Topics in Data Science = Өгөгдлийн шинжлэх ухаан

Big Data – Их хэмжээний өгөгдөл

Монгол Улсын Үндэсний статистикийн хороо, Улаанбаатар хот, Монгол Улс, 2017 оны 5 сарын 8

Ryan Womack

Data Librarian, Rutgers University, https://ryanwomack.com

Хэсэг 1: Танилцуулга

Илтгэгчийн танилцуулга

• Data Librarian

• Эдийн засаг, номын сангийн шинжлэх ухаан, статистикийн магистрийн зэрэгтэй

- Эдгээр салбарын огтлолцол дээр ажилладаг
- Өгөгдөл хайх, ашиглахад нь хүмүүст тусалдаг
- Номын санд зориулан мэдээллийг удирддаг

• Ахисан түвшний эрдэм шинжилгээ, судалгаан дээр ажиллах

Rutgers-ийн талаарх танилцуулга

• Rutgers, The State University of New Jersey

• New Jersey, 8.8 сая хүн амтай, New York болон Philadelphia-ийн хооронд оршдог

• Rutgers 1766 онд байгуулагдсан, 250 гаруй жилийн түүхтэй

• Carnegie ангилал: Судалгаа – Судалгааны ажлыг маш өндөр түвшинд хийдэг, дээд түвшний ангилал (R1)

• 100 гаран судалгааны хөтөлбөртэй

• 68,000 гаруй оюутантай (120 гаруй улсын 7,500 олон улсын оюутантай)

• Дэлхийн топ 100 их сургуулийн нэг (Times Higher Education, Shanghai Ranking, CWUR болон бусад)

Rutgers-ийн талаарх танилцуулга

Зураг

IASSIST-ийн талаарх танилцуулга

• IASSIST – Нийгмийн шинжлэх ухааны мэдээллийн үйлчилгээ, технологийн олон улсын нийгэмлэг

• Мэдээллийн технологи, номын сан, өгөгдлийн үйлчилгээ, судалгаа болон дээд боловсрол, төрийн, ашгийн бус болон хувийн судалгааны салбарт ажиллаж буй 300 гаруй өгөгдлийн мэргэжилтнүүдийн нэгдэл

• IASSIST дараахь зорилгын хүрээнд ажилладаг

- Өгөгдлийн үйлчилгээг хүргэх төгс сүлжээг дэмжих сурталчлах
- Нийгмийн шинжлэх ухааны дэд бүтцийг сайжруулах

- Мэргэжлийн туршлага солилцох боломжийг бүрдүүлэх

• Жил бүрийн конференци нь харилцан туршлага судлах, харилцаа холбоо үүсгэх фоум болдог (АНУ, Канад, Европ, дараагийнх Азид?)

• IASSIST Ази дахь гишүүнчлэлээ нэмэгдүүлэхийг зорьж байгаа.

Танилцуулга

Сургалт нь:

• Хөгжилтэй байна байх гэж найдаж байна

• Их хэмжээний өгөгдлийн талаарх танилцуулга

• Гол технологиудыг авч үзнэ

• Экосистемд оролцогч зарим гол тоглогчидтой танилцуулна

• Их хэмжээний өгөгдөлтэй ажиллахад ямар байх бодит сэтгэгдэл, орчныг бий болгоно

Танилцуулга

Энэ сургалт ... биш байх болно:

• Их хэмжээний өгөгдлийн бүрэн хэмжээний заавар биш

• Програмчлалын гүнгийрүүлсэн сургалт биш

• “Шууд” их хэмжээний өгөгдлийн эксперт болгохгүй

• Онолын мэдлэгийг өгөхгүй

Танилцуулга

Сургалтаас дараахь зүйлсийг олж мэднэ гэж найдаж байна:

• Их хэмжээний өгөгдлийн пакетуудыг өргөн хүрээнд олж мэдэж авна

• Хэрэглүүр, аргуудын зарим жишээ

• Их хэмжээний өгөгдлийн хүч, боломж, хязгаарлалтуудыг ойлгох

• Их хэмжээний өгөгдлийн та хэр хол байна, аль чиглэлд явах шаардлагатайг мэдэхэд тань туслах болно

