
Gépi tanulási módszerek a fotometrikus vöröseltolódás-becslésben

Fizika BSc szakdolgozat

Horváth Bendegúz

az ELTE TTK Fizika BSc hallgatója

Témavezető: Dr. Csabai István egyetemi tanár, Komplex Rendszerek Fizikája Tanszék

Budapest, 2018 május

Kivonat

Ide kell megírnod a kivonatot.

Tartalomjegyzék

1. Bevezetés	3
2. Vöröseltolódás-becslés és gépi tanulás	5
2.1. Fotometrikus vöröseltolódás-becslés	5
2.2. A gépi tanulási módszerek	7
2.2.1. <i>Random Forest</i>	8
2.2.2. Mesterséges neurális hálók	8
2.3. Eddigi eredmények és módszerek áttekintése	11
3. Adatok	13
3.1. Sloan Digital Sky Survey	13
3.2. A tanítóhalmaz elkészítése	14
3.3. Adatexploráció	16
4. Módszerek és eredmények	19
4.1. Neurális hálók és Random Forest kombinálva	19
4.2. Egy mély konvolúciós háló	21
4.3. Mély konvolúciós háló kiegészítve magnitúdóadatokkal	24
5. Összegzés	25
Irodalomjegyzék	27

1. fejezet

Bevezetés

Az Univerzum nagy skálás szerkezetének megértéséhez szükséges, hogy térképet tudjunk készíteni a galaxisok elhelyezkedéséről. Két koordinátát, a galaktikusszélességet és galaktikushosszúságot könnyen megkaphatjuk, viszont a távolság meghatározása már nehezebb feladat. A trigonometrikus parallaxis módszer a legjobb technikákkal is csak galaxison belül működik, a jó *seeing* érdekében pedig ūrtávcső kell. A standard gyertya módszerek pontos távolságértéket adnak, de csak néhány százmillió fényév távolságon belül alkalmazhatóak, ezért kell egy olyan módszer, amivel távolabb is mérhetünk, pontosan.

A XX. század elején Edwin Hubble és Vesto Slipher a galaxisok színképének tanulmányozása során észrevették, hogy a színképek eltolónak a nagyobb hullámhosszak, a vörös színtartomány felé a laboratóriumban mért vonalszerkezethez képest, a galaxis vöröseltolódást szenved. Hubble méréseket készített a galaxisok távolságáról és vöröseltolódásáról, és lineáris összefüggést tapasztalt, amit a később róla elnevezett törvény ír le:

$$v = H \cdot d, \quad (1.1)$$

az arányossági tényező a Hubble-állandó, értéke $H = 73.45 \pm 1.66 \text{ km/Mpc}$. Ez az összefüggés lehetőséget ad a pontos távolságmérésre, ha a galaxisok vöröseltolódását meg tudjuk mérni.

Vöröseltolódás mérésre a leg pontosabb módszer felvenni az objektum spektroszkópiai képét, és megnézni, hogy a vonalak mennyire csúsztak el. A megfelelő minőségű spektrum

felvételéhez akár egy órányi távcsőidő is szükséges lehet, ezért sokáig nem is készült égbolttérkép. Az 1986-ban Margaret Geller és munkatársai által készítet égtérkép csupán egy vékony szeletet fedett le az égből, de már azon is kirajzolódott, hogy a galaxiseloszlás nem egyenletes azon a skálán. Gellerék térképén a legtávolabbi galaxisok körülbelül 200 Mpc távolságnyira voltak tőlünk, ezért indokolt volt elkészíteni egy még távolabb látó térképet. A BEKS égfelmérés szűkebb, 1×1 négyzetfokos tartományban, ceruzaszerűen¹ nézett $1000\text{-}1000 \text{ Mpc}$ távolságba minden irányba, és ezen a skálán is kirajzolódott struktúra az anyageloszlásban, és a sűrűség ingadozás periodikusnak mutatkozott. Nagy távolságokra csökkent az észlelt galaxisszám, ez a halványabb galaxisok nehezebb észlelhetősége miatt volt. A technikai fejlődésnek köszönhetően elkészülhettek olyan kísérleti berendezések, amelyek lehetővé tették valódi háromdimenzióban a galaxisok pontos helyzetének felmérését. A Sloan Digital Sky Survey egyike ezen eszközöknek, első fázisában 1 millió objektum spektrumát vette fel négy év alatt[1]. Ez a sebeség nem kielégítő, ezért felmerült az igény, hogy a vöröseltolódásokat fotometriai úton mérjék. A fotometriával mért vöröseltolódások kicsit pontatlanabbak, de mérésük gyorsabb mint spekroszkópiával, ezenkívül halványabb objektumokat is lehet vele mérni, magasabban van a magnitúdó korlát. Ezen tulajdonságok vonzóvá teszi a vöröseltolódás becslését fotometriai módszerekkel. A technológiai fejlődés nem csak a csillagászatot érintette, egyre erősebb grafikus proceszorok jelentek meg, amik lehetővé teszik gépi tanulási módszerek szélesebb problémakörre való alkalmazását, mint például a vöröseltolódás-becslés fotometriai adatokból.

Dolgozatom célja az általam készített, gépi tanuláson alapuló fotometrikus vöröseltolódás-becslő módszerek bemutatása, amelyek a galaxisok képeit használjaák fel fotometriaia adatként.

¹pencil-beam survey

2. fejezet

Vöröseltolódás-becslés és gépi tanulás

2.1. Fotometrikus vöröseltolódás-becslés

A vöröseltolódás fotometriával történő becslésének kétféle megközelítése van, empirikus és spektrumokon alapuló. Egy spektrumon alapuló módszert először 1962-ban írt le Baum[2], fotoelektronos fotométert használt kilenc sáváteresztő filterrel, amik 3730 Å és 9875 Å közötti hullámhosszú fényt engedtek át. Ezzel a rendszerrel 6 fényes elliptikus galaxis spektrális energia-eloszlását (spectral energy distribution, SED) mérte meg a Virgo halmazból, majd még háromnak egy másik halmazból. A SED-ek átlagát ábrázolta a hullámhossz logaritmusának függvényében, és képes volt észrevenni az eltolódást a két energia sűrűségeletoszlás között, így a második klaszternek a vöröseltolódását is megkapta. Mérése pontos volt, de a módszer arra támaszkodott, hogy a spektrumoknak 4000 Å-nél levágása van, melyet a csillag-atmoszférák fémtartalma okoz, ez az elliptikus galaxisoknál jól látható, de például az aktív csillagkeletkezést mutató irreguláris galaxisokban ez a levágás nem figyelhető meg, ezért ez a módszer csak elliptikus galaxisoknál volt használható[3].

