
Excellent	30	13.64	30	13.64
Fair	18	8.18	48	21.82
Good	57	25.91	105	47.73
Poor	25	11.36	130	59.09
Satisfactory	49	22.27	179	81.36
Very Good	41	18.64	220	100.00
N= 220				

Diagrama de puntos

Los diagramas de puntos consisten en nubes de puntos que relacionan dos variables, haciendo uso de los ejes x e y. Cada eje corresponde al rango de valores de cada una de las variables, por lo que cada uno de los puntos corresponde a una medición concreta. Esto significa que los puntos pueden apilarse, ya que algunas mediciones pueden coincidir.

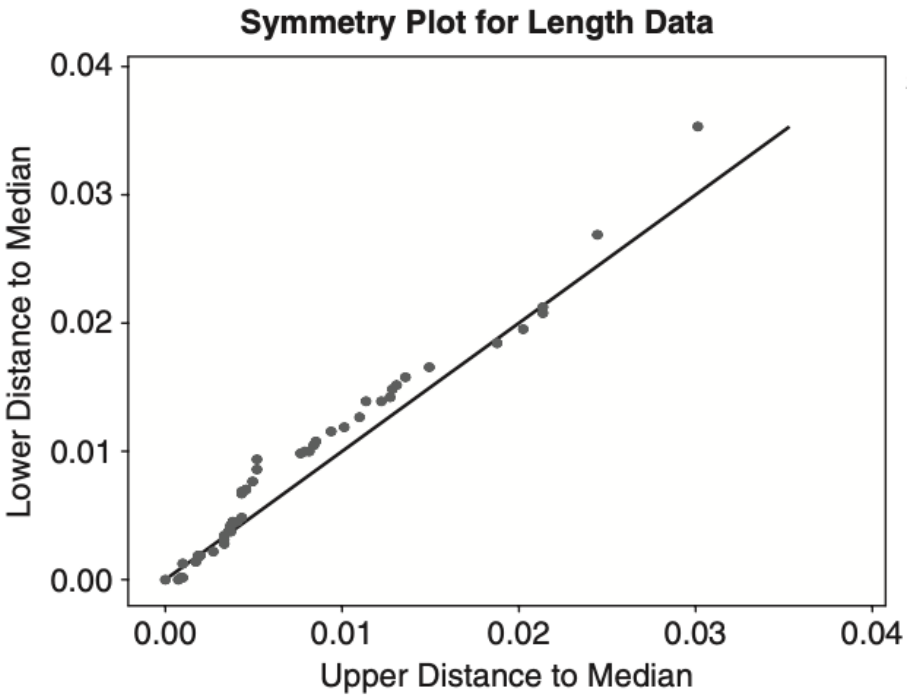


Figure 3.50 Symmetry Plot of Length Data

Diagrama de barras

Es, quizás, uno de los diagramas más conocidos y utilizados para la visualización de datos. Ideal para relacionar un rango de valores categóricos con una variable cuantitativa numérica, como puede ser el precio, la temperatura, etc. En el diagrama de barras, por lo general, el eje x representa el rango de valores, normalmente categórico, frente al eje y, donde se tiene el rango de valores cuantitativo. Además, existen variaciones de este tipo de gráfico, pudiéndose usarse con grupos de barras o en orientación horizontal.