• Цаашид суралцах замыг тань тавьж өгнө

Танилцуулга

Сургалтаас дараахь зүйлсийг олж мэднэ гэж найдаж байна, 2 дугаар хэсэг:

• Hadoop, HDFS, MapReduce, MongoDB

• HBase, Pig, Pig Latin, Hive, Spark, Scala, Sqoop

• Oozie, ZooKeeper, Flume, Ambari, Hue

• Hortonworks, Cloudera, Tessera, RHadoop

• AWS, EMR, EC2, Azure

• DeltaRho, Trelliscope, Lasso, PCA, SVD

Тохируулга

• Script, өгөгдөл зэрэг сургалтын материалуудыг http://github.com/ryandata/bigdata линк дээр байгаа

• script файл нь энд дурдсан тодорхойлолтуудыг харуулах зүйлсийг багтаасан байгаа.

• Зарим алхмыг хийхэд таны өөрийн үйлдлийн систем дээр ажилладагтай ижил алхмууд хийгдэх болно.

Хэсэг 2: Их хэмжээний өгөгдөл

Их хэмжээний өгөгдөл гэж юу вэ?

Утга санаа...

Тодорхойлолт...

Практик хэрэглээ...

Экосистем...

Их хэмжээний өгөгдлийн талаарх төсөөлөл

Их хэмжээний өгөгдлийн бодит байдал

Их хэмжээний өгөгдлийн хэлбэр төрх

Их хэмжээний өгөгдөл бол...

Их хэмжээний өгөгдлийг ихэвчлэн 3 V-ээр тодорхойлдог

1. Velocity - давтамж

2. Variety – олон төрөлт байдал

3. Volume - хэмжээ

4. Ихэвчлэн 4 дэх V буюу үнэн бодит байдал (Veracity)-г нэмж оруулах нь байдаг

Заримдаа үнэ цэнэ (Value) 4 дэх V гэж үздэг

Тооцооллын тодорхойлолт: Их хэмжээний өгөгдөл нь тухайн нэг нөхцөлд тооцоолол хийхэд хэт их мэдээлэл юм.

Хувийн тодорхойлолт: Их хэмжээний өгөгдөл нь таны ажилладагаас илүү том аливаа өгөгдөл юм.

Их хэмжээний өгөгдлийн 4 V

Их хэмжээний өгөгдлийг ашиглан юу хийж болох вэ

Их хэмжээний өгөгдөлтэй ажиллах нэг арга зам нь илүү хүчирхэг техник хангамжтай байх явдал юм.

Зарим асуудлыг үүгээр хялбархан шийдэж болно.

Бүтэцчиглэгдсэн өгөгдөл, сайн томъёолсон загвар, харилцан үйлчлэл

Ж.нь, банкны гүйлгээний төвлөрсөн мэдээллийн сан, зарим асуудлын загварчлал

Өндөр гүйцэтгэлтэй тооцоолол, паралел тооцоолол, их хэмжээний өгөгдлийн сан

Өгөгдөл нь биш процессор нь асуудал болдог

Энэ нь өртөг өндөотэй гэхдээ шинэ зүйл биш

Их хэмжээний өгөгдлийн энэ төрөл энд бидний авч үзэх сэдэв биш.

Их хэмжээний өгөгдлийн хувьд шинэ зүйл юу байна

Интернэтийн үйл ажиллагаа нь сарнисан, бүтэцчилэгдээгүй их хэмжээний өгөгдлийг бий болгодог.

Log файл, хайлтын бүртгэл, хэрэглэгчийн үйл ажиллагаа.

Facebook, Yahoo, Google болон бусад компаниуд энэ ажлыг эхлүүлсэн.

"The Cloud буюу Үүл"

Өгөгдлийн эрдэмтэд болж байна

Энэ сургалт энэ төрлийн өгөгдөлтэй хэрхэн ажиллах арга техникт чиглэгдэх болно.

