Galaxisok vöröseltolódásának becslése színképeik alapján

Rudolf Ádám, ELTE TTK, Fizikus MSc

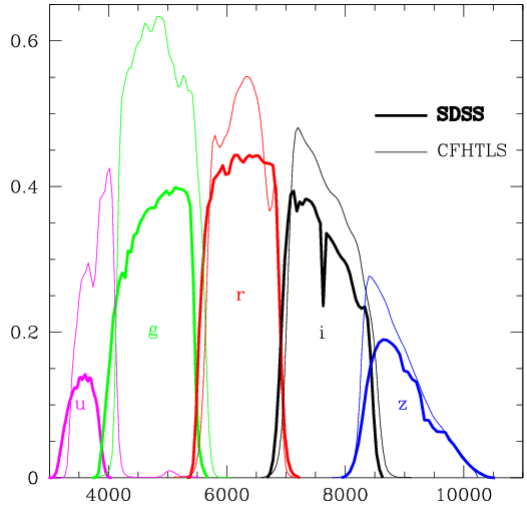
2012.12.1.

Bevezetés

Az adatbányászat általános problémája, hogy egy adathalmazban különböző mennyiségek között valamilyen függvénykapcsolatot fedezzünk fel, és ez alapján egy mennyiséget később jósolni tudjunk a meglévő adatok felhasználásával. A Kaggle internetes portálon ilyen adatsorok jóslására írnak ki versenyeket. A Számítógépes modellezés MSc laboron egyik feladatunk az egyik Kaggle versenyen való részvétel, és annak dokumentálása volt.

1. A probléma

A konkrét feladat galaxisok színképének vöröseltolódásának becslése volt a színkép adatai alapján. A színképeknek nem rögzítik a teljes spektrumát, csak bizonyos színszűrőkkel szűrt kép integrálját. Így minden színszűrőhöz, azaz színhez egy szám van megadva. Ez hasonló az ember színlátásához, ám több színt használnak, és nem pont ugyanazt a hármat. Így az elméletileg folytonos spektrumnak egy tömörített változatát kapjuk meg. A használt 5 szín: u (ultraibolya), g (green, azaz zöld), r (red, azaz piros), i (infravörös), valamint z (távoli infravörös).



1. ábra. Példa az ugriz szűrőcsomag áteresztési karakterisztikájára: a CFHTLS és az SDSS égfeltérképezéseknél hazsnált szűrőcsomagok összehasonlítása. A transzmisszió látható a hullámhossz függvényében.

A versenykiírásban három fájl van megadva: az egyik egy tanítóhalmaz: `train.csv`. Ebben a mérések sorszámain kívül megtalálható az öt szín értéke hibával együtt, valamint a hozzájuk tartozó vöröseltolódás. A `query.csv` fájlban újabb mérési adatok találhatók, csak a vöröseltolódás értéke hiányzik. A tanító halmaz alapján ezeket kell megbecsülni, és elküldeni a Kaggle portálon keresztül, ahol a program ismeri a valódi adatokat, és a következő értéket számolja ki:

$$\frac{1}{N}\sqrt{\sum_{i=1}^{N}\left(z_{e,i}-z_{r,i}\right)^2},$$

ahol N az adatpontok száma, z_e a becsült, z_r pedig a valódi érték a vöröseltolódásra. i az adatpontokat indexeli. A harmadik, `sampleSubmission.csv` fájlban egy ilyen beküldésre való példa látható. Ebbe az adatpont sorszámát, a vöröseltolódást és annak hibáját kell beírni. A letölthető és beküldendő állományok vesszővel elválasztott strukturált szöveg formátumban (`.csv`) vannak.

2. Az adatsor elemzése

R nyelven dolgoztam, mert kiválóan alkalmas adatbányászati problémák megoldására. Azon kívül, hogy könnyedén kezeli a `.csv` formátumot, beépített függvények vannak lineáris modellek illesztésére, döntési fák generálására és az eredmények vizsgálatára.

2.1. Magnitúdóskála

Az adatok magnitúdóskálán vannak megadva, ami az f abszolút fényességtől a következő módon függ:

$$m = -2,5 \log_{10}(f).$$

Az adatok hibája így nem gaussos, ami pl. lineáris regressziónál torzítást okozhat, de ettől most eltekintünk.

A fényességértékeket befolyásolhatja a kérdéses galaxis távolsága, ám mi magára a galaxis saját tulajdonságára vagyunk kíváncsiak, vagyis az abszolút fényességet valahogy ki kell transzformálni az adatokból. E célból a fényességértékek arányaival dolgozom, ami a magnitúdóskálán különbségként jelenik meg. Bármilyen párosítást választhatunk, de tipikusan az $u-g$, $g-r$, $r-i$, $i-z$ értékekkel szoktak dolgozni, a továbbiakban én is ezeket fogom használni.

