Ellenőrző kérdések a BigData elemzési módszerek zárthelyihez 2015

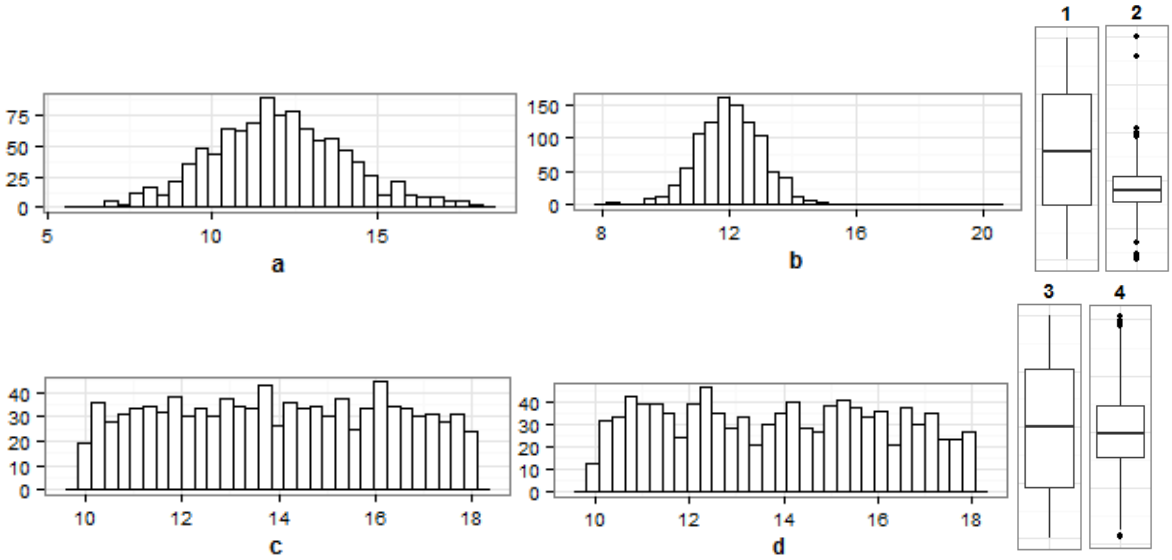
1 Adatelemzési és statisztikai alapok

- 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? 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 a statisztikai elemzés során?
- Mi a dobozdiagram (*boxplot*)? 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?
- Kiszámítjuk egy folytonos változó értékeit tartalmazó adatsor mediánját, móduszát és átlagát. Válassza ki az igaz állításokat!
    - A medián biztosan kisebb az átlagnál.
    - Az átlag legfeljebb kétszerese lehet a mediánnak.
    - A medián és a módusz az adatsor egy-egy kitüntetett értékét jelölik.
    - A módusz megegyezhet a mediánnal.
    - Találunk olyan reguláris kategorikus változót, amelynek mediánja megegyezik az általunk kiszámolt mediánnal.
    - Találunk olyan reguláris kategorikus változót, amelynek módusza megegyezik az általunk kiszámolt mediánnal.

- Egy folytonos változó jellemző értékeit doboz diagrammal (boxplottal) és hisztogrammal is ábrázoljuk. Válassza ki az igaz állításokat!
    - A doboz diagramról mindig könnyedén leolvasható az első kvartilis.
    - A hisztogramról mindig könnyedén leolvasható az első kvartilis.
    - A doboz diagramról mindig könnyedén leolvasható a 40. percentilis
    - A hisztogramról mindig könnyedén leolvasható a 40. percentilis
    - A doboz diagramról mindig könnyedén leolvasható a módusz.
    - A hisztogramról mindig könnyedén leolvasható a módusz.
    - Minden információ ami a doboz diagramról könnyen leolvasható a hisztogramról is könnyen leolvasható, emiatt tekintjük a doboz diagramot a hisztogram egyfajta absztrakciójának.
- Egy adatsor *a*, *b*, *c* és *d* változóját ábrázoltuk hisztogramon és boxploton is, de sajnos a boxplotok címkéi elvesztek, így nem tudjuk, mely ábrák tartoznak ugyanazokhoz a változókhoz. Válassza ki az igaz állításokat!



    - Az 1-es boxplot biztosan a *c* hisztogramhoz tartozik
    - A 2-es boxplot biztosan a *b* hisztogramhoz tartozik
    - A 3-as boxplot biztosan az *a* hisztogramhoz tartozik.
    - A 4-es boxplot biztosan a *d* hisztogramhoz tartozik.