Хэсэг 3: Hadoop + MapReduce

Hadoop

“Cloudera-ийн ахлах архитектур Doug Cutting вэбийн үүсгэсэн өгөгдлийг хэрэгцээтэй зүйл болгох зорилгоор Apache Hadoop-ийг хөгжүүлж, түүнийг удирдах уламжлалт системийн чадавхийг нэмэгдүүлсэн. Hadoop бөөгнөрсөн өгөгдлийг удирдах арга зүйн талаар дурдсан Google-ийн хэвлүүлсэн нийтлэлээс анх санаа нь гарсан бөгөөд тэр цагаас хойш хэдэн зуун терабайт, бүр петабайт өгөгдлийг хадгалах, боловсруулах, анализ хийх стандарт болсоор ирсэн.

Apache Hadoop нь 100% нээлттэй эх үүсвэртэй (open source) бөгөөд өгөгдлийг хадгалах боловсруулах шинэ арга арга замын суурийг тавьсан. Өгөгдлийг хадгалах, боловсруулах үнэтэй техник хангамж, янз бүрийн системүүдийг ашиглахын оронд Hadoop өгөгдлийг хадгалах, боловсруулах хоёуланг нь хийх үнэ өртөг багатай үйлдвэрийн стандарт сервер бүхий их хэмжээний өгөгдлийг паралель (зэрэгцээ) байдлаар боловсруулах боломжийг олгодог бөгөөд хязгаарлалтгүйгээр хэмжээг нь нэмж болно. Hadoop-тай бол ямар өгөгдөл тийм их биш байх болно.”

http://www.cloudera.com/content/cloudera/en/about/hadoop-and-big-data.html [link no longer available]

Hadoop болон HDFS архитектур

• Hadoop нь компьютеруудыг их хэмжээний агуулахын нэгж байдлаар алдаа гарсан тохиолдолд тодорхойгүй нөхцөл байдалд төвлөрсөн удирдлагаар ажиллуулдаг орчин юм. Name story.

• Hadoop орчин нь хэрэглэгчид (харьцангуй) нээлттэй байдлаар кластерийн хэмжээнд ажиллах бусад боловсруулалтын болон аналитик даалгаврыг гүйцэтгэх боломжийг олгодог.

• HDFS (Hadoop Distributed File System) нь Hadoop clusters-ийн үндсэн файл систем болдог.

• Hadoop Architecture

• Бидний гаргаж авах Hadoop cluster-ийн тухайн удирдлагад олон зүйл ордог.

MapReduce

• MapReduce нь Hadoop cluster-ийн олон үйлдлийн загварчлал юм

• Үйлдлийн Map бүрэлдэхүүн хэсэг зангилаа бүрт хүрч, өөрийн даалгаврыг гүйцэтгэж үр дүнг гаргаж өгдөг

• Reduce-ийн бүрэлдэхүүн хэсэг нь бүх зангилаанаас тэдгээрийн үр дүн, нэгтгэлийг цуглуулдаг

• MapReduce функцууд нь ихэвчлэн Java програм шиг бичигддэг бөгөөд өгөгдлийн талаарх мэдлэг, ойлголтыг хоёуланг нь шаарддаг.

MapReduce-ийн жишээ

• MapReduce Illustration

• MapReduce WordCount Program

• MapReduce-ийн товч танилцуулга (эхлэх хамгийн тохиромжтой хэсэг)

• Hadoop ecosystem-ийн дэлгэрэнгүй танилцуулга

Hadoop дээр ажиллах

• Hadoop экосистемийн бүх элемент өөр өөрийн комадын мөрийн хэрэгслээр гарж ирдэг.

• Ж.нь, hadoop file system-д ямар нэг зүйлийг оруулахын тулд hadoop fs -put компаныг ашиглаж болно.

• Энэ сургалтанд үйлдвэрлэлийн системд шаардлагатай зарим командын мөрийг ажиллуулахгүйгээр хийж болох хэрэглүүрүүдийг ашигладаг.

