Procesamiento de información.

Guillermo Ruiz

¿Qué es la **información**? Este concepto ha sido muy usado en la sociedad moderna y tuvo un papel fundamental para su desarrollo. Una sociedad es imposible sin el intercambio de información. Aunque todos tenemos una buena idea de lo que representa esta palabra, es importante definirla de manera precisa, pero primero debemos hablar de **datos**.

Los datos se definen como símbolos que representan propiedades de objetos o fenómenos. Se obtienen de la observación y el intercambio de energía. Los datos son inútiles si no se representan de manera adecuada.

La información se extrae de los datos y se presenta en descripciones que responden preguntas como quién, qué, cuándo y cuántos. Los sistemas de información almacenan, ordenan, buscan y recuperan datos. La información se representa en un espacio menor que los datos.

El **conocimiento** es la generalización de la información para hacer predicciones. Requiere menor espacio que la información. Se puede adquirir de alguien que la tenga, de instrucciones o de la experiencia.

Finalmente, la **sabiduría** es la acumulación de conocimiento que permite su aplicación en nuevos problemas. Otorga la habilidad de ver lo que fue, es y será.

Jerarquía de Datos

Poseer información en el momento correcto es valioso para la toma de decisión por las implicaciones que conlleva. El modelo DIKW (Data-Information-Knowledge-Wisdom) se propuso para delinear una jerarquía de los sistemas de información y los compromisos de cada nivel. Se representa por una pirámide cuya base son los datos. El siguiente nivel es la información, seguido de conocimiento y en la punta, la sabiduría. Para pasar de un nivel a otro se necesita un entendimiento de la situación.

Figure 1: jerarquia_datos.jpg

Gran cantidad de datos

En la actualidad, existe una fuente inagotable de datos: el internet. En él, se crea nueva información todo el tiempo, de toda clase de temas y en formatos muy variados. Las principales fuentes son las redes sociales, en los que los usuarios generan texto, audio, videos e interactúan unos con otros. Las empresas generan otra gran cantidad de datos que recopilan de los usuarios como compras, ventas, intercambios, búsquedas, entre otras. Además, existen cada vez más dispositivos que su función es recopilar, almacenar y procesar datos como las cámaras de vigilancia, equipos de monitoreo del clima o la contaminación, satélites, checadores digitales entre otros.

Toda esta información debe ser almacenada, procesada y analizada para convertirla en conocimiento. A lo largo del curso se verá que los métodos de aprendizaje automático serán un tema central pero que no son lo más importante. La preparación de datos para poder ser leídos por estos algoritmos y la interpretación de los resultados son igualmente importantes.

Información estructurada

Esencialmente existen dos tipos de información, la estructurada y la no estructurada. Los datos estructurados tienen un formato fijo que facilita su lectura y su procesamiento. Por ejemplo, las bases de datos relacionales de un sistema de información son el resultado del análisis de un problema del dominio (bancario, aseguradora, escolar, etc.) y la construcción de las tablas de la base de datos que relacionan la información representan la solución de información a posibles preguntas que los usuarios pudiesen tener, es decir, preguntas como:

cuántas compras realizó el cliente x en cierto día. Las respuestas a este tipo de preguntas se pueden obtener haciendo consultas a la base de datos, a saber, con expresiones del lenguaje SQL.

Una revisión elemental de los conceptos básicos de las bases de datos relacionales y un poco de historia se encuentra en El modelo relacional de base de datos

Información no estructurada

Por otro lado, a la información que no posee elementos como los de las bases de datos se le conoce como información no estructurada y puede tener formatos diferentes. Por ejemplo los documentos escritos, ya que en su interior pueden contener cualquier tipo de información. En este curso daremos énfasis en el procesamiento de información no estructurada y de las técnicas para transformarla a espacios que permita su manipulación y sea fácil de acceder, por ejemplo, el modelo de espacio vectorial ampliamente usado por sistemas de recuperación de información.

Recolección de datos

Los datos se pueden extraer de muchas fuentes como son: archivos de una computadora, recolectados por sensores especiales, de bases de datos relacionales, generada por operadores humanos, realizar encuestas y muchas otras. En el caso de las encuestas, se realizan preguntas puntuales para generar una muestra de la población que se desea estudiar. Por ejemplo, encuestas poblacionales, donde se puede conocer el total de la población de un país, su distribución de hombres y mujeres, y otros datos de interés como el nivel de escolaridad, edad, etc. En este caso, la fuente de información son las personas encuestadas.