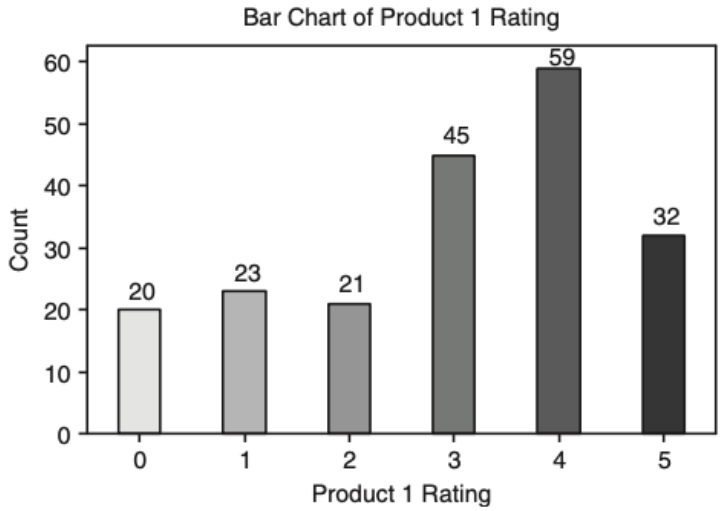
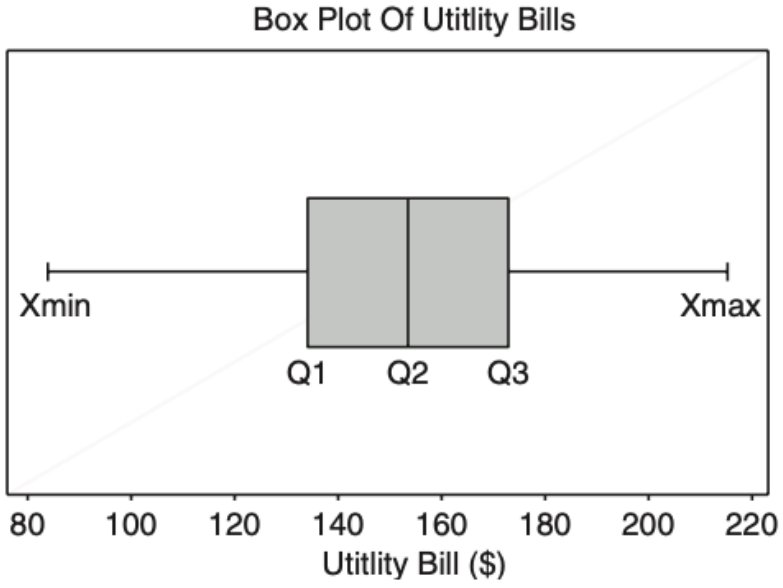


Diagrama de caja

El diagrama de caja muestra, para una variable determinada, el valor más pequeño, el más grande y sus tres cuartiles. Esto se hace mediante un segmento, cuyos extremos representan el valor mínimo y el valor máximo, dada una variable de un conjunto de datos. Los cuartiles, por otro lado, quedan representados mediante una figura rectangular, superpuesta sobre el segmento. El cuartil 1 es aquel lado de la figura que está más cercano al extremo correspondiente al valor mínimo, mientras que el cuartil 3 es el lado opuesto. Generalmente se visualiza el cuartil 2 mediante un punto dentro del rectángulo, lo que también corresponde a la mediana. Es una buena forma de representación visual para tener un resumen de estos valores.



Mapa de calor

Utilizado para representar, mediante colores, el valor de una magnitud en dos dimensiones. El color cambiará su tonalidad o su intensidad conforme el valor de la variable representada. Este tipo de representación es muy útil cuando se quiere representar algo en el espacio 2D. Por ejemplo, la ubicación del puntero del ratón del usuario en una web. El eje x representaría el ancho de la web, mientras que el eje y correspondería al alto. El color podría cambiar en función del tiempo que pasa el puntero del ratón en una coordenada. Se puede usar, por tanto, para identificar los "puntos calientes" de una página web.

Gráfico circular

El diagrama circular, también conocido como gráfica o diagrama de tarta, representa las magnitudes relativas de las partes como un todo. Esto significa que, dadas unas series de grupos de valores, se visualiza la frecuencia relativa con la que se dan cada uno de estos. En definitiva, representan porcentajes y ratios, como, por ejemplo, la dedicación porcentual de un presupuesto a distintos capítulos o asuntos económicos. Por lo general, suelen emplearse colores distintos para cada una de las porciones del gráfico, no siendo muy recomendable particionarla en exceso, ya que es complejo de leer. Para estos casos, se pude utilizar una extensión de una de las porciones, representándola en otro gráfico circular o en una barra.