David C. Koo egy másik módszert, a szín-szín diagrammok módszerét vezette be 1985-ös cikkében[4]. Négy sáváteresztő szűrőt használt, melyeken mért magnitúdók különbségével létrehozott színtereken ábrázolva az azonos típusú, különböző vöröseltolódású galaxisokat jól definiált görbét kapott. Szín-szín diagrammokat bármilyen kombinációjából lehet készíte-

ni három vagy több színnek, az optimális választás a várt vöröseltolódás-eloszlástól függhet, a gyakorlati választás a rendelkezésünkre álló szűröktől[4]. Ez a megközelítés a fotometriai vöröseltolódás-meghatározás fontos eleme lett, ugyanis az egyes galaxisok típusának és színszűrőkön mért fényességének ismeretében meghatározható, hogy milyen vöröseltolódást szeneved a galaxis.

Egy harmadik, gyakran használt módszer a *template fitting* (sablon illesztés), ez a módszer is a spektrális energiasűrűség-eloszlásra támaszkodik, az ismert vöröseltolódású és SED-ű galaxisok SED-jéből felépül egy könyvtár, és az ismeretlen vöröseltolódású galaxisok SED-jét hozzá lehet párosítani hasonlóság alapján a könyvtárban lévőkhöz. A *template* módserek nagyon hasznosak lehetnek új égboltfelmérésnél, ha nem áll rendelkezésre megfelelő mennyiségű spektroszkópiai adat. A módszer használatával, különösen ha elméleti sablonokat használnak[5], a vöröseltolódáson kívül egyéb fizikai tulajdonságát is ki lehet nyerni a galaxisnak. Megfelelő használatához a sablonkönyvtárnak teljesnek kell lennie, hogy össze lehessen kötni mindegyik mérni kívánt galaxis SED-jét egy sablonnal, különben szisztematikus hiba jelenik meg a mérésben, viszont a túl sok sablon degenerációhoz vezethet. A sablon-illesztő és spektrumokon alapuló módszerek egyébb változatait és eredményeik összevetését jól leírják Hildebrandt és társai[6].

Empirikus módszerek alkalmazásához szükség van nagy mennyiségű spektroszkópiával mért vöröseltolódás-adatra és hozzájuk valamelyen, fotometriával mért adatokra. Az egyik legegyszerűbb módszer, hogy a színszűrőkön mért magnitúdóértékekere többváltozós lineáris vagy kvadratikus függvényt illesztenek[7]. Ezeket az adatokat felhasználva a függvény-illesztés helyett lehet gépi tanuló eljárásokat is alkalmazni, például *nearest neighbour*[8], *random forest*[9] vagy neurális hálókat[10]. Ezeknek a módszereknek előnye, hogy használatuk viszonylag egyszerű, hatalmas adathalmazokon is működhetnek, illetve nincs szükség a SED-re se, de nagy mennyiségű és jó minőségű tanulóhalmaz kell az előkészületekhez.

2.2. A gépi tanulási módszerek

A gépi tanuló módszerek már a XX. század közepén megjelentek, de a számítástechnika és a rendelkezésre álló adatok mennyisége még nem állt olyan szinten, hogy töretlenül fejlődhessen. Az akkori megközelítés szorosan összefüggött a mesterséges intelligencia kutatásával, de az 1990-es években az irány eltolódott a gyakorlati jellegű problémák megoldása felé. Ma már egyre több használható adat és fejlett grafikus processzorok mellett sokféle gépi tanuló algoritmus lett implementálva, így a nehézségek a megfelelő módszer megtalálása és alkalmazása az adatokra, illetve az adatok használható formába hozása. A gépi tanulás célja, hogy a gép *megértsze* az adatok szerkezetét, felismerjen egy szabályt és ez alapján jóslatokat tegyen.

Három nagyobb kategóriába sorolhatjuk a módszereket a rendelkezésre álló adatok és problémák alapján. Az egyik csoport a felügyelt tanulás, ilyenkor rendelkezésünkre álló adatok jelöltek, van egy *ground truth*, amit a modell jóslatainak meg kell közelítenie. A felügyelt tanulási módszereket osztályozás és regressziós problémák megoldásához használják, napjainkban az egyre nagyobb felcímkezett adathalmzoknak köszönhetően egyre több problémára tudják alkalmazni. A felügyelet nélküli tanulási módszerek felcímkezetlen adatokkal dolgoznak, feladatuk, hogy felfedjék a az adtokban rejttet struktúrákat. Az *ground truth* hiánya miatt nem lehet jellemezni a tanulás minőségét számértékkal. A fotometrikus vöröseltolódás-becslésben a várt végeredmény jól meghatározott, ezért felügyelt tanulási módszereket alkalmaznak [10], [12]. Ezeknek a módszerek megértéséhez fontosnak tartom bemutatni a munkám során használt eljárások általános a működési elvét, és alkalmazhatóságának határait.

Felügyelt tanulásnál rendelkezésünkre áll egy $(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)$ adathalmaz, ahol x_i egy mintát jelöl és a tulajdonságok(features) számával megegyező dimenziójú vektor, y_i jelöli az osztálycímét vagy a minta értékét regressziós problámákban, dimenziója az osztályok számával megegyező. A feladat, hogy a gép megtalálja azt a leképzést, a tanítóhalmaz mintái alapján, ami a lehető legkisebb hibával képez x_i -ből y_i -be még nem látott minták esetén is. A hiba mérését az adatokhoz és a feladathoz illő metrikával kell végezni. A kutatáshoz kétféle módszert használtam, *random forest regressort* és neurális hálókat.

