Análisis de Grandes Volúmenes de Datos
Carlos Espinoza Peraza - B92786
Reporte 22 de Abril

Resumen de los temas estudiados

Big Data

- Desde hace tiempo se han venido haciendo análisis de datos en campos como la matemática, informática, estadística, ingeniería y otras ciencias.
- Gracias al avance de la tecnología es que se han introducido nuevos factores que llevaron al despegue del término big data. Por ejemplo, uso de sensores, abundancia de datos en línea, bajo costo de recursos y nuevas capacidades de procesamiento.

Recolección de datos

- El primer paso para realizar un análisis es obtener los datos.
- Primeramente se deben realizar preguntas sobre el qué queremos análizar, si contamos con los datos necesarios y en caso de no tenernos donde los podemos obtener.
- Probablemente sea una buena idea buscar datasets preexistentes.

Limpieza de Datos

Generalmente los datos no están listos para ser trabajando directamente. Es necesario hacer una evaluación previa para poder determinar si el dataset es apto para el uso. Estos datasets suelen estar:

- Llenos de ruido (errores o artefactos)
- Llenos de basura
- Incompletos
- Campos obsoletos o redundantes
- Outliers
- Formatos inadecuados para la minería

Es importante recalcar que existe `GIGO: Garbage In Garbag Out` el cual se refiere a que probablemente los algortimos o modelos a los que serán sometidos los datos van a dar un resultado, pero depende del buen trabajo realizado en esta etapa, que esa información sea verídica y acertiva para responder a nuestro objetivo inicial de análisis.

Para tratar con estos problemas existen una variedad de métodos estadísticos para analizar la viabilidad de los datos por analizar, además de que son utilizados como guía para la optimización del dataset.

Dentro de los métodos de limpieza están:

> **Identificación de columnas con valores únicos:** Se deben eliminar estas columnas o `features` que no aportan datos relevantes dentro del modelo

> **Considerar la utilidad de las columnas con pocos valores únicos:** Que una columna o `feature` contenga pocos valores únicos, no significa necesariamente que deba ser eliminada. Por eso es bueno realizar el cálculo de cuál es el procentaje de valores únicos en relación a la cantidad de filas. Con este dato, bajo ciertos límites, se pueden detectar rápidamente columnas deben ser revisadas con mayor detalle.

> **Identificar filas que contengan datos duplicados:** Las filas que tienen datos duplicados probablemente sean poco útiles y hasta peligrosos pues llevan a una mala evaluación del modelo.

> **Método de la desviación estándar (Outliers):** En el caso de tenerse un dataset con una distribución similar a la distribución gausseana, es posible realizar un análisis utilizando la desviación estándar para encontrar aquellos datos que se encuentran alejados de la mayoría de observaciones.

> **Método del rango intercuartílico (Outliers):** Este consiste en la diferencia de los percentiles 75 y 25, resultando en el 50% del medio de los datos, lo que también es conocido como `el cuerpo de los datos`.

> **Detección automática de outliers (Outliers):** Para esto se utiliza el aprendizaje de máquina, el cual, mediante *one-class classification* acomoda en un modelo los datos normalizados y con esto logra predecir la categoría de las nuevas entradas. En otras palabras, para que este método funcione se le debe cargar un conjunto de datos normales para que a la hora de recibir nuevas entradas, este sea capaz de separarlos en datos normales, outliers o anomalías.

Comentarios sobre la materia estudiada

Esta es la primera vez que se le da tanto énfasis al proceso en general de como empezar a hacer un análisis. Generalamente, todo este trabajo es realizado previamente y se obvia cuales fueron los métodos utilizados para el tratamiento de impurezas.

Con respecto a la forma en la que se impartió la materia, fue bastante bueno tener presentaciones que no estuvieran cargadas de texto, además de prácticas que nos ponen a trabajar a un nivel más detallado lo visto en las presentaciones. Con esta metodología queda mucho más claro tanto lo que se ve en la clase, como lo que se ejemplifica en los tutoriales.

Dudas sobre la materia estudiada

¿Cómo sabe uno hasta que punto es necesario seguir puliendo los datos?

¿Hay alguna técnica que sea tipo estándar o recomendada en la mayoría de casos?

Posible uso como profesional

Probablemente le encontraría usos en cualquier campo o empresa en el que vaya a trabajar. Como hemos visto en este y otros cursos, el análisis nos permite tomar decisiones informadas. Por lo tanto, podría utilizarse para dar explicaciones a algunos fenómenos vistos en la industria. También, como algunos ejemplos específicos, podría ser

que se realicen análisis sobre la forma en la que las personas realizan compras y a partir de esto generar informes sobre cuales son los productos más vendidos, clasificándolos mediante distintos parámetros (tiempo, cantidades, ubicación, entre otros).

Problemas que podría resolver con las técnicas estudiadas

Las técnicas utilizadas en los ejercicios nos permitirían obtener información más acertiva sobre cuestiones tan cotidianas como el rendimiento académico de los estudiantes de computación. Se podrían hacer descartez sobre que información puede no ser útil de analizar. Además, es como una capa más de calidad que nos permitirían tener una mayor confianza y seguridad de que se está atacando el problema de `GIGO` a la hora de aplicarse el análisis.