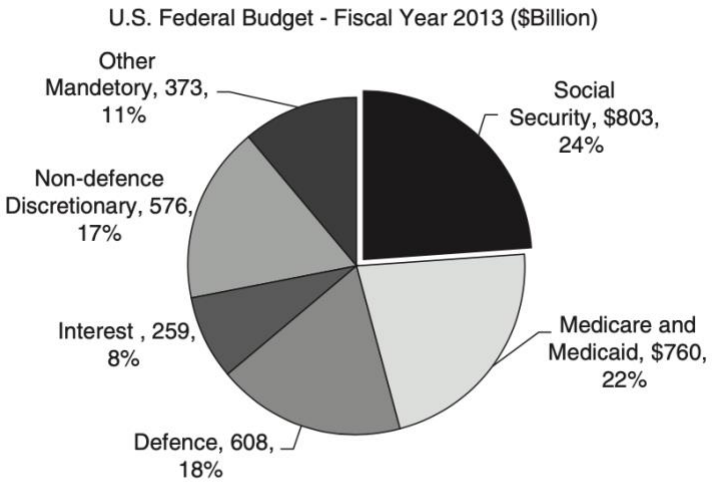


Figure 3.31 U.S. Federal Budget

Gráfico de líneas

El gráfico de líneas es una forma de representación gráfica sencilla. Generalmente se utiliza para reflejar el cambio de un valor en el tiempo, siendo el eje x una secuencia de valores temporales (horas, días, meses, años, etc.), y el eje y la medición de una variable en ese momento temporal concreto. Al final, se obtiene un punto, como el diagrama de puntos, pero estos puntos quedan unidos de manera secuencial, formando una línea. En estos gráficos se pueden incluir más de una línea, de distinto color, que representen la misma medición en diferentes condiciones. Esta forma es muy útil para hacer comparativas de la evolución y la tendencia de una variable en el tiempo.

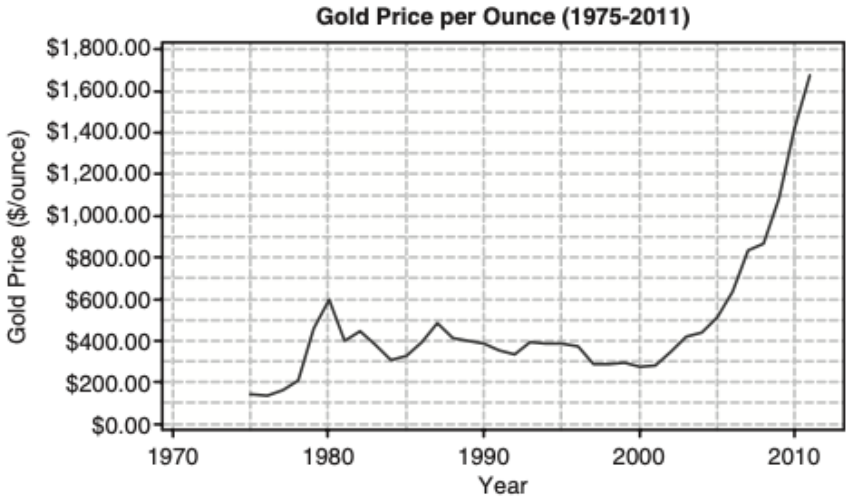
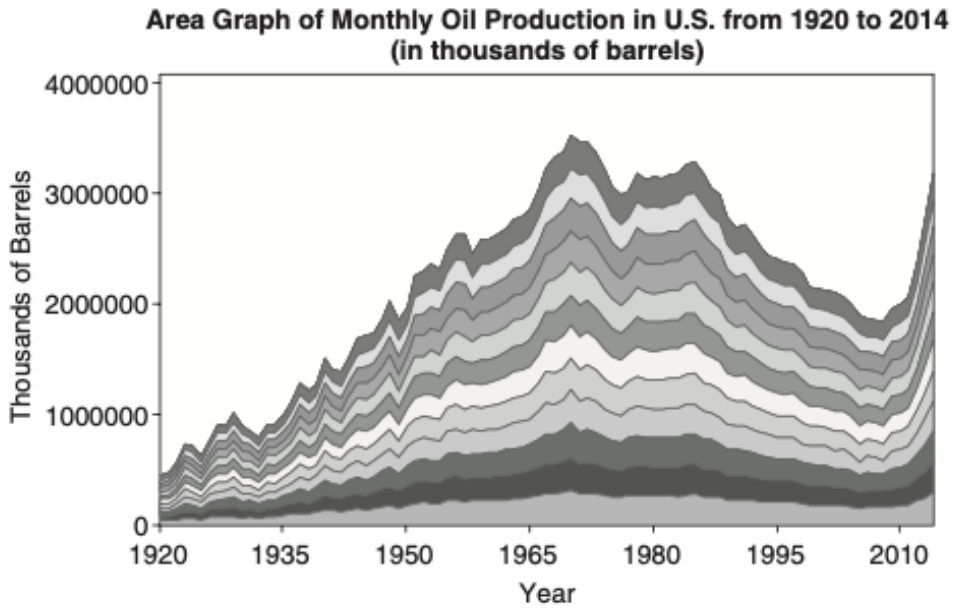