2.2.1. Random Forest

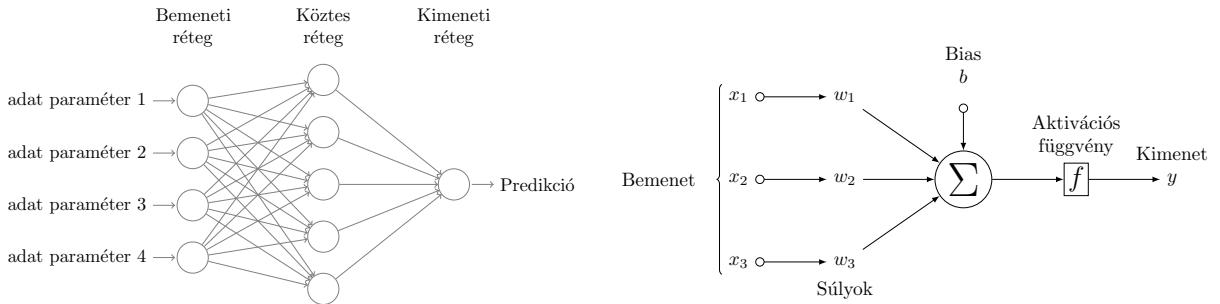
A *random forest* algoritmus egyik jó tulajdonsága, hogy alkalmazható klasszifikációs és regressziós problémákra is, alapját a döntési fák sokasága képezi, amiket összefűzve pontosabb becslést tud adni. A döntési fák felépülése a gyökér csomópontnál kezdődik, ami tartalmazza a tulajdonságokat és a célértéket. A csomópontnál meglévő jellemzőket felosztják az így létrejövő utódpontok között, igyekezve a hibát minimalizálni. A folyamatot tovább iterálva létrejön egy rétege a csomópontoknak, amelyek így egy fát alkotnak. A fa kialakulását szabályozni lehet, paraméterek lehetnek, hogy hány hány rétegű, milyen mély legyen a fa, minimum hány elem kerüljön egy levélbe¹, minimum hány felé legyen osztva a minta egy csomópontnál, vagy milyen módon mérje a szétválasztás minőségét. Véletlenség az erőbe úgy kerül, ha az egyes fáknál a szétválasztás a csomópontoknál nem a lehető legjobb *split* szereint történik, hanem véletlenszerűen kiválasztott részét kapják meg a tulajdonságoknak. Erre azért van szükség, mert a fontosabb tulajdonságok dominálnának mindegyik fa döntésében, így az egyes fák predikciói korreláltak lennének [13]. Az erőbe több fát adva az algoritmus nem tanul túl, becslései jobban konvergálnak a kívánt végeredményhez, de határértéke van a hibának [14]. Előnye még, hogy felfedi a prediktálás szempontjából fontos tulajdonságokat, sok attribútummal rendelkező adathalmazoknál ez segíthet az összefüggések megértésében. A fő korlátai, hogy a túl sok fa lassú prediktálást eredményez, illetve regressziós problémák-nál a modell nem tud extrapolálni, csak a tanulóhalmaz értékkészletein belüli eredményeket ad. Komplexebb vagy zajosabb adatoknál érdemes máshogy próbálkozni, ezen esetekben például mesterséges neurális hálókkal jobb eredményt lehet elérni.

2.2.2. Mesterséges neurális hálók

A mesterséges neurális háló egy leegyszerűsített modellje a biológiai neurális hálónak, megtartva annak jó tulajdonságait és a tanulás mechanizmusát. Alap építőkövei a neuronok és a súlyok², a neuronok a súlyokon keresztül vannak összekötve egymással és ezek adják

¹levél a döntési fának az a része, amiből nem nő ki több ág

²biológiai analógiája az axonok



2.1. ábra. Egy nerális háló vázlatos modellje és egy neuron aktivációjának a folyamata.

meg, hogy az egyik neurontól a másik milyen súllyal kapja meg az értékét. A neuronok rétegekbe vannak rendezve: bemeneti réteg, köztes réteg és kimeneti réteg³, a köztes réteg több rétegből is állhat. Az egyazon rétegen lévő neuronok nincsenek összekötve egymással, értéküket az alattuk lévő neuronuktól kapják a súlyokkal számolva. Matematikai formában az értékátadás:

$$z_i = w_{i1}x_1 + w_{i2}x_2 + \dots + w_{in}x_n + b_i, \quad (2.1)$$

az egyenletben z_i a rétegen az i -edik neuron, w_{ij} az x_j előző rétegbeli neuron és z_i neuron összekötő súly értéke, b_i pedig a *bias* vektor i -edik eleme.

Ez átírható egy egész rétre:

$$\underline{z} = \mathbf{W}\underline{x} + \underline{b} \quad (2.2)$$

A \mathbf{W} a w_{ij} súlyokból képzett súlymátrixot jelöli. Ahhoz, hogy a háló bonyolultabb problémákat is meg tudjon oldani, szükséges nemlineárítást vinni a rendszerbe. Ehhez aktivációs függvényt alkalmazunk, ami eldönti, hogy a neuron aktivizált legyen vagy sem. Gyakran alkalmazott aktivációs függvény a sigmoid:

$$S(z) = \frac{1}{1 + e^{-z}} \quad (2.3)$$

Illetve a *rectified linear unit*(ReLU):

$$R(z) = \max(0, z) \quad (2.4)$$

³input layer, hidden layer és output layer

A ReLU előnye, hogy a súlyok optimalizálásánál történő deriválásoknál az eltünő gradiens problémája nem áll fenn, így gyorsabb tanulást eredményez és a sigmoidhoz képest nem olyan szűk intervallumon ad vissza értékeket. A bemenő adatok ilyen transzformációkon mennek keresztül, míg kimenő adatként össze lehet hasonlítani a várt kimenettel. Az összehasonlítás a hibafügvénnyel történik, ami kiszámolja a két érték közötti távolságot valamelyen metrikával. Legtöbbször a hibafüggvényt több különböző adat becslése után értékeljük ki, a háló *batch*-okban kapja meg az adatot. Régessziós problémákban kedvelt hibafüggvény a *mean squared error*:

$$MSE = \frac{1}{n} \sum_{i=0}^n (y_i - p(x_i))^2 \quad (2.5)$$

a képletben a $p(x_i)$ az x_i adatkból prediktált érték, n pedig a *batch* méret. A hiba minimizálása egy optimalizációs probléma, a p függvény a súlyuktól függ, amiket úgy kell megváltoztatni, hogy a hiba minimális legyen. A súlyok optimalizálása *backpropagation*-nel történik, a hibafüggvényt deriválva a súlyok szerint a láncszabálytal, az utolsó rétegtől kezdve az elejéig, megkapjuk a gradiensmátrixokat rétegenként, amit szorozva egy kis számmal(*learning rate*), levonva a súlymátrixokból megkapjuk az új súlyokat. A súlyok frissítése *batch*-onként történik, így a *batch*mérettel és a *learning rate*-tel is lehet a tanulást gyorsítani, majd pontosabbá tenni [16].

A hálók architektúrájának kialakítására nincs általános szabály, intuícióinkra és hasonló problémáinkon jól szereplő hálók mintájára kell hagyatkozni, tanulásából követkztetéseket levonni és alakítgatni. Általánosságban a komplexebb problémákhoz *mélyebb*, több neuronból álló hálók kellenek, de ekkor a beállítandó paraméterek száma is nagyobb lesz, a tanítás lassabb lesz a megnövekedett számítási igény miatt. Képi adatoknál a paraméterek száma nagyon magas lenne, valamint a képnek nem az egész része érdekes, csupán részletek, és ezek a részletek bárhol lehetnek a képen, ezért nem célszerű teljesen összekötött neurális hálókat használni. Ezeknek a problémáknak a megoldására találták ki a konvolúciós neurális hálókat.

