Válasz opponensi véleményre – doktori disszertáció

Bíráló: Dr. Kovács Edith Alice
Cím: Gépi tanulással támogatott túléléselemzés
Szerző: Csalódi Róbert
Témavezetők: Dr. Abonyi János
Dr. Ruppert Tamás

Tisztelt Dr. Kovács Edith Alice!

Ezúton szeretném megköszönni konstruktív és rendkívül hasznos véleményét az értekezésem kapcsán. Alaposan áttekintettem az összes megjegyzést és kérdést, amikre részletesen válaszolok a levélben.

Köszönettel
Csalódi Róbert

1. A 3. fejezetben lévő prosztata-rákos példán, a kúrák hatását a klaszterekbe való sorolás elemzéséhez, a Jelölt, kibővítette az Shannon entrópia kiszámolásával (3.23 képletből hiányzik egy egy (-1) szorzó-tényező). Ennek alapján a magas dózis relatív kevésbé hat a klaszterekbe való tartozásra. Ugyanennél a résznél pozitívumként kiemelem, hogy a Jelölt felhívja a figyelmet, a körültekintő értelmezésre, esetleges háttérváltozók hatására.

Kérdés a Jelölthöz: Milyen érték körüli lenne az entrópia, ha egy tényező felvett értéke nem lenne hatással a klaszterbe való tartozásra, az 5 klaszter esetén?"?

Nagyon szépen köszönöm a kérdését. Ugye ebben az esetben az egyes klaszterekbe tartozó valószínűségek azonosak, melyek összege egyet kell, hogy kiadjon. Így valamennyi klaszterbe tartozás valószínűsége egységesen 0.2-nek felel meg. Tehát $H = -5 * 0.2 * \log_2(0.2) = 2.3219$ ebben az esetben az entrópia, ami maximális. Minél inkább magasabb az entrópia értéke annál bizonytalanabb a klaszterbe tartozás.

2. A 3. fejezetben lévő Covid-halálozás alkalmazás esetén, a Jelölt készített egy színezett ábrát (3.14 ábra). A feltüntetett számok alapján, amik a 100K emberre jutó halálozások szerepelnek. Észrevehető, hogy a klasztereken belül ezek egyáltalán nem mondhatók homogéneknek, pedig ezt várnánk. Lila klaszter (2-es) van 20 alatti és 200 feletti értékeket is tartalmaz, piros klaszter sem tekinthető homogénnek, mégha Kínát nem is tartjuk mérvadónak.

Kérdés a Jelölthöz: Mi lehet ennek a magyarázata?

Nagyon szépen köszönöm ezt a fontos észrevételt. Amikor először elkészítettem ezt az ábrát jómagam is hosszan dilemmáztam ezen kérdésen. Alapvetően a különböző klaszterekhez tartozó adatok valószínűségi eloszlását írom le Weibull modellel és vizuálisan az

eloszlásfüggvények alapján, valamint a paraméterek segítségével hasonlítom össze, hogy az adott klaszterbe tartozó országok esetében várhatóan hogyan alakul a százezer főre vetített halálozási ráta. Weibull eloszlásnál ezen paraméterek esetében a minták előfordulási valószínűsége egy adott klaszterben nagy terjedelmen magas, amit a 3.13-as ábrán lévő hisztogramok is szemléltetnek, és ez okozza ezt a látszólagos inhomogenitást. Ha csak önmagában a halálozási rátákat klasztereznénk könnyen lehet, hogy a minták kisebb szórással szeparálódnának és nem lenne a halálozási rátáknak ekkora terjedelme. Viszont a klaszterbe való tartozást azonos súllyal befolyásolják a magyarázó változók is. Így előfordulhat, hogy látszólag alacsony halálozási rátájú ország egy magasabb halálozási rátájú klaszterbe kerül mert a magyarázó változóinak karakterisztikája oda húzza el a besorolást. Így a módszer választ tud adni az olyan kérdésekre, hogy az olyan országokban, amelyeknél egy magyarázó változó értéke várhatóan magasabb, ott várhatóan hogyan alakul a halálozási ráta.

3. A 3. fejezet, 19. oldalán, a módszer egy megszorításaként, az is megemlítendő, hogy a figyelembe vett attribútumokról feltételezendő, hogy egymástól független föltéve a klasztert. Megjegyzem, hogy más eljárásokban, mint a naive Bayes is élnek ilyen megszorító feltételezéssel, és hatékony.

Kérdés a Jelölthöz: Mit gondol, hogyan lehetne azt tesztelni, hogy vajon a figyelembe vett változók elegendők ahhoz, hogy egy hatékony eljárást kapjunk?

Nagyon szépen köszönöm a kérdését. A dolgozatban bemutattam az Akaike Információs Kritériumot, ami ugye egy számszerű értéket ad az illesztés jósága és a változók száma alapján. Ahogy a bírálat összegzésében írta, az AIC kiváló a különböző identifikált esetek összehasonlítására. Így magát a kritériumot felhasználtam arra, hogy kiválasszam a rendelkezésre álló változók közül melyek bevonása és mekkora klaszterszám mellett eredményezi a legjobb AIC értéket. Ezzel a rendelkezésre álló lehetőségekből kihoztam az optimális beállítást, egy véleményem szerint hatékony és kézenfekvő megoldással. Például a COVID-os esettanulmány során kezdetben 20 magyarázó változóm volt, ami lecsökkent 9-re. Az optimális beállítást egyébként a különböző kombinációk manuális kipróbálásával értem el, aminek automatizálása mindenképpen tudna még a hatékonyságon növelni. Ezt tudná segíteni egy érzékenységvizsgálat, amivel lényegében a változók előszelekcióját végeznénk el.