



Matemáticas

Computación II

6º Semestre

Unidad 1. Introducción a la minería de datos

Clave

05142423/06142423

Universidad Abierta y a Distancia de México





Computación II

Unidad 1. Introducción a la minería de datos

Índice

<i>Presentación.....</i>	<i>3</i>
<i>Desarrollo de temas</i>	<i>4</i>
<i>Grandes Datos y Causalidad</i>	<i>4</i>
<i>Conceptos básicos</i>	<i>6</i>
<i>Análisis exploratorio de datos</i>	<i>9</i>
<i>Cierre de la unidad</i>	<i>10</i>
<i>Fuentes de consulta.....</i>	<i>10</i>



Computación II

Unidad 1. Introducción a la minería de datos

Presentación

La minería de datos se ha convertido en un tópico recurrente hoy en día, debido en buena parte a la accesibilidad a la tecnología que genera una gran cantidad de estos últimos. Todos podemos acceder a equipo de cómputo, celulares, GPS, los cuales generan información masiva que una vez estructurada puede ser interpretada y usada en consecuencia. Pero para poder hacer esto, requerimos una forma inteligente, por decirlo de forma muy vaga o estandarizada por ponerlo de otra forma ya que es muy común perderse la imagen del bosque por estar viendo los árboles. Es decir, es fácil perderse en los detalles cuando lo que queremos es la imagen general que estos datos en su contexto global pueden ofrecer.

En esta unidad abordaremos algunos aspectos de un tema que está de moda que es el de la minería de datos, entendida esta como la búsqueda de patrones o estructuras en los datos que podamos recolectar o medir sobre un fenómeno o experimento particular con el fin de que nos permitan hacer una explicación o una predicción en el mejor de los casos. Presentaremos una serie de algoritmos básicos y muy conocidos que pueden ilustrar el funcionamiento básico de la minería de datos.

Competencia específica

Identifica las variables más representativas que prometen representar correlaciones causales en un conjunto de datos a través del análisis exploratorio de datos.

Logros

- Comprender la naturaleza de los grandes datos.
- Comprender el concepto de los análisis exploratorios de datos.
- Determinar una descripción estadística genérica sobre un conjunto de datos.



Computación II

Unidad 1. Introducción a la minería de datos

Desarrollo de temas

La minería de datos, como se estableció en la presentación de esta unidad, la entenderás como el proceso de buscar estructura y patrones en los datos generados a partir de alguna medición que genera una cantidad tan grande de datos como para no poder extraer relaciones funcionales entre ellos de manera trivial.

A lo largo de esta unidad repasarás algunos conceptos para entender este tema que está tan de moda, particularmente entender ¿qué es la minería de datos? así como otra rama de la computación íntimamente ligada a la minería y derivada de la inteligencia artificial que es el aprendizaje computacional.

Un aspecto fundamental que tienes que repasar es la relación de la minería con la estadística, ya que notarás una gran relación sobre todo cuando la estadística donde lo que usualmente hacemos es recibir un conjunto de datos para representarlos a través de diferentes métricas asociadas a dicho conjunto. Y en ese aspecto la minería y estadística son indistinguibles, se usan muchos conceptos de esta última en la minería, pero a su vez, esta, presenta una característica distinta que es la de echar mano de cualquier recurso para entender los datos que estemos estudiando y, aunque no es cuantificable, su notoria naturaleza gráfica, eso es, en los métodos de la minería de datos es muy común encontrar que muchos de ellos echan mano de métodos gráficos para presentar sus resultados. Esto podrá parecer poco relevante en el momento, incluso no es realmente una distinción ya que se puede hacer lo mismo en estadística, excepto que en la minería es una práctica común hacer uso de métodos visuales primero y analíticos después con el fin de ir descubriendo la estructura que tienen los datos a diferencia de la estadística donde se plantean las hipótesis antes para comprobarlas después. La minería de datos intenta, pues, dar un método más robusto de explicación de los datos.