A konvolúciós hálókban a rétegek között nincs teljes összeköttetés, a neuron csak az alatta lévő neuronnal, és annak szomszédaihoz kötődik. Az objektumfelismerő-osztályozó problémáknál az objektum bárhol előfordulhat a képen, ezért a neuronokat összekötő súlyoknak is

eltolás invariánsank kell lennie, ezért egy rétegben minden neuron ugyanazokkal súlyokkal összegzi az allata lévő neuron szomszédainak kimenetét. A figyelembe vett szomszédonk száma meghatároz egy *filter* méretet, a *filter* technikailag egy tenzor, aminek elemei a súlyok. A bemeneti képet végig páztázza a filter, és végrehajta az összegzéseket, az egyes kimenetek egy *aktivációs térképet alkotnak*⁴, amire alkalmazhatjuk a következő réteg konvolúciós réteget. Ez a műveleletet kétdimenziós diszkrét konvolúció a matematikában. A kimeneti *aktivációs térkép* formája egy n széles és hosszú, c csatornával rendelkező képen alkalmazott $f \times f$ méretű, s ugrással mintavételező *filter*-rel, a képre p *padding*⁵-et alkalmazva a konvolúció során a következő képpen alakulnak:

$$a = \frac{n + 2p - f}{s} + 1, \quad (2.6)$$

a képleteben a létrejövő *aktivációs térkép* szélessége, k *filtert* alkalmazva $a \times a \times k$ alakú *aktivációs térképet* kapunk. Az *aktivációs téképek* méretét csökkenteni kell az első kettő dimenziójában, k a *filterek* számával lesz egyenlő. Erre egy módszer a *pooling*. A *poolingoknak* sok fajtája van, legnépszerűbb a *Maxpooling* és az *AveragePooling*, alapja, hogy a *filterekhez* hasonlóan végig pásztázza a képet egy *kernel*, *Maxpooling* esetében a *kernel* méreten belüli elemek közül a maximális nagyságút adja át az *aktivációs térképnek*, *Averagepoolingnál* pedig a kernelen belüli elemek átlagát adja át.

2.3. Eddigi eredmények és módszerek áttekintése

⁴szokásos feature map-nek is nevezni.

⁵kép szélén nullákkal létrehozott keret

3. fejezet

Adatok

Az eredményes gépi tanuláshoz a nagy mennyiségi és jó minőségű adat majdnem annyira fontos, mint a jó modell megválasztása. Általában nem áll rendelkezésünkre egyből az algoritmusnak adható adat, a nyers adatokat előbb preprocesszállni kell, ki kell nyerni a fontos tulajdonságokat, melyek előzetes ismereteink alapján fontosak lehetnek, és össze kell állítani az adathalmazt, amelyet majd szétválasztunk tanuló- és teszthalmazra.

A kutatómunkámhoz sok, jó minőségű galaxis képére volt szükségem, és mindegyik galaxisnak a spektroszkópiával mért vöröseltolódására is, ezekhez az adatokhoz a Sloan Digital Sky Survey adatbázisában fértem hozzá.

3.1. Sloan Digital Sky Survey

A Sloan Digital Sky Survey az eddigi legrészletesebb égboltfelmérés, mély, több színávos képekkel az ég egyharmadáról, 500 millió asztrofizikai objektumról, 3 millió felvett spektrumadattal. A felmérést 2000-ben kezdte, több ciklusban, 2014-ben kezdődött a negyedik fázis (SDSS-IV), és már elkezdődtek a megbeszélések az SDSS-V elindításáról[17]. A megfigyelési adatokat egy külön erre a célra megépített 2.5 m széles optikai teleszkóp szolgáltatja az Apache Point Obszervatóriumból, Új Mexikóban. Öt színszínben mér, a látható fény és az infravörös tartománya között, u , g , r , i és z színszűrőkkel (ultraviolet, green, red,

near infrared és infrared), amik 3000 \AA és 10000 \AA közötti tartományt fedik le. A különböző szűrőkön keresztül a CCD chipek egymás után rögzítenek, 71.2 másodperc késéssel, r , i , u , z és g sorrendben, így a különböző színű képeken előfordulhat, hogy egy objektum kicsivel arrébb. Ennek volt előnye is, mozgó objektumokat könnyebben lehetett azonosítani, például létre tudtak hozni aszteroida katalógusokat. Az SDSS képek alapvető egysége a *field*, ami 10-szer 13 szögperces szeletet tartalmaz az égből, és 1489-szer 2048 pixelt tartalmaz és az egyes *fieldeket* három jelzőszám teszi egyedivé, a *field number*, a *run*, ami a szkennelés száma, a *camera column*, ami a megmutatja melyik oszlop CCD kamera készítette a képet egy 1 és 6 közötti szám, mindegyik oszlop egy *fieldet* készít, ami 128 pixel szélességen fed át a szomszédossal. Az elkészült felvételek az SDSS képfeldolgozó pipelinejába kerülnek, ahol kalibrált FITS formátumú képet csinálnak, és a katalógushoz hozzá adják a képi paramétereket.

A megfigyeléseket folyamatosan végzik, az adatbázisokat viszont csak évente frissítik, ezeket a *data release*-nek (DR) hívják és tartalmazzák az előző megfigyelésekkel származó adatokat is. Az SDSS virtuális oszervatórium keretrendszerben működik, adatai publikusak. Az objekumok és paramétereik relációs adatbázisba vannak rendezve, a SQL nyelvet használva lehet lekérdezéseket indítani.

3.2. A tanítóhalmaz elkészítése

A tanítóhalmaz elkészítését a *SciServeren*[18] végeztem, amely *Jupyter Notebook* keretrendszerben nyújtott adatfeldolgozó és számoló felületet, valamint közvetlen elérést az SDSS *DR7-es fieldekhez*. A szerveren egy *API* segítségével közvetlenül lehetett *python* környezetben SQL lekérdezéseket végezni az SDSS adatbázisából, az így kapott táblákat elmenteni. Két tanítóhalmazt állítottam össze, az 1-es adathalmaz 150 000 képből áll, a galaxisok minden egyik színszűrőn mért magnitúdója 15^m -nál halványabb és 22^m -nál fényesebb. A legtöbbször ezt az adathalmazt használtam tanításra és tesztelésre, viszont a vöröseltolódás-eloszlásában lévő csúcs miatt létrehoztam egy 2-es adathalmazt is, melyben a vöröseltolódás eloszlás kicsit egyenletesebb létrehozása során nem csináltam felső határt a magnitúdónak, de a vöröselto-

lódásokat határok közé szorítottam, több lekérdezés eredményéből állt össze az adattábla.,

Feltéteként szabtam meg, hogy a galaxisok vöröseltolódás 0.05-nél nagyobb legyen, hogy a nagy fényességük miatt a képeken szaturációt okozó csillagok ne keveredjenek bele, valamint 1-nél kisebb legyen a vöröseltolódás, hogy az SDSS adatbázisában tévesen galaxisnak osztályozott kvazárok se kerüljenek bele. A képméret kiválasztásánál figyelembe kellett venni, hogy a túl nagy kép túl sok haszontalan információt is tartalmazhat, valamint a közel szomszédos objektumok is félrevezethetik a neurális hálót, de a túlságosan kicsi képméret miatt lehet lemarad fontos információ. Ezért vizualizálva a galaxisokat különböző képmérettel, az optimális választásnak az 50×50 pixeles méret tűnt.