• A toy Hadoop cluster, Yahoo Hadoop cluster

• Yahoo болон Facebook эдгээр технологийн эхэн үеийн томоохон хэрэглэгчид байсан.

• Google өөрийн гэсэн файл системтэй. Google BigQuery-г үзнэ үү.

Хэсэг 4: Pig болон Hive

Pig

Өмнө нь дурдаж байсанчлан Java, Python эсвэл бусад програмчлалын хэл дээр MapReduce программыг бичихэд өндөр түвшний мэргэжлийн ур чадвар шаардлагатай бөгөөд ихээхэн цаг хугацаа ордог.

• Pig нь энэ хүндрэлийн заримыг арилгахаар хийгдсэн.

• Pig-ийг Yahoo дээр хөгжүүлсэн!

• Pig-ийн програмчлалын хэл нь Pig Latin.

• Pig нь зарим талаараа нээлттэй бүтэц, syntax-тай (Hive-тай харьцуулахад).

• Pig энгийн кодны цаадах үйл ажиллагаа зэрэгцээ гүйцэтгэдэг.

Hue

• Pig дээр бидний авч үзсэн жишээний хувьд бид браузер дээр ажилладаг AWS EMR-ийн Hue administration interface-ийг ашиглах болно.

• Complete Works of Shakespeare дээр үг тоололтыг ажиллуулдаг

• Default login: hue, password: 1111

• Pig Editor-тай.

• Мөн Grunt гэж нэрлэх command line interface-тэй.

• Hue-ийн бүх хэрэглүүрийг багтаасан орчныг http://demo.gethue.com линкээр харж болно

Hive

• Hive програмын хэлийг Facebook дээр хөгжүүлсэн.

• Hive нь стандарт SQL ихээхэн төстэй. Hive Cheat Sheet дахь SQL-ийг үзнэ үү.

• Pig шиг энгийн кодны цаад талд Hadoop cluster-тай хамт ажиллаж нарийн төвөгтэй байдлыгшийдвэрлэж өгдөг.

• Гэхдээ log файлууд нь нарийн төвөгтэй зүйлсийг нь харуулдаг бөгөөд алдаа гарсан тохиолдолд шууд харж болно!

• The manuals at the Hive-ийн сайт дээрх гарын авлагууд нь програмчлалын хэлний хамгийн цогц эх үүсвэр болдог.

• Түүнчлэн, бид Beeswax editor-ийг ашиглахын тулд Hive-ийг хэрэглэдэг.

Spark

• Spark нь Hadoop-аас сүүлд гарсан бөгөөд өргөн хүрээнд ашиглагдаж байна. Spark-д Hadoop-ийн өгөгдлийн агуулахыг ашиглаж болно, гэхдээ өөрийн гэсэн бие даасан системтэй.

• SQL төрлийн query-г илүү функц төрлийн програмчлалын хэлтэй хослуулж ашиглаж болно.

• Java, Python болон Scala функцуудыг дэмждэг. Spark нь Scala програмчлалын хэл дээр бичигдсэн байдаг.

• Ерөнхийдөө Hadoop-ээс илүү хурдан.

• Spark demo нь командын мөрийн хэрэглүүрүүдийг ашиглахыг шаарддаг.

– Энэ нь Mac/Linux орчинд илүү хялбар хэдий ч хэрэв Windows дээр Cygwin-тэй бол адилхан үйлдлийг гүйцэтгэж болно.

• EdX-ийн Apache Spark дээр их хэмжээний өгөгдлийн танилцуулга.

Spark-ийн талаарх материалууд

• Spark on CDH

• Stanford Spark Class

• Spark on an EMR Cluster

• Шугаман регрессийн кодыг Spark дээр дээд түвшинд тооцоолох зааврын 234 дүгээр хуудсыг үзнэ үү

Хэсэг 5: Hadoop-ийн экосистемийн хэрэглүүрүүд

Administration

• Бид Hue-тэй танилцсан.