Grandes Datos y Causalidad

Correlación no implica Causalidad: Una reflexión sobre la ética en la Minería.

Un aspecto extremadamente importante que seguramente ya revisaste en cursos de Estadística es la relación entre correlacionar un grupo de variables y obtener una relación funcional entre ellas.



Computación II

Unidad 1. Introducción a la minería de datos

En la sección “Para Saber Más” puedes leer (Morozov, 2013) en el que se hace un excelente análisis de lo que significa hacer un análisis de un conjunto de datos y cómo podría hacerse muy fácilmente una interpretación errónea de este análisis. Otro conjunto de ejemplos visuales se puede ver en (Vigen, 2014) en el que se intenta dejar claro que el que dos valores se parezcan en el tiempo no implica que uno sea causa de otro.

Y es que bajo un análisis de regresión encuentres que dos variables estén fuertemente correlacionadas no implica que una sea función de la otra, bien puedes estar viendo una simple casualidad en la que dos variables se expresaron con un conjunto de valores en el tiempo en que fueron registradas.

Es importante hacer un llamado en este punto a la responsabilidad ética que se tiene cuando se analizan datos. No sólo de ética profesional sino de también de responsabilidad social. Hoy en día es casi natural recabar datos desde muchas fuentes, y muy diversos, particularmente datos personales que pudieron haber sido o no otorgados con el consentimiento de su poseedor. En el caso de que así fuera es importante apelar a la responsabilidad de que tan sensibles son los datos analizados, y más aún apelar a la responsabilidad de las conclusiones a las que se puede llegar a través de un análisis de minería de datos con ellos.

A lo largo de esta unidad vamos a revisar los temas de análisis exploratorio así como el planteamiento de la pregunta que queremos hacer sobre los datos. Esto nos servirá para poder escoger alguno de los múltiples algoritmos que existen para modelar la información que tengamos.

Para entender estos conceptos puedes revisar las siguientes ligas.

Morozov, E. (2013) El Reto de manejar grandes bases de datos. *El País*.

http://elpais.com/elpais/2013/06/24/opinion/1372068111_079679.html

Observa los ejemplos de la siguiente liga. ¿Qué puedes pensar a partir de esas gráficas y la insistente presentación en los medios de comunicación sobre todo tipo de genes (genes de la obesidad, risa, divorcio, etcétera)?

Vigen, T. (2014) *Correlaciones espurias En línea*. Consultado el 05 de marzo de 2026.

<http://tylervigen.com/spurious-correlations>



Computación II

Unidad 1. Introducción a la minería de datos

Hasta este punto sólo hemos descrito el panorama general en el análisis de datos, y algunos aspectos cruciales para su correcta interpretación: Distinguir la correlación de la acusación. Tal vez en el curso de estadística hayan hecho también mucho énfasis en este aspecto, pero podrás observar con el paso del tiempo que nunca está de más.

A continuación, veremos otros conceptos introductorios complementarios como el tipo de preguntas que vamos a hacer o el tipo de algoritmos que vamos a usar.

Conceptos básicos

La primera fase y la más obvia es el proceso de obtención de los datos que se desean analizar. Los métodos son variados y van desde el diseño y levantamiento de encuestas, pasando por la realización de experimentos específicos o la investigación y recabación de datos a partir de otros conjuntos de datos. Este proceso específico escapa al ámbito de esta asignatura, pero sí nos permite plantear y describir lo que es un **conjunto de datos** (*dataset* es el término en inglés). Los conjuntos de datos son, precisamente eso, un arreglo con los datos organizados en una matriz donde cada renglón representa un dato y cada columna representa un atributo específico. Cada uno de estos atributos pertenecen a un dominio específico, así como a una categoría estadística concreta, es decir, pueden ser valores reales o categóricos.

Los **atributos** pueden ser la altura, la temperatura, el panorama del día, etcétera. Los atributos y conjuntos de atributos contribuyen a describir una instancia específica. Lo que nos lleva a la definición de **instancia** que es un arreglo de atributos ordenados de forma específica que ejemplifica como son los datos de los cuales hay que discernir una estructura.