A *SciServer*-en az egyes *field*ek olyan könyvtárstruktúrába vannak rendezve, hogy a *run*, *rerun*, *camcol*, *fieldID* és színszűrő alapján kereshetők. Ezért az SQL lekérdezésnél a magasság, vöröseltolódás és objektum azonosító addatai mellett lekérdeztem *field* könyvtárban megtalálásához szükséges adatokat, valamint az egyes színszűrőkkel készített képen hol van a galaxis közepe¹. A galaxis közepének lehelyezkedésére az volt a kikötés, hogy ne a kép szélénél 25 pixeles szomszédságában helyezkedjen el, így a kivágásnál nem fog gondot okozni a hiányzó információ. Az így kapott táblákkal előtudtam hívni a FITS formátumú *field*eket, melyekből ki tudtam vágni a galaxis 50×50 méretű képét, levontam a *Softbias-t*², és az eredeti könyvtárstruktúrához hasonló módon elmentettem a képet, egy *python dictionary*be pedig az elérési útvonalat és az objektum azonosítóját, így azonosító alapján előhívható volt a kivágott kép. Ezt mindegyik szűrővel készült képre külön megcsináltam, majd az összes képet egy nagy tömbbe tettem a könnyebb mozgatás és elérés érdekében. A tömb dimenziója objektumok száma \times 12 500 lett, a különböző színű, azonos objektumról készült képeket egymás után tettem, így az adatokat tartalmazó tábla sorának sorszámaival lehetett megtalálni a képeket. Ez a tárolási mód okozhatna olyan hibát, hogy elcsúsznak, és nem a megfelelő indexnál lesznek a képek, de ezt véletlen mintavételezéssel megjelenítve a képet és az azo-

¹ sor és oszlop pixelben, a különböző színszűrőkön készített képeken máshol volt az objektum közepe a SDSS képalkotása miatt.

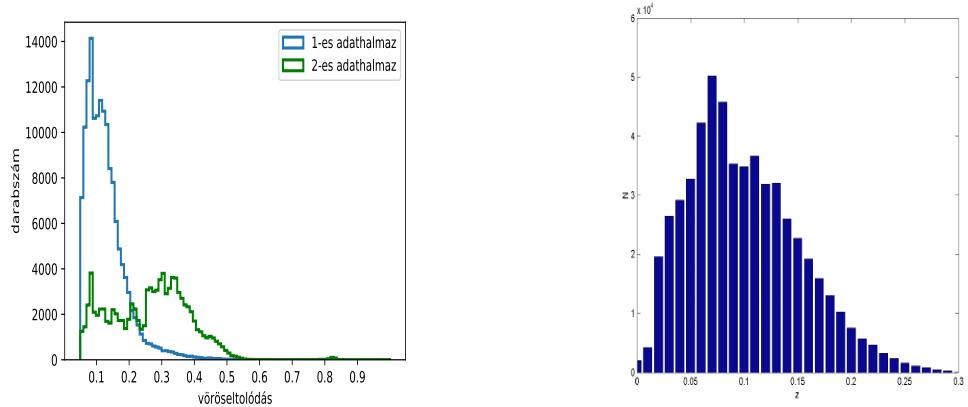
² a *field*ekhez hozzáadtak egy 1000-es nagyságrendű számot a kép készülése után, hogy ne legyen negatív pixelérték sehol.

nosítót, össze tudtam hasonlítani az SDSS képeivel. A képi adattömb az 1-es adathalmazhoz 14GB méretű lett, aminek a betöltéséhez nagy memóriájú gép kellett, a betöltött tömb formája már könnyen alakítható volt a feladat igényeihez.

3.3. Adateexploráció

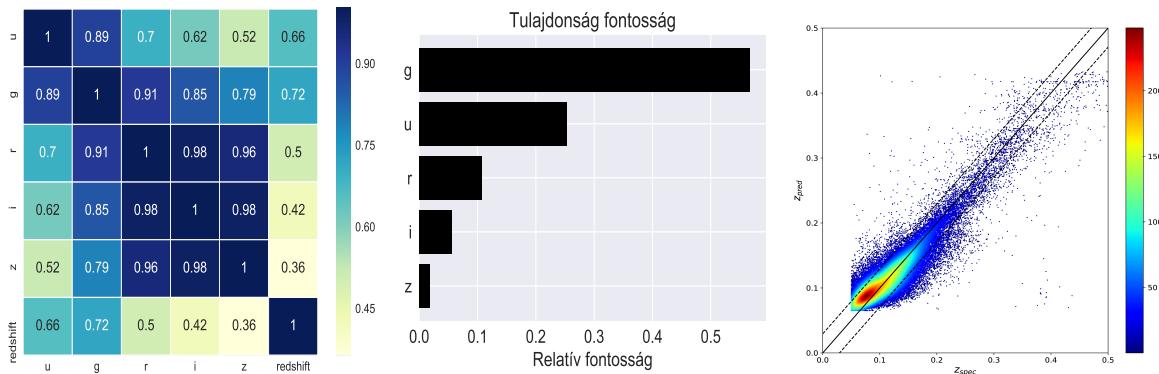
Mielőtt elkezdenénk a tanítást, érdemes megvizsgálni a tanulóhalmazunk statisztikai jellemzőit, megnézni mennyire jól reprezentálja a valóságot, valamint feltárni az összefüggéseket. Ehhez az SDSS *DR7*-es adatbázisának vöröseltolódás-eloszlásával hasonlítottam össze a tanulóhalmaz eloszlását, alkalmaztam egy *Random Forest regreesort* az *ugriz* magnitúdóértékekre, és készítettem egy *korrelációs térképet* a magnitúdókkal és vöröseltolódásokkal.

Az SDSS *DR7*-ben a galaxisok vöröseltolódásának mediánja $z \sim 0.07$ [19], a teszthalazomé $z \sim 0.11$, ami a 0.05-ös vöröseltolódásbeli vágásomnak tudható be. Összehasonlítva



3.1. ábra. Az általam előállított tanulóhalmaz és az SDSS *DR7* vöröseltolódás-eloszlása. A jobboldali kép forrása: [19]

a 3.1. ábrán a baloldalon az 1-es adathalmaz hisztogramját a jobboldalival, a két eloszlás jó hasonlóságot mutat, a vágásban van különbség, a legtöbb galaxisnak mind a kettő esetben 0.3-nál kisebb a vöröseltolódása. .



3.2. ábra. A korrelációs mátrix,a Random forest tanulásában a különböző tulajdonságok relatív fontossága és a magnitúdókból becslése. A szaggatott vonal a rmse nagyságát ábrázolja.