• Oozie бол Hadoop-ийн workflower scheduler юм. Түүнийг Hue-г ажиллуулж байхдаа харсан.

• Zookeeper нь хуваарилагдсан аппликэшнүүдийг зохицуулах, удирдахад туслах зохицуулалтын үйлчлигээ юм.

• Ganglia бол Hadoop clusters-ийн хяналтын систем.

• Ambari бол янз бүрийн хяналт, удирдлагын хэрэглүүр дэх вэбэд суурилсан интерфэйс юм. Бид түүнийг Hortonworks Sandbox on Azure дээр илүүү дэлгэрэнгүй авч үзэх болно.

Өгөгдлийн бусад үйл ажиллагааны хүрээ (framework)

• Cassandra бол баганан гүйцэтгэлд чиглэгдсэн их хэмжээний өгөгдлийн сан юм.

• HBase сарнисан өгөгдлийг хадгалах Google's BigTable дээр загварчлагдсан их хэмжээний өгөгдлийн агуулах юм.

• MongoDB бол их хэмжээний үйлдлийг гүйцэтгэх зориулалттай NoSQL өгөгдлийн сан юм. [Болгоомжтой хандана уу! Аюулгүй байдалтай холбоотой томоохон алдаа гарч байсан]

• Sqoop бол Hadoop болон уламжлалт хамаарлын өгөгдлийн сан хооронд өгөгдөл дамжуулах хэрэгсэл юм.

• Flume бол ажиллаж байгаа системээс Hadoop-ийн өгөгдлийн агуулах руу өгөгдлийн автоматаар шилжүүлэх (server logs зэрэг) хэрэгсэл юм.

• Их хэмжээтэй өгөгдлийн аливаа даалгаврын хувьд The presence of so many tools and techniques (Impala, Mahout) for almost any big data task is what has made “Hadoop” a go-to solution for big data needs.

ACID болон BASE-аас гадна

“Төгсгөлийн нийцлийн сервис нь уламжлалт ACID (Atomicity, Consistency, Isolation, Durability) баталгаажуулалтын талаас авч үзвэл ихэвчлэн BASE (Basically Available, Soft state, Eventual consistency) утгыг нийлүүлэгч гэсэн ангилалд багтдаг. Төгсгөлийн нийцэл нь хэрэв тухайн өгөгдөлд шинэчлэл хийгдээгүй бол уг өгөгдөлд хандах эцсийн бүхий л хандалт нь хамгийн сүүийн шиэнчилсэн утгыг гаргаж өгөхийг баталгаажуулдаг тооцооллын хуваарилалтад ашигладаг нийцлийн загвар юм. Төгсгөлийн нийцэл нь хуваарилалтын системд өргөн ашиглагддаг "

Үүлэн тооцооллын R, 210 дугаар хуудас

Хэсэг 6: Бусад үйлчилгээ үзүүлэгчид

Microsoft Azure, Google Cloud

Бусад cloud үйлчилгээ нь их хэмжээний өгөгдлийг дэмждэг.

• Microsoft Azure ихээхэн хүнд scripting-тэй хэдий ч Hortonworks-д үнэгий ашиглах эрхийг олгодог. Мөн Portal.azure.com-г үзнэ үү.

• Google Cloud-ийг Hadoop clustersaap хангахад ашиглаж болно.

• Эдгээр үйлчилгээ нь тусламжийн өгөгдөл багатай, илүү үсрэлтүүдийг хийдэг Amazon илүү шинэлэг, хүч багатай байдаг.

Cloudera

• Cloudera (http://www.cloudera.com)

• Hadoop-т суурилсан их хэмжээний өгөгдлийн шийдлийг санал болгогч. 2009 онд байгуулагдсан.

• CDH (Cloudera Distribution with Hadoop)-ийг хуваариладаг. Энэ нь хэд хэдэн төслийг дэс дараатай “тавиур”-т эргэлдүүлж, хуваарилж, дэмжлэг үзүүлдэг гэсэн үг юм. HBase болон бусад Apache-ийн төслүүдэд дэмжлэг үзүүлдэг.

