University of Puerto Rico
Mayagüez Campus

Department of Computer Science and Engineering

CIIC 5995/8995: Big Data Analytics
Final Exam

Name: _____ TOTAL: /100

Student ID: _____

Section: _____ **June 29, 2017**

Answer the following questions.

I) **(25 pts)** **Problem I: MapReduce**: Briefly explain the concept of MapReduce, and provide an example on what problems you can solve with it.

Map Reduce es un modelo de programación y una implementación asociada para procesar y generar grandes conjuntos de datos. Los usuarios especifican dos funciones que son: Map y Reduce. Los usuarios especifican a una función Map, toma un conjunto de datos en lo que los elementos se dividen en tuplas (pares key/value) y la función de Reduce toma la salida de un Map como entrada y combina los datos tuplas en un conjunto más pequeño de tuplas.
Un problema que se puede resolver con Map reduce es el de proceso de una tienda, por ejemplo se tiene información de ventas de los últimos cuatro años y se quiere verificar la cantidad de ventas que se han realizados en todos ese tiempo y que productos fueron los más vendidos. Sería recomendable usar Map reduce. Para Map se ordenaría por cada tipo de producto, fechas, precio. Luego al aplicarse Reduce sumando todos los puntos y llegar a un monto total por cada uno(producto, fecha y tiempo) luego verificar la cantidad de ventas y los productos de acuerdo a ese tiempo.

II) **(25 pts)Problem II: Spark RDD**: Briefly explain the concept of the Spark RDDs and contrast their performance with that of MapReduce.

La principal abstracción de Spark es Resilient Distributed Dataset (RDD). Un RDD está particionado entre los distintos workers de de Spark y permite almacenar datos en memoria y persisten según los requisitos, esto permite un aumento masivo en el trabajo de procesamiento por lotes, los RDDs logran la tolerancia de fallos . Los usuarios pueden almacenar de forma explicita un RDD en la memoria entre maquinas y reutilizarlo en varias operaciones paralelas similares a Map Reduce. Un RDD impulsa el cálculo de los datos como en MapReduce.

III) **(25 pts)Problem III: Supervised Learning**: Briefly explain the concept of Supervised Learning, and provide an example on what problems you can solve with it.

Aprendizaje supervisado necesita una data de entrenamiento. Consiste en hacer predicciones a futuro basadas en comportamientos o características que se han visto en los datos ya almacenados (el histórico de datos). El aprendizaje supervisado permite buscar patrones en datos históricos relacionando todos campos con un campo especial, llamado campo objetivo.

Por ejemplo, si se desea contar con información de productos vendidos (se tendría un conjunto de ejemplos sobre estos productos), los correos productos se etiquetarían como “calidad” o “no calidad” por parte de los usuarios. El proceso de predicción se inicia con un análisis de qué características o patrones tienen los productos ya marcados con ambas etiquetas. Se puede determinar, por ejemplo, que un producto de calidad es aquel que tiene una marca en especifica y ese sería uno de los patrones. Una vez determinado el patron (esta fase se llama “de aprendizaje”), los correos nuevos que nunca han sido marcados como no calidad se comparan con los patrones y se clasifican (se predice) como “calidad” o “no calidad” en función de sus características.

IV) **(25 pts)Problem IV: Unsupervised Learning**: Briefly explain the concept of Unsupervised Learning, and provide an example on what problems you can solve with it.

Aprendizaje no supervisado no necesita una data de entrenamiento. Aprendizaje no supervisado es un método de aprendizaje automático donde un modelo es ajustado a las observaciones. Además, el aprendizaje no supervisado trata los objetos de entrada como un conjunto de variables aleatorias, siendo construido un modelo de densidad para el conjunto de datos.El aprendizaje no supervisado puede ser usado en conjunto con la Inferencia bayesiana para producir probabilidades condicionales (es decir, aprendizaje supervisado) para cualquiera de las variables aleatorias dadas.

Por ejemplo si se necesita una aplicación de reconocimiento de caras podríamos pasar la fotografía como un mapa de bits pero esto seria muy costoso computacionalmente, pero sin embargo si pasáramos una serie de valores como anchura de ojos, anchura de boca, tamaño de frente, etc., esto nos podría clasificar la cara en función de sus parecidos, usando un aprendizaje no supervisado.