La definición de **concepto** es la más abstracta ya que es la relación o estructura en los datos que queremos abstraer. Por ejemplo, si tuviéramos un conjunto de datos que describiera a un conjunto de mamíferos en términos de su tamaño, temperatura sanguínea, tipo de alimentación, tipo de reproducción, el **concepto** de carnívoro vendría dado por todos los ejemplos en los que se pudiera distinguir el tipo de alimentación basada en una dieta cárnica, los herbívoros serían los mamíferos cuya dieta se basara en hierbas, etcétera. Para una mayor descripción sobre estos conceptos puedes consultar (Witten, 2005).

Muchos de los algoritmos que revisarás requieren de un proceso de **entrenamiento** ya que el algoritmo está diseñado para abstraer el concepto de un conjunto de datos por definir, para lo



Computación II

Unidad 1. Introducción a la minería de datos

cual tiene que ajustar alguna serie de parámetros que dependen de cada conjunto de datos específicos.

Estos procesos de entrenamiento constan de una fase de entrenamiento, esto es, requerimos dividir el conjunto de datos de entrada en dos partes, una que es el **conjunto de entrenamiento**, concretamente hablando consta de una serie de **ejemplos** para determinar las reglas que deben generalizar y el conjunto de validación que es una serie de instancias en las que probamos la validez de las reglas encontradas.

Lo que usualmente queremos es usar un algoritmo y construir un modelo que pueda abstraer de nuestro conjunto de entrenamiento la estructura subyacente para podamos describir estos datos de forma genérica. La forma de construir este modelo de manera correcta es un proceso cuasi artesanal para el cual existen varias técnicas, pero a grandes rasgos depende mucho del conocimiento profundo del problema. Decimos que nuestro modelo está **subentrenado** si detenemos el proceso de entrenamiento antes de que el indicador de generalización (alguna medida sobre que tan bien clasifica ejemplos nuevos) rebase algún umbral predefinido.

Decimos que el modelo está **sobreentrenado** cuando esperamos que esta medida de eficiencia sea muy alta pero sólo hacemos el entrenamiento con un subconjunto de datos, es decir, cuando incorporamos la idiosincrasia del conjunto de entrenamiento en el modelo. Este fenómeno de sobreentrenamiento lo podemos reproducir fácilmente cuando probamos nuestro modelo con los mismos ejemplos con los que lo entrenamos. Para más información puedes revisar (Witten, 2005).

Datos e información

A lo largo de la unidad pasada y en esta en particular trataremos el tema de datos e información. En la asignatura pasada nos preguntamos sobre los modelos de cómputo en abstracto y determinamos que eran las estructuras de datos. Aquí nos hemos hecho la pregunta sobre obtener información útil a partir de un conjunto de datos, es en ese sentido que vale hacerse la pregunta sobre que significa información y la pregunta retórica que plantearemos a continuación ayuda a determinar el concepto de información.

Imagina que recibimos una señal extraterrestre ¿Cómo podemos determinar si es un mensaje producto de una inteligencia extraterrestre o simplemente ruido cósmico?



Computación II

Unidad 1. Introducción a la minería de datos

En ese sentido podemos decir que la información es un conjunto de valores compuesto por datos estructurados de alguna forma tomados de una población. No toda la información se ve igual y depende mucho del problema que se tenga a la mano para determinar cuanta información requerimos. Entender los valores de las variables para entender la dinámica y problemática de un salón de clases varía en órdenes de magnitud que el de toda una universidad y esto a su vez que el conjunto de datos que pueda generar un país entero.

El tipo de datos para describir la dinámica de un proceso físico es enteramente distinto que el tipo de datos requeridos para entender dinámicas sociales. Aunque al final es posible que el tipo de dinámica sea cualitativamente parecido.

Es decir, el contexto en el que están inscritos los datos puede determinar que tipo de información tenemos a la mano, este mismo contexto nos va a ayudar a determinar que pregunta es la queremos hacerle a los datos recabados.