• Cloudera Live-ийн "Try the Demo" дээр дарах юмуу эсвэл шууд дараахь линкээр орж болно http://demo.gethue.com

• Hue нь Hadoop-ийн удирдлагын интерфэсийн нэг юм.

Hortonworks

• Hortonworks (http://www.hortonworks.com)

• 2011 онд Hadoop-ийн багийн инженерүүд 24 Yahoo! дээр үүсгэн байгуулсан. Cloudera-тай адилаар Hortonworks Hadoop аппликэшний “тавиур”-ыг удирдаж дэмжлэг үзүүлдэг. Зарим талаараа linux хуваарилалттай ижил. Hortonworks нь Hive болон бусад Apache төслүүдийг тэргүүлэгч нийлүүлэгч.

• Hortonworks Virtual Machines-ийн Sandbox mode-д болон Azure cloud-д байдаг (нэг сарын үнэгүй триал хувилбартай).

Ambari revisited

[Хэрэв та өөрөө хийж үзэхийг хүсвэл]

Hortonworks-ийг ашиглан Virtual Machine-аас татаж авах эсвэл Azure cloud trial хувилбараар аль алинаар нь Ambari-г ажилуулж болно.

Эхлүүлээд дараа нь browser interface-д хандана.

• 127.0.0.1:8888 нь VM дээрх (Azure дээр бол өгөгдсөн URL-ийг ашиглан) эхлүүлэх interface-ийг гаргаж өгдөг

• 127.0.0.1:8000 нь interface-ийг гаргаж өгдөг

• 127.0.0.1:8080 нь Ambari interface-ийг гаргаж өгдөг (login: admin, password: admin)

Хэсэг 7: R болон их хэмжээний өгөгдөл

R болон их хэмжээний өгөгдөл

• R нь open source бүхий статистикийн програмчлалын тэргүүлэгч платформ бөгөөд open source бүхий их хэмжээний өгөгдлийн програм хангамжийн иж бүрдэл байдлаар ашиглагддаг.

• Зарим R-ийн төслүүд болон өргөтгөлүүд том файлд хандах, паралел тооцоолол хийх болон HPC-ийн бусад асуудлыг шийдвэрлэж өгдөг.

• Эдгээрийн олонх байдаг хэдий ч (Task Views дээр жагсаасан) энд авч үзэхгүй.

• Бид кластер дахь Hadoop загварын их хэмжээний өгөгдөлтэй зохицдог хоёр төслийн талаар авч үзэх болно:

– RHadoop

– DeltaRho

RHadoop

• RHadoop нь Revolution Analytics-ийн хөгжүүлсэн таван пакетын багц юм:

– ravro, rhbase, rhdfs нь холбогдох форматаар өгөгдлийг бичих, унших пакетууд юм.

– rmr нь R-аар үйл ажиллагааны хүрээг зураглан багасгах интерфэйс болдог

– plymr нь rmr, “plyr meets MapReduce”-ийн програмын шаардлагыг багасгах дээд түвшний функцуудын цогц болдог

• Бусад пакетад SparkR, RHive, RCassandra болон бусад пакетууд ордог

• Үүлэн тооцооллын R дээрх 203 дугаар хуудсыг үзнэ үү

Cloud дээрх R

• cloud дээр R-ийг ажиллуулах нэг арга зам нь Amazon Machine Image (AMI) юм.

• Эдгээр урьдчилан бэлтгэн инсталлууд нь R-ийг ажиллуулах, AWS-ийг хэдхэн секундын дотор хялбархан ажиллуулахад тусална.

• Тухайлсан хэрэгтэй хувилбаруудын хувьд Louis Aslett's RStudio Server хувилбаруудыг үзнэ үү.