Otra fuente de información valiosa es internet y, en especial, las redes sociales donde hay información opinable, donde las personas expresan sus opiniones sobre productos, servicios, personas, entidades, temas, etc., lo cual es una fuente de datos valiosa de la percepción de las personas, útil para la toma de decisiones para mercadotecnia, análisis de reputación, política pública, etc.

Análisis de datos

El análisis de los datos puede ser simple como calcular histogramas o más elaborados con técnicas de agrupamiento o clasificación de datos, encontrar correlaciones entre los datos, etc. Cada objeto estudiado se describe mediante **variables** o atributos que representan sus propiedades. Todas las variables de un objeto se llama **registro**. A la colección de todos los registros se le conoce como **conjunto de datos**, que generalmente se representa por una tabla donde cada fila es un registro y las columnas son los atributos.

Existen diferentes tipos de variables, según el tipo de datos que contienen. Los tipos más comunes son: Variable nominal. Indica una categoría a la que pertenece el objeto. La categoría pueden ser palabras o números. Sus elementos no tienen un orden. Variable binaria. Toma los valores de 0 ó 1. Variables Ordinales. Los valores que toma tienen un orden. Variables de intervalo. Son variables numéricas donde cada número representa un intervalo.

Aprendizaje Supervisado

En el aprendizaje supervisado se tiene los datos y cada uno de ellos tiene asociado un valor objetivo. Puede ser una etiqueta, o un valor numérico. Por ejemplo, en una colección de textos podemos asignarles una etiqueta que nos diga el tema que abordan. Al historial de un paciente, podemos asignarle un valor (por ejemplo, del 1 al 10) que nos diga qué tan sano se encuentra.

La **clasificación** es uno de los problemas más comunes en el aprendizaje supervisado. En él, se tiene que asignar la etiqueta correcta a los datos. Los algoritmos de aprendizaje se alimentan de una gran colección de datos que ya fueron previamente etiquetados por expertos. Los algoritmos buscan patrones que les permitan clasificar de manera correcta datos nunca antes vistos ellos. A este proceso se le llama **entrenamiento**.

Cuando el algoritmo debe asignar un valor continuo a los datos, se le llama **regresión**. Por ejemplo, predecir el valor de cierta moneda, o la temperatura del día siguiente.

Aprendizaje no supervisado

En el aprendizaje no supervisado, se tienen los datos como en el caso anterior, pero no tienen asociado ningún valor. El objetivo entonces es agrupar los datos parecidos y diferenciarlos del resto. A este proceso se le llama **clustering**. Los clusters formarán una partición de los datos y cada cluster es un conjunto que contiene objetos similares entre ellos pero distintos a los que hay en los demás clusters. Por ejemplo, tenemos un conjunto de datos que contiene artículos de noticias provenientes de diarios nacionales. Sólo tenemos los textos pero no están etiquetados o separados por tema. El objetivo es crear un algoritmo que sea capaz de separarlos por tema, que ponga todos los de política en un conjunto, en otro los de deportes, y así para los temas de finanzas, internacionales y salud. En total 5 clusters y no nos importa el orden, el primer cluster puede ser finanzas, salud o cualquier otro.

En este problema tenemos el inconveniente de que debemos elegir el número de clusters de antemano. En el caso anterior fueron 5 clusters pero no queda claro qué debería hacer el algoritmo con las noticias policíacas, ¿se deben poner en política, o salud? Elegir el número correcto de clusters es vital para el buen funcionamiento del algoritmo pero no hay una clara estrategia para decidirlo.

Preparación de datos

Algunas veces los datos con los que vamos a trabajar ya se encuentran listos para ser usados, lo único que tenemos que hacer es descargarlos. Otras veces los datos que necesitamos no existen y debemos recolectarlos. Después se tienen que hacer una serie de pasos para que sean útiles, como son limpieza, revisión de errores y valores faltantes, eliminación de duplicados entre otros.

Presentación (visualización) de datos

Para que los datos tengan un sentido, deben ser presentados de una manera adecuada, como puede ser una gráfica. Generalmente, las gráficas se adaptan adecuadamente para dar una percepción de la estructura, de los patrones, tendencias y relaciones en los datos. Las tablas también son una forma de presentar los datos de manera concentrada.

Descripción de los datos

En estadística podemos distinguir dos tipos de datos, los datos cuantitativos y los datos cualitativos. Los cuantitativos se usan para los cálculos numéricos y para definir, por ejemplo, los ejes en las gráficas. Los cualitativos se usan para las agrupaciones de las muestras (observaciones) en tablas o gráficas.

Los estadísticos muestrales más comunes para caracterizar los datos se enfocan en las medidas de ubicación, por ejemplo, el promedio, la mediana, la moda; y medidas de dispersión como el rango, varianza y desviación estándar.