Gráfico de área

Es un tipo de combinación de gráfico de líneas y diagrama de barras, muy utilizado en series temporales. Representa la variación de las cantidades de una variable conforme pasa el tiempo, por lo que se puede utilizar para conocer la evolución o la tendencia de dicho cambio. La línea marca la unión de los momentos de medición, por ejemplo, los meses o años, en el eje x; mientras que el área coloreada que queda encerrada bajo la línea corresponde a la cantidad de la variable que se mide. Podría decirse que es un tipo de gráfico de barras continuo. Suele usarse mucho mediante áreas apiladas, asignado un color a cada una de estas, e interpretándose como un todo, pudiéndose observar las diferencias. Por ejemplo, representando la producción de barriles de petróleo a lo largo de los meses, durante varios años, para distintas empresas estadounidenses.



Criterio de uso

Tras conocer los principales elementos que entran en juego en la visualización de datos y que deben tenerse en cuenta para realizar una correcta interpretación y comunicación de estos, en esta sección se establecen criterios que se pueden implementar en cualquier análisis exploratorio de datos para cualquier dataset.

Visualización del dataset en tabla

En primer lugar, cualquier análisis dará comienzo con la lectura de un dataset. El primer paso en visualización debe ser observar los datos en su conjunto, esto es, la tabla en la que se pueden ver todas las columnas y ejemplos. A través de esto, utilizando una representación tabular, se puede hacer una comprobación rápida del número de variables, conocer sus nombres, el tipo de dato de cada una y hacer una revisión general del volumen de casos.

Muchas librerías permiten la generación de data frames a partir de la fuente de los datos (generalmente un archivo csv). A través de estas librerías se puede visualizar el data frame, el cual ya está presentado de manera tabular y se puede establecer un parámetro para ver solo los *n* primeros ejemplos del conjunto de datos.

Búsqueda de outliers con diagramas de caja

Una vez se conocen las dimensiones del dataset, lo normal es comenzar a preprocesar el conjunto de datos limpiando aquellos ejemplos en los que se detectan datos nulos y outliers. Para poder localizar los outliers, los diagramas de caja son muy recomendables. Con ellos se pueden detectar los valores atípicos, los consideramos como outliers, y eliminarlos.

Se establece un diagrama de caja por cada variable o columna del dataset que vaya a ser tenida en cuenta en el análisis exploratorio. También se pueden visualizar todas juntas, en un multidiagrama.

Histogramas de barras

A través de los histogramas, representados mediante diagramas de barras, se puede conocer la distribución de las variables del conjunto de datos y qué frecuente es que se dé un valor u otro. Con esto se podrá valorar si se van a disponer ejemplos sesgados, porque no existe cierto equilibrio entre las distintas muestras, por ejemplo, para entrenar un modelo predictivo.

El histograma se genera para cada una de las variables del dataset que se vayan a tener en cuenta.

Proporción de casos con gráficos circulares

En el caso de variables categóricas, se pueden establecer sumatorios que cuenten el total de casos de cada tipo y puedan ser representados en gráficos circulares, junto al porcentaje correspondiente para cada dato. Con este tipo de gráficas también será posible conocer el volumen de casos para determinados valores y detectar algunas descompensaciones en el dataset. Además, en la fase de

análisis, gracias a estas gráficas se pueden obtener respuestas interesantes a preguntas que se puedan ir planteando.

Relación de variables con diagramas de puntos

Una forma de comprobar el tipo de relación que puede existir entre variables del dataset es a través del diagrama de puntos. Enfrentando dos variables, cada una de ellas asignada a los ejes x e y, se puede obtener la forma que sigue la distribución. De manera visual se puede valorar si siguen una relación lineal o logarítmica, o simplemente no se puede certificar relación alguna. Gracias a este análisis se pueden seleccionar atributos que serán usados para entrenar un modelo predictivo, por ejemplo.