A korrelációs mátrix elemeiből az látszik, hogy az egymás melletti színek magnitúdója erősen korrelál a szomszédossal. A magnitúdok és a vöröseltolódás nem korrelál ilyen szinten, legmagasabb korrelációs együtthatója a g magnitúdóknak van a vöröseltolódással. Ezek után alaklmaztam az adatok közül az első 100 000-re a *Random forest*-et, majd prediktáltattam az utolsó 50 000 mintán. A véletlen erdő $rmse = 0.029$ jósággal prediktált, a becsült vöröseltolódás-értékeket a 3.2 ábra bal szélén ábrázoltam a spektroszkópiai vöröseltolódások függvényében. Az ábrán látható, hogy a pontok nagyrésze rafekszik az ideálist jelző vonalra, viszont ahol a legsűrűbb a pontok elhelyezkedése, szisztematikus hibája van a rendszernek, felül becsül. A legfontosabb paraméterek a becslés során a g magnitúdóérték bizonult, ennek relatív fontossága 0.59 volt, ami jóval magasabb, mint a második legfontosabb u értéknek, ami 0.24. A tulajdonságok relatív fontosságának sorrendje összhangban van az egyes fényeségek vöröseltolódással vett korrelációjával.

A magnitúdó adatok elemzése azért indokolt, mert a bemeneti adataink az öt csatornás képek lesznek, és a magnitúdók az egyes csatornákhoz köthető származtatott adatok. A legtöbb gépi tanuláson alapuló vöröseltolódás-becslő módszer ezeket az adatokat használja, ezért ha a képekből ki tudjuk nyerni gépi tanulással megfelelő pontossággal a fényességértékeket, alkalmazhatunk már bevált módszert rá.

4. fejezet

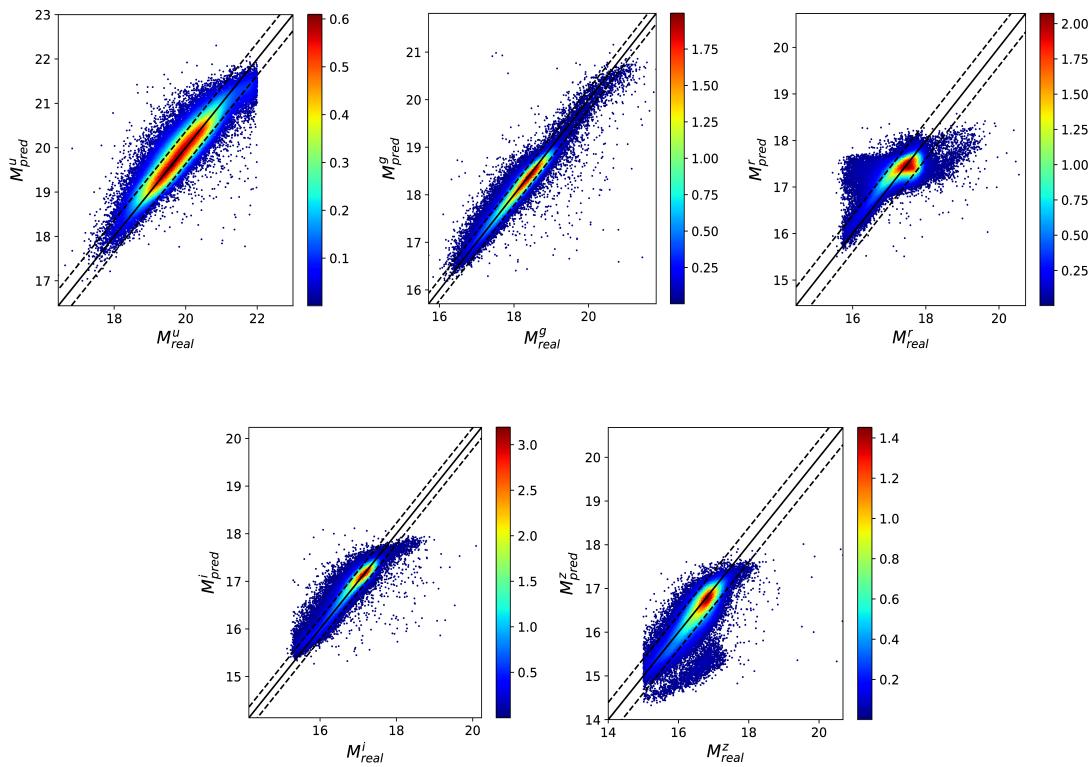
Módszerek és eredmények

A kutatáshoz az egyes gépi tanulási módszereket *python* nyelven a *keras* (neurális hálók) és *sci-kit learn (random forest)* könyvtárakkal valósítottam meg. A tanulás gyorsítása érdekében a tanítást a *Google Cloud* felhő alapú számítási szolgáltatást nyújtó platformon végeztem NVIDIA Tesla K80-as GPU-val.

4.1. Neurális hálók és Random Forest kombinálva

A *random forest* algoritmus jól teljesít a magnitúdókból prediktálás során, így célszerű megvizsgálni, hogy mennyire működik jól, ha a magnitúdóértékek is becslésből származnak. A képekből való fényességbecsléshez mind az öt színszűrőhöz elkészítettem egy-egy konvoluciós neurális hálót, ezeket a hálókat az első 50 000 képen tanítottam be. A következő 50 000 képpel prediktáltam magnitúdóértékeket, ezeket a becsült magnitúdókat használtam, a *random forest regressor* tanításához. A harmadik 50 000 képnek is megbecsültem a magnitúdóit, és ezekből predikált a véletlen erdő vöröseltolódás-értékeket, így nem volt átfedés a tanuló- és teszthalmazok között. A neurális hálók tanításánál 80 *epoch*-ot használtam minden egyik hálóra, de nem egyben ment végig a tanulás, a *learning rate*-t csökkentettem, ha a hibafüggvény értéke nem csökkent, illetve a *batch* méretet is növelte közben. Az egymás melletti¹ mag-

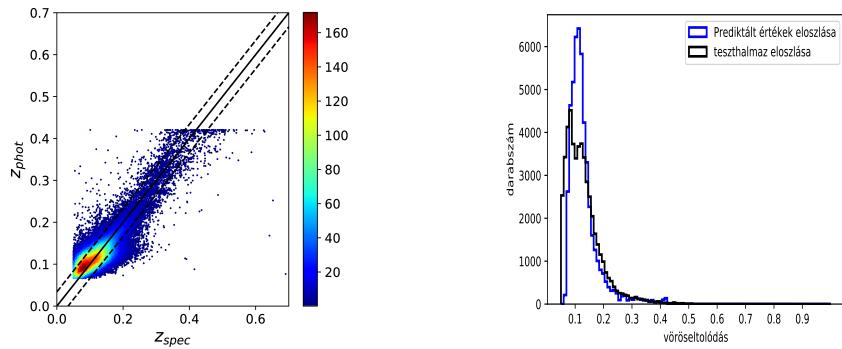
¹ hulámhossz szerint



4.1. ábra. Az egyes neurális hálók magniúdópredikciója ábrázolva a valódi magnitúdóértékek függvényében.

nitúdók korrellálása, illetve a célváltozók és a bemeneti adatok hasonlóság miatt az volt a feltételezésem, hogy ugyanolyan architektúrájú neurális hálókat lehet használni minden egyik fényességértek számításához.