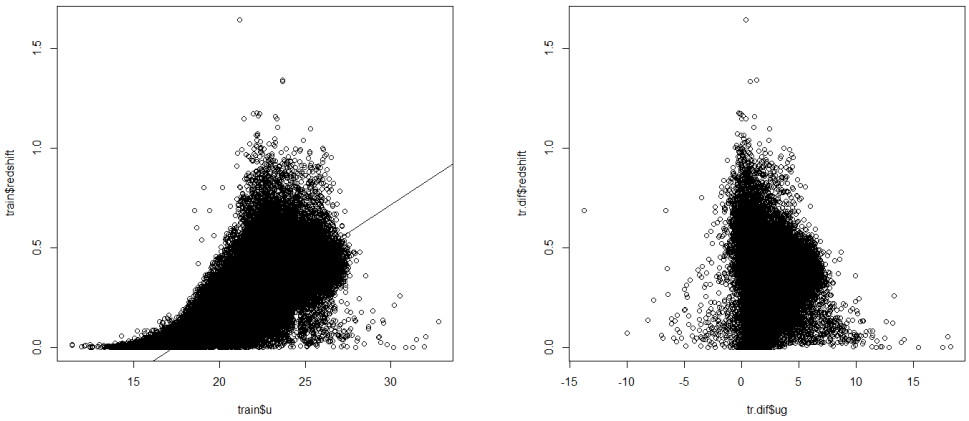
Az adatok hibájával nem foglalkoztam, hibaként a vöröseltolódás értékek gyökét adtam meg.

2.2. Vizuális vizsgálat, fontos változók

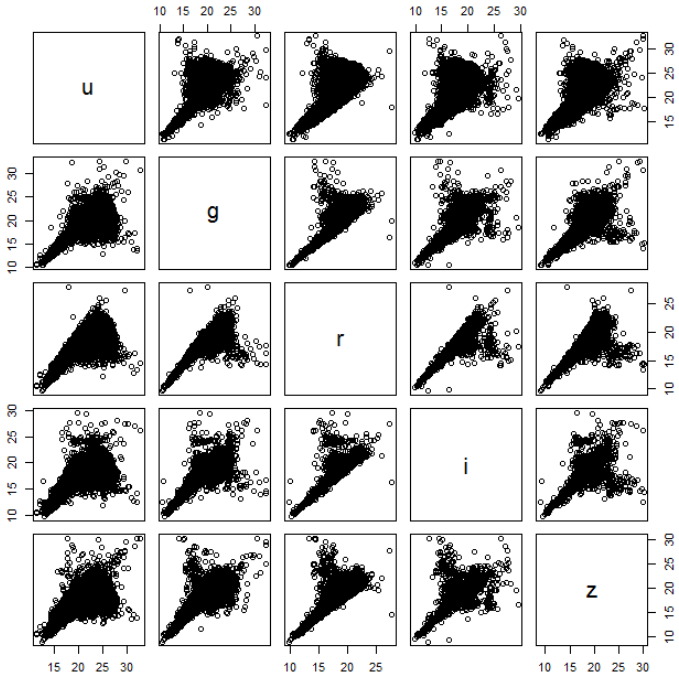
Első vizsgálatként az adatokat érdemes kirajzoltatni egymás függvényében, hogy vizuálisan valamilyen fogalmat nyerjünk egy lehetséges függvénykapcsolatról. A 2. ábrán látható a vöröseltolódás baloldalon u, jobboldalon pedig $u-g$ függvényében. A baloldali ábrán látható egy csak u függvényében illesztett egyenes is. Az $u-g$ függvényében ábrázolt adatsor közelebbinek tűnik a lineárishoz.

Hogy láthassuk a változók közötti korrelációkat, kirajzolhatjuk a szórási mátrixot, amely egy ábrasorozat, amiben a kiválasztott változók közül mindegyiket ábrázoljuk mind függvényében. Ez arra lehet jó, hogy lássuk, mely paraméterek állnak egymással kapcsolatban, és ez segítsen kiválasztani, hogy a modellalkotásba melyikeket szeretnénk bevenni. A nyers színadatokra ezt a 3. ábra tartalmazza.