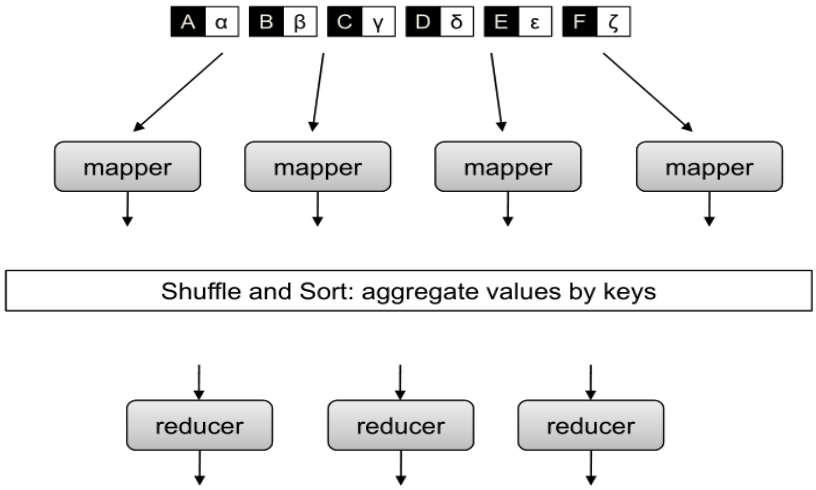
3 Nagyméretű adatok vizualizációja

- Mik a disztributív, algebrai, holisztikus típusú statisztikai aggregátorok? Hová tartozik a szórás, az IQR és a percentilis?
- Mi a *small multiplies* elv a vizualizációban? Miért különösen fontos ez a nagyméretű adatsorok vizualizációja témakörben?
- Mi az alapvető különbség a Map&Reduce és a Divide&Recombine minták között?
- Mondjon néhány alapvető megközelítést/tippet/trükköt nagyméretű adatsorok vizualizációjára!
- 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.)

4 A MapReduce algoritmusszervezési minta

- Mi a horizontális és mi a vertikális skálázási megközelítés? Az, hogy a BigData házi feladatok háromfős 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.

- 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. 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

5 Adatfolyam-feldolgozás

- Ismertesse az adatfolyam-feldolgozás elemi blokkjának tekintett “stream processor” mintát! Hogyan történik ezekkel a bejövő adatfolyamok feldolgozása? Mit értünk adatfolyam-feldolgozás esetén az alábbi fogalmakon: ismeretlen mintavételezési gyakoriság (unknown sampling frequency), korlátozott számítási/tárolási kapacitás (limited computational/storing resources), ad-hoc és állandó lekérdezések (ad-hoc and standing queries)?
- Milyen problémák merülhetnek fel adatfolyamok mintavételezésénél? Vázoljon egy lehetséges megoldást kulcs-érték párok kulcsok fölötti mintavételezésének problémájára?
- Mik a Bloom filterek? Térjen ki a bitvektor és a hash függvények szerepére a megközelítésben! Hogyan alkalmazzuk őket halmazba tartozás közelítő ellenőrzésére adatfolyam-feldolgozásban?

- Web crawlerünkben Bloom-filtereket alkalmazunk a már látogatott URL-ek felismerésére. Bloom filterünk két hash függvényt használ, a következő paraméterekkel működik:
  - N = 11
  - Input: egész számok (az URL-eket reprezentálandó)
  - h1(x): a páros bitekből képezett y mod N (tehát h1(585) = h1($\mathbf{1}0\mathbf{1}0\mathbf{0}1\mathbf{0}0\mathbf{1}\mathbf{1}_2$) = 01001_2 mod 11 = 9 mod 11 = 9)
  - h2(x): a páratlan bitekből képzett y mod N (tehát h2(585) = h2($\mathbf{1}0\mathbf{1}0\mathbf{0}1\mathbf{0}0\mathbf{1}\mathbf{1}_2$) = 10010_2 = 18 mod 11 = 7)

A rendszerünkben eddig a következő műveleteket hajtottuk végre: **Beszúr(25), Beszúr(159)**. Mit ad vissza a **KERES(118)** művelet? Interpretálja a végeredményt!

6 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 (outlierek) közötti különbség? Szemléltesse a különbséget példával!
- Milyen alapvető módszerei vannak az offline outlier detektáló algoritmusok adatfolyamokra történő közvetlen adaptálásának?
- Az ábrán szemléltetett kétdimenziós adatsoron lefuttattuk az alábbi outlier detektáló algoritmusokat: DB, LOF, féltér-mélység keresési és BACON. A végeredményt vizualizáltuk, minden esetben a nagyobb outlier score-ral rendelkező adatpontokat színeztük pirosra. Sajnos elfelejtettük, melyik ábra melyik algoritmushoz tartozik. Párosítsuk össze az algoritmust az ábrákkal!