Az egyes hálók bemenetként megkapták az $50 \times 50 \times 1$ méretű, az adott színszűrőn keresztül készített képet inputként. Az első réteg egy konvolúciós réteg volt, 32 darab 2×2 -es filterből állt, alkalmaztam rá a *ReLU* aktivációt és utána egyből egy 2×2 *MaxPooling*-ot. Utána a következő konvolúciós réteg 16 db 2×2 -es filterből állt, *sigmoid* aktivációval, majd egy 2×2 -es *MaxPooling* következett. A kijövő értékek kilapítása után következtek a teljesen összekötött rétegek, az első réteg 1024 neuronból állt, *sigmoid* aktivációval. Utána 15%-os *Dropout* regularizáció következett, ami a túllillesztés elkerülésére szolgált. Ezt követően 512 neuron, *Relu* aktiváció és 10%-os *Dropout*, 256 neuron *ReLU*-val és végül az egy darab ne-



4.2. ábra. A Random forest prediktált vöröseltolódás-értékei a spektroszkópiai vöröseltolódások függvényében és a két eloszlás.

uron. A hibafüggvény *mean squared error* volt. A *random forest regressor* 250 fából állt, maximális mélysége 15, egy levélbe minimum 120-elemnek kellett kerülnie, egy szétválasztáshoz legalább 28 minta volt szükséges és a szétválasztás jóságát *mean squared error*-ral mérte.

A 4.1 ábrán látható, hogy az r illetve z magnitúdókat nagy hibával becsülté, de az adat-exploráció felfedte, hogy a *random forest*-nek ezek az adatok kevésbé fontosak. A magnitúdóbecslésekben az $rmse$ -k a következők lettek: $rmse_u = 0.372$, $rmse_g = 0.220$, $rmse_r = 0.397$, $rmse_i = 0.223$, és $rmse_z = 0.393$, az alsó indexek a színt jelölik. Ekkora torzítással a magnitúdóadatokon a *random forest* $rmse = 0.0338$ pontossággal tudott becsülni. A becsült és a spektroszkópiai vöröseltolódás-eloszlásokban látszik a különbség, a prediktált eloszlás görbéje szűkebb, és a csúcs kicsit el van tolódva, ez szisztematikus hiba. A 4.2 bal oldali ábráján látszik, hogy 0.4-es vöröseltolódás körül vágása van, ez a véletlen erdők extrapolációs képességének hiánya.

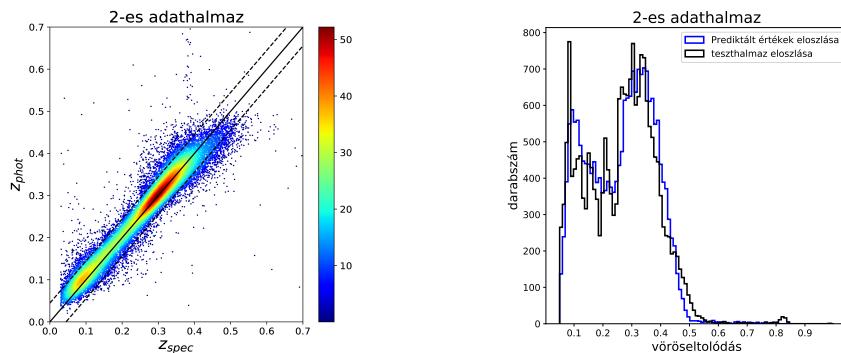
4.2. Egy mély konvolúciós háló

A *random forest*nél láthatóan működött a kinyert adatokból a vöröseltolódás-becslés, de kézenfekvőbb olyan módszert megvalósítani, ami kitalálja, hogy mi a fontos tulajdonság a

prediktálás szempontjából, és ezt ki is vonja automatikusan a képből.

A képi adatokból jól prediktáló konvolúciós hálók² architektúrájához hasonló modellt célszerű készíteni kiindulásként, de ezeket mind osztályozó, objektumfelismerő problémák megoldására használták, tanítására több adat és számítási kapacitás állt rendelkezésre, ezért a vöröseltolódás-becslő háló nem lesz olyan *mély*. A háló tanítását a 2-es adathalmaz előző 80 000 képével végeztem, majd a maradék 20 425 képpel teszteltem. Valamint végeztem tesztelést az összehasonlítás édekében az 1-es adathalmaz 100 000-től 1500000-ig elhelyezkedő képeivel is.

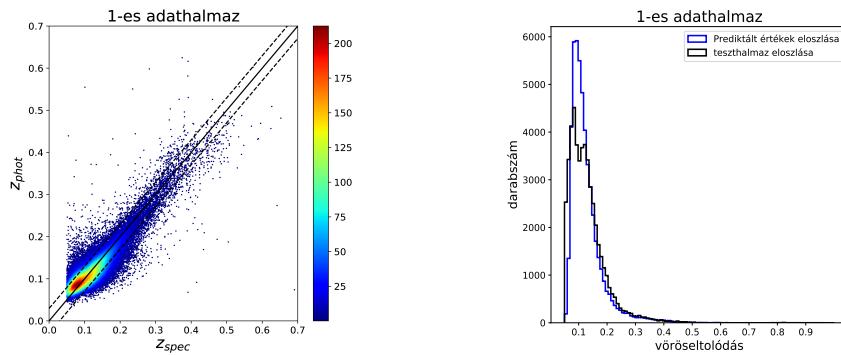
A modell architektúrája a következő volt: a bemenet az $50 \times 50 \times 5$ méretű kép volt, utána kétszer egymás után, 48 darab 3×3 -as filterből álló konvolúciós réteg, majd 2×2 Averagepooling. Utána következett 3 egymás utáni 96 3×3 filterből álló réteg, majd megint egy 2×2 -es Averagepooling, majd kétszer 200 darab 3×3 filter, 2×2 Averagepooling, egy 200 filteres, 3×3 konvolúciós réteg, egy BatchNormalizáció, majd megint egy 200 darab, 3×3 -as filteres réteg. Aztán 2×2 -es Averagepooling, majd egy 200 és egy 20 darabos 3×3 filteres réteg, utána 72 darab 1×1 filteres dimenzió csökkentő konvolúciós réteg. Mindegyik



4.3. ábra. A konvolúciós neurális háló becslései a spektroszkópiai vöröseltolódás-értékek függvényében a 2-es adathalmazon, és a vöröseltolódás-eloszlások.

rétegen *ReLU* aktivációt alkalmaztam. A teljesen összekötött rész 4096 neuronnal kezdődött, majd 400 neuron a következő rétegen, minden *ReLU* aktivációval, végül az 1 neuronból álló

²például VGG16, GoogLeNet



4.4. ábra. A konvolúciós neurális háló becslései a spektroszkópiai vöröseltolódás-értékek függvényében a 1-es adathalmazon, és a vöröseltolódás-eloszlások.

kimeneti réteg, aktiváció nélkül.