• Бусад тохируулгын заавар нь байдаг

• Сүүлийн үр дүнг Google-ээс харж болно

Trelliscope (DeltaRho, Tessera байсан)

DeltaRho-г Purdue, Pacific Northwest National Laboratory болон Mozilla хөгжүүлсэн. 2014 оны 11 дүгээр сард гарсан уг төсөл олон гэрээ байгуулж чадсан.

• R-ийн орчинд ажиллуулахад DeltaRho кластерийн дагуу ажиллах өөрийн командтай бөгөөд энэ орчинд анализ хийх ачааллыг бууруулж өгдөг.

• datadr пакет нь Hadoop-ийн хялбарчилсан интерфэйс болдог MapReduce-тэй төстэй байдлаар “хувааж, эргүүлэн нэгтгэдэг”.

• RHIPE нь Hadoop-тэй шууд харилцаж ажилладаг пакет юм.

• DeltaRho олон тооны хувьсагч, ажиглалтыг харуулж чадах Trelliscope гэсэн өөрийн визуалчлалын интерфэстэй.

• Эхлүүлэхийн тулд quickstart-ийг ажиллуулна

• Live demo нь байгаа.

Хэсэг 8: Өндөр хэмжээст болон сарнисан өгөгдөл

Өндөр хэмжээст өгөгдөл

• Олонх өгөгдлийн шинжилгээ өндөр хэмжээтэй холбоотой байдаг.

• Хувьсагчдын тоо нь ажиглалтын тооноос их байвал түүнийг өндөр хэмжээст өгөгдөл гэдэг (p>n).

• Тодорхой асуудал - trying to determine which of 10,000 генийн SNP (single nucleotide polymorphisms)-ийг тодорхойлох нь ховор тохиолддог хавдрын 100 тохиолдлыг тайлбарлахад тусалдаг.

• Principal Component Analysis (PCA) нь вариацийг тайлбарлахын тулд өгөгдлийн сан дахь хамгийн гол элементүүдийг сонгон авах сонгодог, дээр үеийн арга бөгөөд одоо өргөн хэрэглэгдсээр байна [R-д prcomp, SAS PROC FACTOR].

Өндөр хэмжээст өгөгдөл, үргэлжлэл

• Lasso бол хувьсагчийн тоог бууруулах, хувьсагчийг сонгон авах үйлдлийг зэрэг гүйцэтгэдэг өөр нэг нийтлэг арга юм.

Математикийн тэгшитгэл

• Статистикийн шинжлэх ухааны олон улсын нэвтэрхий толь “Торгуулийн абсолют хэмжээг тооцох нь” (“Absolute Penalty Estimation”, International Encyclopedia of Statistical Science)-г үзнэ үү.

• Бид (зохиомлоор) олон тооны хувьсагчийг ашиглан торгуулийн хэмжээг тодорхойлдог, тиймээс загварыг үр дүнд чухал нөлөө бүхий хувьсагчдыг л багтаасан байхаар багасгах шаардлагатай болдог.

• Нарийн тооцооллын боломж нь Lasso-г ихээхэн алдартай болгожээ [R-д lars, SAS GLMSELECT].

• Олон, олон бусад хувилбарууд гарсан гарч байсан, ж.нь. ℓ 1 / ℓ 2 торгуулийн журам. Орчин үеийн олон хувьсагчийн статистик техникүүд болон Өндөр хэмжээст өгөгдлийн статистик (Modern Multivariate Statistical Techniques and Statistics for High-Dimensional Data)-ийг үзнэ үү.

Сарнисан өгөгдөл

“Тоон шинжилгээнд санрисан матриц гэдэг нь ихэнх элемент нь тэг байдаг матриц юм. Тодруулбал, хэрэв ихэнх элемент нь тэг биш бол тухайн матрицийг нягтарсан гэж үздэг. Матриц дахь тэг элементийн тоог нийт элементийн тоонд харьцуулсан харьцааг сарнилт (нягтрал) гэж нэрлэдэг” [Wikipedia]

• Жишээ нь, Netflix-ийн киноны үнэлгээний өгөгдөл. Бүх киноны багахан хувийг аливаа нэг хэрэглэгч үнэлсэн байдаг.