Látszik, hogy a színértékek jól korrelálnak, de nagy fényesség esetén már szórni kezdenek. A korreláció alátámasztja azt, hogy a színképek hasonló



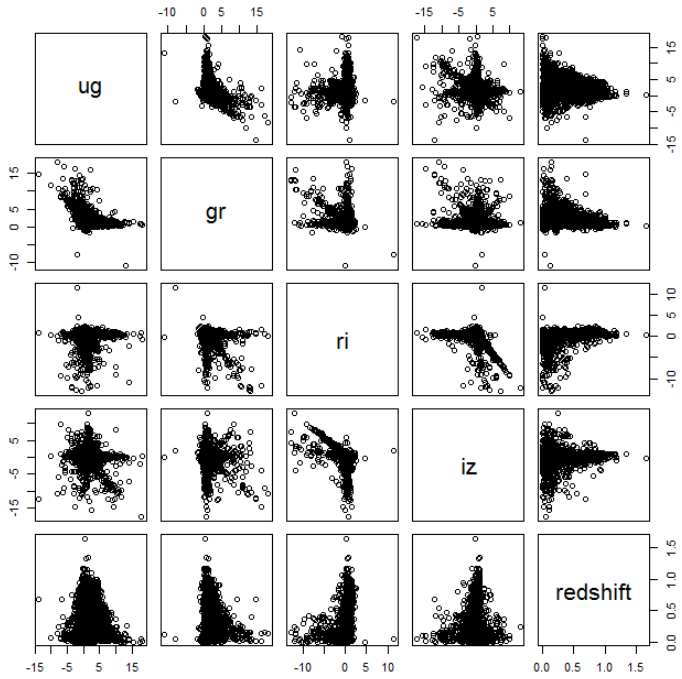
2. ábra. A vöröseltolódás az u, valamint az $u-g$ függvényében. látszik, hogy lineáris modellel közelíthetőek az adatok, de nagy a szórásuk, vagyis nagy hibára kell számítani.



3. ábra. A színek szórási diagramja. Látható, hogy kis értékekre mindegyik korrelál, de nagyobb magnitúdóra elkezdenek szórni.

típusú galaxisokra hasonlóak. Ezt a korrelációt transzformáljuk ki az adatokból a különbségek képzésével.

A különbségekre kirajzoltathatjuk ugyanezt az ábrát, ez a 4. ábrán látható. Itt az átfedések miatt a szomszédos adatoknál megjelenik egy - 45 fokos egyenes, és a 0-s egyenesekhez közel is sok adat kerül, de ezeken kívül trendek sokkal kevésbé figyelhetők meg. A továbbiakban mind a négy változót használni fogom a modellekben.



4. ábra. A különbségek szórási diagramja. Az identitás ellentettje sok helyen megjelenik, ahogy a 0 értékek körül is sok adat van. Azon kívül nem látszik egyértelműen korreláció.

3. Paraméterek becslése

3.1. Lineáris modell

Elsőnként lineáris modell illesztésével próbálkoztam. Az R `lm` függvényének segítségével lineáris modellt illesztettem mind a négy különbségi adatsorra, hogy megjósolja a vöröseltolódásokat. A modell adatai:

```
Call:
lm(formula = redshift ~ ug + gr + ri + iz, data = tr.dif)

Residuals:
     Min       1Q   Median       3Q      Max
-2.67014 -0.04200 -0.01336  0.02700  2.27024

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0618531  0.0003471 -178.18   <2e-16 ***
ug          -0.0163730  0.0001838  -89.10   <2e-16 ***
gr           0.1677741  0.0003338  502.62   <2e-16 ***
ri           0.1762213  0.0006883  256.01   <2e-16 ***
iz           0.0304457  0.0005404   56.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

Residual standard error: 0.07352 on 436270 degrees of freedom
Multiple R-squared: 0.5935,     Adjusted R-squared: 0.5935
F-statistic: 1.593e+05 on 4 and 436270 DF,  p-value: < 2.2e-16
```

A paraméterekhez tartozó hiba mindenhol maximum % nagyságrendű, és a szignifikancia szintek is jók. Ez is alátámasztja, hogy minden változót használnunk kell, különben jelentős információt vesztenénk.

Látszik, hogy lineáris modellünk működik, bár hagy némi kívánnivalót maga után.

Az elért pontszámom ezzel: 0,07325

A pontszámon van még mit javítani.