También se puede hacer uso de este diagrama de puntos para tratar de representar la misma relación teniendo en cuenta distintos casos categóricos, los cuales se diferencian por el color de los puntos. Gracias a esto, se pueden detectar clusters de ejemplos.

Correlación con mapas de calor

Otra forma de estudiar y analizar la relación entre variables de un conjunto de datos es a través del estudio de la correlación. Para esto, una vez se hace el cálculo de la correlación de variables, esta puede representarse mediante un mapa de calor, siendo muy fácil interpretarla de forma visual. Gracias a esta gráfica se puede hacer la selección de variables que se usarán para entrenar modelos predictivos.

Caso práctico: Bokeh library

A partir del marco teórico esbozado, donde se han descrito los elementos gráficos que protagonizan la visualización de datos y se han trazado una serie de criterios de uso para la realización de un análisis exploratorio, ahora se ponen en práctica con la librería de visualización Bokeh.

Este ejercicio se incluye completo en un notebook de Jupyter.

Conclusiones

Para este trabajo se habían fijado los siguientes objetivos:

- Describir los diferentes tipos de formas gráficas que se utilizan para la visualización de datos.
- Establecer una serie de criterios para el correcto uso de la visualización.
- Desarrollo de ejemplos y casos de uso.

Se puede concluir que todos ellos han sido satisfechos a través de un estudio del estado del arte, haciendo una aproximación a la visualización desde un punto inicial de aprendizaje, conociendo los diferentes elementos técnicos necesarios para una buena y correcta ejecución, estableciendo criterios de uso para según qué situaciones y finalizando con una puesta en práctica de todo lo aprendido.

Además, se ha optado por una librería como Bokeh, la cual ofrece bastante libertad a un nivel medio-alto, ya que no cuenta con métodos previamente definidos que permitan la obtención directa de visualizaciones, sino que es necesario su construcción. De esta forma, se pone en práctica la teoría y se proponen e implementan una serie de funciones que pueden ser utilizadas para un análisis exploratorio de datos.

Como conclusión, también se ha llegado a la reflexión de que es lógico que las librerías que implementan visualización, y que incluyen ya una serie de funciones predefinidas, tengan tan buena acogida en la mayoría de los proyectos. Es muy cómo contar con herramientas que ya hacen buena parte del trabajo que aquí se ha realizado para implementar funciones que no se incluyen en Bokeh.

Por último, también se ha podido conocer la importancia y el valor añadido de la visualización en cualquier análisis exploratorio de un dataset. Ofrece muchas pistas sobre el objeto de estudio y ayuda a plantear preguntas y formular hipótesis que puedan ser posteriormente evaluadas. De esta forma, es lógico que sea un proceso técnico que esté presente en todo proyecto KDD y a lo largo de todo su ciclo de vida.

Referencias

- A Complete Guide to Data Visualization in Python With Libraries, Chart, Graphs & More. (s. f.). Simplilearn. https://www.simplilearn.com/tutorials/python-tutorial/data-visualization-in-python
- Bokeh. First Steps. (s. f.). https://docs.bokeh.org/en/2.4.3/docs/first_steps.html
- Bokeh. Boxplot. (s. f.). https://docs.bokeh.org/en/latest/docs/gallery/boxplot.html
- Cómo mejorar tu web gracias a los mapas de calor. (s. f.). Devservice.es. https://www.devservice.es/blog/mapas-calor-web/
- Codecademy. Exploratory Data Analysis: Data Visualization. (s. f.). https://www.codecademy.com/article/eda-data-visualization
- Guía de visualización de datos: definición, ejemplos y recursos de aprendizaje. (s. f.). Tableau. https://www.tableau.com/es-es/learn/articles/data-visualization
- Iliinsky, N., & Steele, J. (2011). Designing Data Visualizations: Representing Informational Relationships. Van Duuren Media.
- Sahay, A. (2016). Data visualization, volume i: Recent trends and applications using conventional and big data. Business Expert Press.
- Wilke, C. O. (2019). Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures (1.). O'Reilly Media.