Como un primer acercamiento a la caracterización de los datos, el conteo e histogramas ayudan a describir los datos cuantitativamente, y nos da una visión inmediata de las características de los datos. Entre mayor es el tamaño de las observaciones, la distribución de los valores de los datos se apreciará mejor.

Teoría de la información

Un escrito en español tiene una estructura y sigue las reglas gramaticales del español. Esto es cierto en determinada medida, ya que las normas no siempre se respetan en textos informales o con errores y los idiomas están en constante evolución, lo que les permite adaptarse a los cambios en la sociedad. Aún con estos cambios, es fácil distinguir cuando un texto está en español o en algún lenguaje extranjero. Si vemos a los caracteres de un escrito como una función de distribución, podemos observar que no todos tienen la misma probabilidad de ocurrir, pues dependen de los caracteres que los anteceden.

Si en un texto en español se enmascara un carácter y se pide a una persona (que habla español) que lo adivine, seguramente acertará en menos de 3 intentos. Esto nos dice que predecir el siguiente carácter es una tarea no muy complicada. Algo similar sucede con las palabras.

Cuando tenemos el texto de un documento, podemos identificar que unas palabras son más importantes que otras. Palabras como los artículos (el, la, los, …) nos ayudan en la lectura y comprensión del texto pero no son muy informativas. Palabras que no ocurren con frecuencia suelen ser las más significativas, por ejemplo, si vemos las palabras "transformación, matriz, conmutativo", seguramente estamos ante un texto de álgebra lineal.

Dado un mensaje, diremos que su cantidad de información será alto si su contenido es sorprendente y será bajo si su contenido es predecible. Por ejemplo, si recibimos un mensaje que dice que mañana no vamos a ganar la lotería, diremos que tiene muy poca información, ya que no ganar la lotería es lo más probable. Por el contrario, un mensaje que diga que mañana vamos a ganar la lotería sería muy valioso.

Podemos aplicar este principio para definir de manera numérica la cantidad de información en un mensaje de acuerdo a que tan predecibles son las palabras que contiene. Eso es lo que nos dice la entropía de un texto.

Entropía

Es una medida de la incertidumbre de una variable aleatoria X. Mide el grado de sorpresa al ocurrir un evento. Mientras más raro el evento ocurrido, mayor será la sorpresa. La entropía de una variable aleatoria discreta se define como

$$H(X) = -\sum_{x \in X} p(x) \cdot \log(p(x))$$

donde $p(x)$ es la función de probabilidad; en computación se usa log base 2 y así, la entropía dice el menor número de bits que se necesitan para mandar un mensaje. En la definición anterior usamos la convención $0 \cdot \log(0) = 0$

Ejemplo: Sea E el resultado de dos lanzamientos de una moneda y sea X la variable aleatoria de contar el número de "águilas" (A) del resultado. Sea S el espacio muestral \$S = {AA, AS, SA, SS} \$

$$p(0) = P(X = 0) = \frac{1}{4}$$
$$p(1) = P(X = 1) = \frac{1}{2}$$
$$p(2) = P(X = 2) = \frac{1}{4}$$

$$p(3) = P(X = 3) = 0$$
$$\vdots$$

Podemos calcular la entropía del lanzamiento de las dos monedas:

$$H(X) = -\sum_{x \in X} p(x) \log p(x) = -\left[\frac{1}{4}\log\frac{1}{4} + \frac{1}{2}\log\frac{1}{2} + \frac{1}{4}\log\frac{1}{4}\right] = 1.5$$

También la entropía puede considerarse como la cantidad de información promedio que contienen los símbolos usados. Los símbolos con menor probabilidad son los que aportan mayor información. Por ejemplo, si consideramos a las palabras de un texto como sistema de símbolos, las palabras frecuentes como son las **stopwords** (artículos y preposiciones, entre otras) aportan poca información a diferencia de palabras de contenido que describen al documento. De ahí que si eliminamos palabras que aportan poca información, el documento se comprime y se sigue teniendo cierto grado de comprensión. La entropía es máxima cuando todos los símbolos aportan la misma cantidad de información. Por ejemplo, para encontrar la entropía del siguiente texto "cantata" encontramos que tenemos un total de 7 elementos de los cuales.