• Матрицийн тэгүүдийг бөөгнөрүүлэх, тэднийг багцалсан хэлбэрт оруулах [R-д Matrix болон sparseMatrix пакет].

• Singular value decomposition (SVD) болон бусад математикийн/тооцооллын техник нь асуудлыг шийдвэрлэхэд ашиглагддаг [R-д svd эсвэл irlba in R].

• Судалгааны идэвхтэй чиглэл.

• Хэрэв жишээ авч үзэхийг хүсвэл AWS дээрх сарисан матрицийн цуглуулаг байгаа.

Хэсэг 9: Практик дахь их хэмжээний өгөгдөл

Практик дахь их хэмжээний өгөгдөл

Бодит их хэмжээний өгөгдлийн санд хандах нэг жишээг ажилуулж үзье.

• 500 TB гаран вэц хандалтын өгөгдөл Common Crawl project дээр байгаа, мөн Amazon S3 дээр байгаа.

• Энэ нь индекс ашиглах түргэн арга зам юм. The Support Library илүү идэвхтэй ашиглалтын боломжийг олгодог.

• Зарим нэг сурах бичиг энд байгаа.

• Энэ линкийн хувьд boto хэрэгтэй болно (pip-ийг ажилуулж boto-г суулгана).

Практик дахь их хэмжээний өгөгдөл, үргэлжлэл

• AWS заавартай, бас браузер хайлтын интерфэйстэй 1000 Genomes project зэрэг нийтийн бусад өгөгдлийн санг агуулдаг.

• Кластер үүсгэх нь өгөгдөлд анализ хийх MapReduce төрлийн функцийг ашиглах боломжийг олгоно.

• R-ийг ашиглан EC2 дээр дэлхийн цаг уурын мэдээлэлд дүн шинжилгээ хийх AWS-ийн өөр нэг жишээ.

Хэсэг 10: Дүгнэлт

Дүгнэлт

• Өнөөдрийн жишээнүүд их хэмжээний өгөгдлийн тооцооллын талаарх төсөөлөл, ойлголт өгөх зорилготой.

• Эдгээр технологийг үр дүнтэй ашиглахад оролдлого, сургалт шаардлагатай.

• Томоохон өгөгдлийг хадгалах ажлыг хэлүүлэхэд цаг хугацаа, мэргэжлийн туслалцаа хэрэгтэй.

• Хадгаладсан өгөгдөлд анализ хийхэд мэргэжлийн туслалцаа, цаг хугацаа хэрэгтэй.

• Хүсэл эрмэлзэл бүхий багийн хүчин чармайлт хэрэгтэй.

• Бодож үзэх зүйл: Өгөгдлийн шинжлэх ухааны 50 жилийн түүхтэй

Төгсгөл

• Асуулт байна уу?

• Маш их баярлалаа!

• Намайг дагаарай...

– https://youtube.com/librarianwomack

– https://www.linkedin.com/in/ryanwomack

– https://twitter.com/ryandata

– https://ryandata.wordpress.com

Ном зүй

1. Peter Buhlmann and Sara van de Geer. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, 2011.

2. Michael Frampton. Big Data Made Easy: A Working Guide to the Complete Hadoop Toolset. Apress, 2015.

3. Thilina Gunarathne. Hadoop v2 MapReduce Cookbook. Second Edition. Packt, 2015.

4. Richard Hill, Laurie Hirsch, Peter Lake, and Siavash Moshiri. Guide to Cloud Computing: Principles and Practice. Computer Communications and Networks. Springer, 2013.

5. Alan J. Izenman. Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer, 2008.

6. A. Ohri. R for Cloud Computing: an Approach for Data Scientists. Computer Communications and Networks. Springer, 2014.

7. K. G. Srinivasa and Anil Kumar Muppalla. Guide to High Performance Distributed Computing: Case Studies with Hadoop, Scalding, and Spark. Computer Communications and Networks. Springer, 2015.