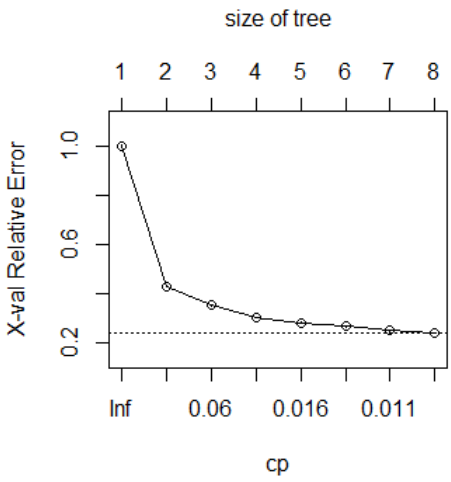
3.2. Regressziós fa

Az adatok becslésének egy elterjedt módja a döntési fák generálása. A döntési fa egy speciális fa gráf, ahol a belső pontok, azaz elágazások valamelyik változó egy értékéhez tartoznak, és attól függően, hogy az adott változó nagyobb-e, vagy kisebb egy megadott értéknél, kerülünk különböző ágakba. Több döntés után valamilyen végpontba jutunk, ami a becsülendő változónak ad valamilyen értéket.

A döntési fa oly módon áll elő, hogy az adathalmazunkat valamely érték alapján részhalmazokra bontjuk, majd ezt a felbontást ismételjük, a részhalmazokon, amíg a végpontokhoz már csak egy adott érték tartozik, vagy egy előre megadott hibánál kisebb a hiba.

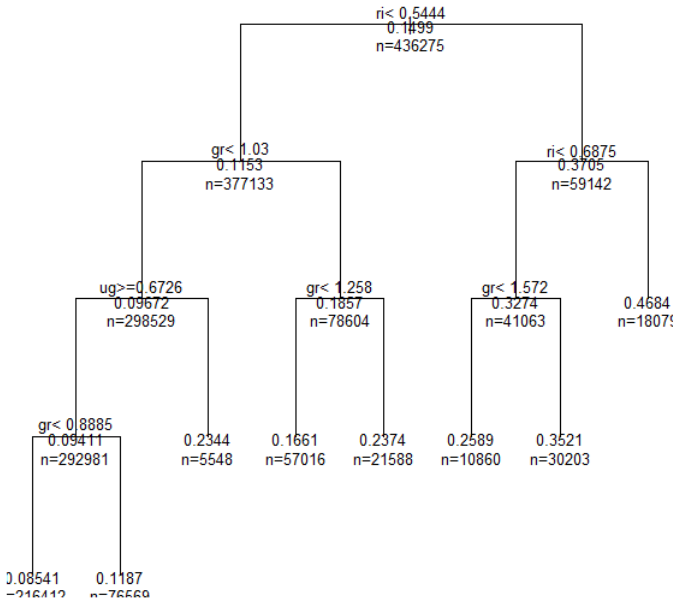
Az R nyelv `rpart` csomagjának függvényei ilyen problémák kezelésére valók. Az `rpart` függvény egy ilyen döntési fát generál le, és a modellt elmenthetjük egy objektumba, hogy megvizsgáljuk, vagy prediktáljunk a segítségével egy más adatsorra. A kimenet típusától függően többféle modellt is generálhatunk. Mivel a kimenetünk egy folytonos változó, nekünk a regressziós fára lesz szükségünk. A program kereszt validációs módszerrel számolja a hibát. Ez azt jelenti, hogy a halmazt különböző módokon kettévágja, és a modell alapján az egyikből próbálja megjósolni a másikat. Ebből számol hibát.

Szintén a négy különbségi változó figyelembevételével regressziós fát alkottam. A program 7 vágást csinált. A kereszt validációból számolt hiba az osztópontok számának függvényében az 5. ábrán látható.



5. ábra. A regressziós modell hibája az osztópontok számától függően.

A töntési fa a 6. ábrán látható.



6. ábra. A regressziós fa felépítése.

8 kimenetel talán kevésnek tűnhet egy majdnem félmilliós, bonyolult adatsor megjóslására, de a modellben durva nemlinearitás van, és a tesztek alapján a jóslóképessége megfelelő. Megjegyzendő, hogy végül az *ug* változót nem is használja.

Az így kapott adatsorra számolt pontszám: 0.05649.

Ez lényeges javulást jelent a lineáris modellhez képest.

4. Eredmények összefoglalása

Kétféle modell segítségével kíséreltem meg jósolni a vöröseltolódásokat a galaxisok színképe alapján. Egy 4 dimenziós lineáris modellel, valamint regressziós döntési fával. A kapott pontszámok:

- Lineáris modell: 0,07325
- Regressziós fa: 0,05649

Habár az adatokat a verseny letelte után küldtem, a portál lehetőséget adott rá, hogy lássam, milyen pontszámot kaptam volna. Ez alapján a harmadik helyre kerültem volna.