- `c` aparece una vez
- `a` aparece 3 veces
- `n` aparece una vez
- `t` aparece dos veces

Si X es la v.a. de elegir una letra al azar de la palabra *cantata*, entonces

$$P(X = c) = 1/7$$
$$P(X = a) = 3/7$$
$$P(X = n) = 1/7$$
$$P(X = t) = 2/7$$

La entropía de X sería

$$H(X) = -\sum_{x \in X} p(x) \log p(x) = -\left[\frac{1}{7}\log\frac{1}{7} + \frac{3}{7}\log\frac{3}{7} + \frac{1}{7}\log\frac{1}{7} + \frac{2}{2}\log\frac{2}{2}\right] = 1.84237$$

y para el caso del texto *sara* se tendría una entropía de 1.5, menor entropía, lo cual nos sugeriría que hay menor riqueza **léxica**, de acuerdo a ese conjunto de símbolos. Podemos considerar a la entropía como una medida de la riqueza del vocabulario cuando lo aplicamos a documentos.

```
Texto:  cantata
Entropía: 1.8423709931771084

Texto:  sara
Entropía: 1.5
```

Entropía relativa

La entropía relativa o divergencia de Kullback-Leibler es una medida de distancia entre dos distribuciones, $D(p||q)$. No es propiamente una métrica, pues es no simétrica y no cumple la desigualdad del triángulo. Mide la similitud o diferencia de dos funciones de distribución.

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

En la definición anterior usamos la convención $0 \log(0/0) = 0$. Resulta que la entropía relativa es no negativa y es cero si y sólo si $p(x) = q(x)$. Es útil si queremos saber qué tan bueno es un modelo al predecir una distribución: Si $p(x)$ es la distribución de los datos observados y $q(x)$ es una aproximación obtenida por un modelo, entonces la entropía relativa es el promedio de la diferencia en el número de bits necesarios para codificar p usando q.

Información mutua

La información mutua $I(X, Y)$, es una medida de la cantidad de información que una variable aleatoria contiene de otra variable aleatoria, es decir, es la reducción de la incertidumbre de una variable aleatoria debido al conocimiento de la otra variable.

$$I(x_i, y_i) = \log \frac{p(x_i, y_i)}{p(x_i)p(y_i)}$$

La información mutua es la entropía relativa entre la distribución conjunta y el producto de la distribución $p(x_i)p(y_i)$. Donde $p(x_i)$ y $p(y_i)$ se pueden obtener de las frecuencias relativas de los datos.

La frecuencia relativa n_i es el cociente entre la frecuencia absoluta f_i y el número total de datos N

$$p(x_i) = n_i = \frac{f_i}{N}$$

La información mutua se ha aplicado en el descubrimiento de las asociaciones de palabras que para las personas son de sentido común, pero para una computadora no lo son. En el dominio médico, considere las asociaciones de palabras doctor-paciente y doctor-panadería. Será más probable la primera asociación que la segunda. En este caso, la información mutua nos ayuda a descubrir si la asociación es significativa o no, $I(X, Y) > 0$

$$I(\texttt{doctor}, \texttt{paciente}) = \log \frac{p(\texttt{doctor}, \texttt{paciente})}{p(\texttt{doctor})p(\texttt{paciente})}$$

donde $p(\texttt{doctor}, \texttt{paciente})$ es la probabilidad de que co-ocurran las dos palabras en una ventana de palabras definida. Por ejemplo, usando una ventana de dos ($w = 2$), es decir, de bigramas, para el siguiente texto tenemos el conteo de 2 para la pareja (bigrama) $(\texttt{doctor}, \texttt{paciente}) = 2$

Texto:

El doctor y paciente se saludaron en el hospital. El doctor y paciente no están de acuerdo

Preprocesando el texto convirtiendo a minúsculas y removiendo *stopwords*, el texto quedaría:

```
doctor paciente saludaron hospital doctor paciente no estan acuerdo
```

```
Que tiene como unigramas a:
```

```
{'doctor': 2,
 'paciente': 2,
 'saludaron': 1,
 'hospital': 1,
 'no': 1,
 'estan': 1,
 'acuerdo': 1}
```

```
Y los bigramas son:
```

```
{'doctor paciente': 2,
 'paciente saludaron': 1,
 'saludaron hospital': 1,
 'hospital doctor': 1,
 'paciente no': 1,
 'no estan': 1,
 'estan acuerdo': 1}
```

```
La información mutua queda:


{'doctor paciente': 2.3398500028846247,
 'paciente saludaron': 2.3398500028846247,
 'saludaron hospital': 3.3398500028846247,
 'hospital doctor': 2.3398500028846247,
 'paciente no': 2.3398500028846247,
 'no estan': 3.3398500028846247,
 'estan acuerdo': 3.3398500028846247}
```

Note que la IM de los bigramas que aparecen una sola vez es alta, por lo que es común agregar la condición de mostrar sólo los bigramas que aparecen cierto número de veces en el texto.