

Ellenőrző kérdések a BigData elemzési módszerek zárthelyihez - 2016

1 Adatelemzési és statisztikai alapok

- Milyen típusú változótípusokat különböztetünk meg? Hol van ezeknek szerepe? Milyen típusú változók fordulhatnak elő egy olyan adatsorban, amely egy magyarországi lakosok vásárlási szokásait felmérő, alábbi pontokat tartalmazó kérdőívből született:
 - nyilatkozó neme, életkora, lakóhelye, legmagasabb iskolai végzettsége;
 - vásárlási gyakorisága, hetente hányszor vásárol X terméket;
 - a standard vagy a prémium alterméket szereti?
- Mi a strukturált/nemstrukturált/szemistukturált adat? Mondjon példát mindhárom típusra!
- Mi a felderítő és mi a megerősítő statisztikai elemzés? Mondjon példát mindkét megközelítésre!

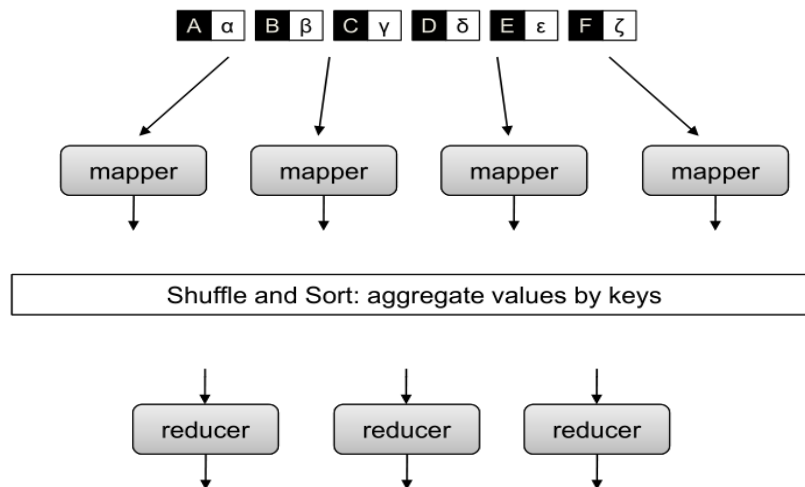
2 Vizuális analízis

- Mik a fő különbségek az EDA és a CDA között az adatelemzés során?
- Mi a dobozdiagram? Minek a szemléltetésére használjuk? Ábrán szemléltesse, hogy a dobozdiagram hogyan reprezentálja egy megfigyelés-halmaz alapvető leíró statisztikáit!
- Mi a dobozdiagram mediánjának, „bajszainak“ és „sarokpontjainak“ (*whiskers and hinges*) kapcsolata a normális eloszlás paramétereivel? Diskutálja, hogy alkalmas-e a dobozdiagram más eloszlások szemléltetésére is, és ha igen, milyen korlátokkal!
- Mi a SPLOM? Miért használjuk a vizuális EDA során? Mik alkalmazásának legfőbb korlátai?
- Mozaik-diagram: szöveges definíció, szemléltetés ábrával, jellemző alkalmazási esetei és alkalmazásának korlátai.
- Párhuzamos koordináták: szöveges definíció, szemléltetés ábrával, jellemző alkalmazási esetei és alkalmazásának korlátai.
- Mit jelent, ha egy párhuzamos koordináta diagram két szomszédos tengelye között futó szakaszokra illesztett egyenesek egy pontban metszik egymást?
- Mit jelent, ha egy párhuzamos koordináta diagram két szomszédos tengelye között futó szakaszokra illesztett egyenesek két pontban metszik egymást? Milyen hipotézist állítana fel ebből a megfigyelésből?

3 Nagy méretű adatok vizualizációja

- Mik a disztributív, algebrai, holisztikus típusú statisztikai aggregátorok? Hová tartozik a szórás, az IQR és a percentilis?
- Milyen típusú vizualizációkat alkalmazunk szokásosan a Big Data vizualizáció során?
- Ismertesse és indokolja, hogy a vizuális analízis klasszikus diagram-típusai közül melyek alkalmazása válik nehézkessé, illetve értelmetlenné „Big Data” kontextusban!
- A Big Data vizualizáció során a megjelenítést és a diagram-leíró adatok kiszámítását jellemzően szétcsatoljuk és az előbbit is több lépésben végezzük. Vázolja fel és ismertesse a Big Data vizualizációs „csővezeték” (*pipeline*) főbb lépéseit!

4 A MapReduce algoritmus-szervezési minta



- Hogyan érjük el a MapReduce séma alkalmazásánál az adat és kód kolokációját?
- Mi a “shuffle and sort” fázis feladata a MapReduce végrehajtás során?
- A kiterjesztett MapReduce sémában mi a “combiner” feladata? Miért érdemes alkalmazni?
- Tároljunk a HDFS-ben fix formátumú CSV állományokat, melyek n folytonos változó feletti megfigyeléseket írnak le egy időbélyeggel kiegészítve. Azaz az állomány sorai így néznek ki: $timestamp, x1_value, x2_value, \dots, xn_value \setminus n$ Adjon Mapper és Reducer pszeudokódot az egyes megfigyelt változók időbeli maximum-helyének meghatározására!

5 Adatfolyam-feldolgozás

- Ismertesse az adatfolyam-feldolgozás elemi blokkjának tekintett “stream processor” mintát! Hogyan történik ezekkel a bejövő adatfolyamok feldolgozása?
- Milyen problémák merülhetnek fel adatfolyamok mintavételezésénél? Kulcs és érték mezőkre osztható feldolgozandó n -esek esetén hogyan valósítaná meg a kulcstér feletti mintavételezést? (Azaz a kulcsok halmazán mintavételezünk – pl. felhasználók, ha $(user, search_query)$ alakú megfigyeléseink vannak - és minden a mintába eső kulcshoz tartozó n -est továbbbenedünk.)
- Mik a Bloom filterek? Hogyan alkalmazzuk őket halmazba tartozás közelítő ellenőrzésére adatfolyam-feldolgozásban?

6 Spark

- Ismertesse a Spark számításszervezési-modelljének alapelemét, a *Resilient Distributed Dataset*-et (RDD)!
- Mit jelent, az hogy a Spark támogatja az ezekből szervezett számításleíró gráf lusta kiértékelését (*lazy evaluation*)?
- A Spark alapvetően két fajta *operációt* támogat, a *transzformációkat* és az *akciókat*. Definiálja az előbbi kategóriát és adjon három példát!
- A Spark alapvetően két fajta *operációt* támogat, a *transzformációkat* és az *akciókat*. Definiálja a második kategóriát és adjon három példát!

7 BD ML

- Röviden ismertesse a gépi tanulás (machine learning) alapfeladatit!
- k-means klaszterezés: alapötlet, definíció, (szekvenciális) pszeudokód
- Bináris osztályozási döntések jóságának mérése: a konfúziós mátrix (igazságmátrix, eset-kontroll tábla; *confusion matrix*)
- Lineáris regresszió: alapötlet, definíció
- Lineáris regressziós feladat átírása „summation form”-ba. Értékelje a summation form kiszámításának párhuzamosíthatóságát!