Ellenőrző kérdések a BigData elemzési módszerek zárthelyihez

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/szemistrukturált adat? Hogyan hidalná át a nemstrukturált adatok feldolgozásának problémáját?
- Mi a populáció? Mi a minta, mikor tekintünk egy mintát reprezentatívnak?
- Mi a felderítő és mi a megerősítő statisztikai elemzés? Mondjon példát mindkét megközelítésre!
- Mi a null- és mi az alternatív hipotézis? Tudjuk, hogy egy standard Tic Tacos doboz drazséinak száma normális eloszlást követ, melynek várható értéke 36; mi azt sejtjük azonban, hogy az egyetemi büfében kapható dobozban kevesebb van, ezt szeretnénk bizonyítani a fogyasztóvédelmi hivatalnak. Vásárolunk néhány dobozzal és kiszámoljuk a drazsék számának átlagát, majd lefuttatjuk a statisztikai tesztünket. Mi lehet itt a null- és az alternatív hipotézis, hogy a fogyasztóvédelemhez írott érvelésünkben hitelesek legyünk?
- Mi a megfigyelési tanulmány és mi az irányított kísérlet?
- Legyen adott az alábbi újsághír: „Brit tudósok 100, harmincéves nő dohányzási szokásait vizsgálták. 50 nő az elmúlt 10 évben napi 1 doboznyit fogyasztott el, míg a másik 50 egyetlen szálat sem. Mind a 100 nőnek megmérték a tüdőkapacitását és úgy találták, hogy van korreláció a tüdőkapacitás és a dohányzási szokások között.“. A brit tudósok irányított kísérletet végeztek, vagy egy megfigyelési tanulmánynak voltak tanúi?

2 Vizuális analízis

- Mik a fő különbségek az EDA és a CDA között a statisztikai elemzés 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? Diszkutálja, 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 párhozamos koordináta diagram két szomszédos tengelye között futó szakaszokra illesztett egyenesek egy pontban metszik egymást?

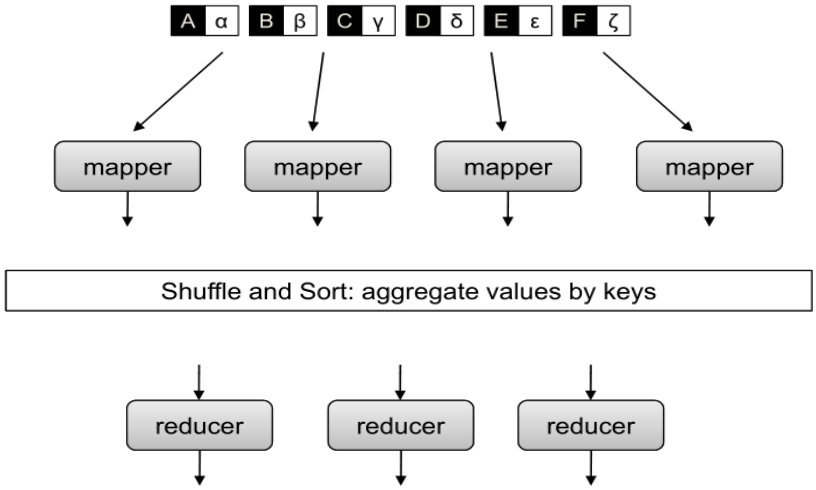
- Mit jelent, ha egy párhozamos koordináta diagram két szomszédos tengelye között futó szakaszokra illesztett egyenesek két pontban metszik egymást? Milyen hipotézist állítana fel ebből a megfigyelésből?

3 Nagy méretű adatok vizualizációja

- Mik a disztributív, algebrai, holisztikus típusú statisztikai aggregátorok? Hová tartozik a szórás, az IQR és a percentilis?

4 A MapReduce algoritmusszervezési minta

- Mi a horizontális és mi a vertikális skálázási megközelítés? Az, hogy a BigData házi feladatok kétfős csapatokban oldhatóak meg, az melyik típusú erőforrás-bővítési mechanizmust követi?
- Legyen adott a következő input: 5 p dimenziós A, B, C, D, E középpontról szeretnénk eldönteni, hogy az általuk definiált Voronoi cellákban a szintén bemenetként érkező n darab p dimenziós adatpontból hányat tartalmaznak. Adjon MapReduce stílusú megoldást a problémára! A megoldási mód tetszőleges lehet, írjon például pszeudokódot, vagy töltse ki az alábbi ábrán a



mapper doboz kimenetét és a reducer ki- és bemenetét! Ügyeljen arra, hogy megoldása kellően konkrét legyen, a megoldási elgondolást tartalmazó szöveges megoldást nem fogadjuk el.

- 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?
- PageRank megvalósítása MapReduce-szal
- Tároljunk a HDFS-ben fix formátumú CSV állományokat, melyek n folytonos változó feletti megfigyeléseket írnak le egy időbélyeggel kiegészítve. Adjon Mapper és Reducer pszeudokódot az egyes megfigyelt változók időbeli maximum-helyének meghatározására!
- k-means klaszterezés megvalósítása MapReduce segítségével: algoritmusszervezés szöveges ismertetése, map és reduce pszeudokódok
- Tároljunk HDFS-ben egy folytonos megfigyelt változó feletti, időbélyeggel ellátott megfigyeléseket (pl. "timestamp, value" szerkezetű CSV). Hogyan állítaná elő a megfigyelések hisztogramját MapReduce algoritmusszervezéssel? (Pszeudokódot is kérünk.)

- Tároljunk HDFS-ben két folytonos megfigyelt változó feletti, időbélyeggel ellátott megfigyeléseket (pl. “timestamp, var1, var2” szerkezetű CSV). Hogyan állítaná elő a megfigyelések hőtérképét (*heatmap*) MapReduce algoritmusszervezéssel? (Pszeudokódot is kérünk.)

5 Mintavételezés és anomáliadetektálás

- Mutassa be a 3 tanult mintavételezési technikát és illusztrálja példával ezek működését pl. egy közvélemény-kutató cég esetén, ahol a bemeneti populáció Magyarország teljes lakossága!
- Mit nevezünk kollektív anomáliának? Mi a viselkedési és kontextus anomáliák közötti különbség?
- Használható-e a boxplot EDA során anomáliadetektálásra? Miért?
- Mik a korlátai az euklideszi távolságnak? Mondjon példát más távolságfüggvényekre és azok lehetséges alkalmazási területeire!

6 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 kulcstér feletti mintavételezést? (Azaz a kulcsok halmazán mintavételezünk és minden a mintába eső kulcshoz tartozó n-est továbbengedünk.)
- Mik a Bloom filterek? Hogyan alkalmazzuk őket halmazba tartozás közelítő ellenőrzésére adatfolyam-feldolgozásban?