Algunas de las preguntas que los datos nos a permitir determinar y encontrar su respuesta, se encuentran las siguientes:

Descriptiva: Describir un conjunto de datos...

Exploratoria: Encontrar relaciones entre los diferentes atributos

Inferencial: A partir de un subconjunto de datos de una población poder inferir algo de la población en general

Predictiva: Generar un modelo de un fenómeno a partir de un conjunto de variables.

Causal: Determinar las causas que provocan que una variable cambie cuando afectamos otra. Usualmente la observamos en medicina, biología y ciencia básica.

Mecanicista: Dar una descripción completa de algún fenómeno u objeto. Este es el tipo de pregunta que requiera la mayor cantidad de datos y cuya complejidad es inmensa. Usualmente la observamos en ingeniería o ciencia básica.

Aunque usualmente no abordarás con tanto detalle los ejemplos y el caso de uso en esta unidad, es de mucha utilidad determinar al menos.



Computación II

Unidad 1. Introducción a la minería de datos

En esta sección repasamos los conceptos básicos que caracterizan a las posibles preguntas que se pueden hacer a un conjunto de datos. El plantear el tipo de información que deseamos obtener en buena medida determina los subsecuentes análisis que deseamos aplicar.

Análisis exploratorio de datos

Considerando los conceptos y definiciones redactadas en el tema anterior es que vamos a revisar el tema del análisis exploratorio de datos. Sobre todo, porqué hicimos hincapié en la distinción entre la tarea estándar que define a la minería de datos, esto es, la de ir construyendo una hipótesis a partir de los datos y no con una suposición a priori.

Si queremos construir un modelo sobre predictivo sobre las preferencias de consumo de algún producto en una población más nos vale tener en cuenta aspectos como la edad promedio, desviación estándar, y muchos otros descriptores estadísticos en los que puedas pensar que te ofrezcan un resumen genérico de los datos.

Pero no es sólo el uso de estos resúmenes estadísticos de los que hay que echar mano, sino también de resumirlos en herramientas gráficas. Para tal efecto existen gráficas que pueden resumir muy bien esta información.

Por ejemplo, lo primero que podríamos hacer es hacer un conjunto de histogramas que describan los datos, pero el determinar el tamaño de los percentiles puede implicar un sesgo en la conformación de estos histogramas. Para esto podemos hacer un *boxplot* (gráfica de cajas) que es una gráfica que determina la media y desviación estándar de cada atributo.

Otros *plots* interesantes pueden ser una gráfica rango-frecuencia, este tipo de gráficas implican ordenar nuestros datos por rango y observar cómo es que la frecuencia con la que son representados en los datos va cambiando de mayor a menor. El principio detrás de este tipo de gráficos es la regla 80/20 de Wilfrido Pareto.

Existen muchos otros análisis exploratorios que se pueden hacer, por ejemplo, una gráfica de telaraña. Lo importante de este análisis es que al hacer varios de estos puedas tener una idea más clara de que es la información subyacente en los datos, para así determinar que posible modelo o algoritmo vamos a usar para describirlos.



Computación II

Unidad 1. Introducción a la minería de datos

Revisa los siguientes recursos:

Meyer, P. (13 de Agosto de 2015). *Exploratory Data Analysis*. [Video]. YouTube. <https://www.youtube.com/watch?v=zHcQPKP6NpM>

National Institute of Standards and Technology. (s.f.). *Engineering Statistics Handbook. What is EDA?* <http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm>

Cierre de la unidad

En esta unidad revisamos los principios básicos del análisis de los datos, que van desde los conceptos para establecer un vocabulario común hasta una fase denominada “Análisis exploratorio” que está pensada para que algunas estadísticas sencillas de obtener vayan dibujando las hipótesis que deseemos comprobar posteriormente con algoritmos más complejos, en vez de intentar hacer modelos o regresiones numéricas innecesarias.

Fuentes de consulta

Básica:

MacKay, D. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Whitten, I. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.