A tanításhoz a megnövekedett paraméterszám miatt több *epoch*ra volt szükség, mint a magnitúdóbecslésnél, összesen 100 *epoch*-ra, a *learning rate*-et csökkentettem, ha nem csökkent a hiba értéke, illetve a *batch size*-t növelte. A 2-es adathalmazon a 20 425 galaxison tesztelve a $rmse = 0.0447$ lett. A 4.3 ábra bal oldalán látható, hogy a prediktált értékeket jelölő pontok szimmetrikusak a 45° -os egyenesre a 0.2 és 0.45 közötti vöröseltolódás-tartományon, nincsen szisztematikus hibája. A 0.2-nél kisebb vöröseltolódásokat gyengén felülbecsli, a 0.5-nél nagyobb vöröseltolódásúkat pedig alul. Érdemes volt megnézni, hogyan teljesít a háló az 1-es adathalmazon, ahol több alacsony vöröseltolódású galaxist tartalmaz a tesztszett valamint összehasonlítható így a *random forest* módszerrel.

A 2-es adathalmazon tanult hálót 10 *epoch* erejéig tanítottam még az 1-es halmon előtt kiértékeltem volna, mert az eltérő tanító- és tesztszett eloszlás problémát jelenthet.³ A tesztadatokon $rmse = 0.0296$ hibát produkált, ami jobb mint a *random forest*é, viszont az alacsony vöröseltolódásoknál még mindig enyhén felülbecsül.

Ez a módszer jól alakalmazható a feltételezéssel...

³ A felügyelt tanulásnál gyakran alkalmazott az az előfeltételezés, hogy a célváltozó tanító- és teszthalmazának az eloszlása megegyezik. Elterő esetben, ha tudjuk, hogy a tanuló- és tesztminták nem azonos eloszlásból származnak *covariate shift* módszert lehet alkalmazni a korrigáláshoz [20].

4.3. Mély konvolóciós háló kiegészítve magnitúdóadatokkal

5. fejezet

Összegzés

Irodalomjegyzék

- [1] Z. Frei and A. Patkós, Inflációs Kozmológia: (Typotex, Budapest, 2005).
- [2] Baum, W. A.: 1962, Problems of Extra-Galactic Research, Proceedings from IAU Symposium no. 15. Edited by George Cunliffe McVittie. International Astronomical Union Symposium no. 15, Macmillan Press, New York, p.390
- [3] “Photometric Redshifts.” NASA/IPAC Extragalactic Database - NED, ned.ipac.caltech.edu/level5/Glossary/Essay_photredshifts.html.
- [4] Koo, D. C. “Optical Multicolors - A Poor Person’s Z Machine for Galaxies.” The Astronomical Journal, vol. 90, 1985, p. 418., doi:10.1086/113748.
- [5] Bruzual, G., and S. Charlot. “Stellar Population Synthesis at the Resolution of 2003.” Monthly Notices of the Royal Astronomical Society, vol. 344, no. 4, 2003, pp. 1000–1028., doi:10.1046/j.1365-8711.2003.06897.x.
- [6] Hildebrandt, H., et al. “PHAT: PHoto-ZAccuracy Testing.” Astronomy & Astrophysics, vol. 523, 2010, doi:10.1051/0004-6361/201014885.
- [7] Connolly, A. J., et al. “Slicing Through Multicolor Space: Galaxy Redshifts from Broadband Photometry.” The Astronomical Journal, vol. 110, 1995, p. 2655., doi:10.1086/117720.
- [8] Csabai, I., et al. “The Application of Photometric Redshifts to the SDSS Early Data Release.” The Astronomical Journal, vol. 125, no. 2, 2003, pp. 580–592., doi:10.1086/345883.

- [9] Carliles, S., et al. “Random Forests For Photometric Redshifts.” *The Astrophysical Journal*, vol. 712, no. 1, 2010, pp. 511–515., doi:10.1088/0004-637x/712/1/511.
- [10] Collister, Adrian A., and Ofer Lahav. “ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks.” *Publications of the Astronomical Society of the Pacific*, vol. 116, no. 818, 2004, pp. 345–351., doi:10.1086/383254.
- [11] Csabai, I., et al. “Multidimensional Indexing Tools for the Virtual Observatory.” *Astronomische Nachrichten*, vol. 328, no. 8, 2007, pp. 852–857., doi:10.1002/asna.200710817.
- [12] Collister, Adrian A., and Ofer Lahav. “ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks.” *Publications of the Astronomical Society of the Pacific*, vol. 116, no. 818, 2004, pp. 345–351., doi:10.1086/383254.
- [13] Ball, Nicholas M., and Robert J. Brunner. “Data Mining And Machine Learning In Astronomy.” *International Journal of Modern Physics D*, vol. 19, no. 07, 2010, pp. 1049–1106., doi:10.1142/s0218271810017160.
- [14] “Random Forests Leo Breiman and Adele Cutler.” Statistics at UC Berkeley, www.stat.berkeley.edu/~breiman/RandomForests/.
- [15] Sadeh, I., et al. “ANNz2: Photometric Redshift and Probability Distribution Function Estimation Using Machine Learning.” *Publications of the Astronomical Society of the Pacific*, vol. 128, no. 968, 2016, p. 104502., doi:10.1088/1538-3873/128/968/104502.
- [16] L., Samuel, et al. “Don’t Decay the Learning Rate, Increase the Batch Size.” SAO/NASA ADS: ADS Home Page, 1 Nov. 2017, adsabs.harvard.edu/cgi-bin/bib_query?arXiv%3A1711.00489.
- [17] Zasowski, Gail. Science Blog from the SDSS, blog.sdss.org/2018/02/21/sdss-v-is-underway/.

- [18] “SciServer – Collaborative Data-Driven Science.” SciServer, www.sciserver.org/.
- [19] Verevkin, A. O., et al. “The Non-Uniform Distribution of Galaxies from Data of the SDSS DR7 Survey.” *Astronomy Reports*, vol. 55, no. 4, 2011, pp. 324–340., doi:10.1134/s1063772911020089.
- [20] Mcgaughey, Georgia, et al. “Understanding Covariate Shift in Model Performance.” *F1000Research*, vol. 5, 2016, p. 597., doi:10.12688/f1000research.8317.3.

Nyilatkozat

Név: Horváth Bendegúz

ELTE Természettudományi Kar, szak: Fizika BSc

Neptun azonosító: ZNL3LK

Szakdolgozat címe: Fotometrikus vöröseltolódás becslés

A **szakdolgozat** szerzőjeként fejyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló munkám eredménye, saját szellemi termékem, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest 2018. május?.
