

Osztályozás

Csima Judit

BME, VIK,
Számítástudományi és Információelméleti Tanszék

2015. február 26.

Osztályozás, classification

- adott egy rekordokból álló halmaz, a rekordoknak attribútumaik vannak
- az egyik attribútum a célváltozó, ez kategorikus attribútum, ez reprezentálja, hogy melyik osztályba tartozik az adott rekord
- cél, hogy egy olyan modellt építsünk fel, ami képes megjósolni a célváltozó értékét, ha a többi attribútum értéke adott

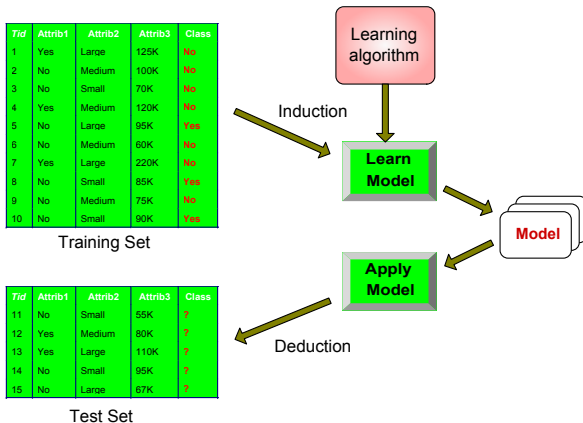
Példák

- egy beteg paraméterei alapján eldönteni, hogy jó- vagy rosszindulatú daganata van-e
- bankkártyás tranzakció adatait vizsgálva eldönteni, hogy van-e csalás vagy nincs (szabályos-e a tranzakció)
- fehérjék szerkezetét előre jelezni (alfa-hélix, béta-sheet, egyéb) az aminosav-sorrend alapján
- adóbevallásban szereplő értékek alapján megtippelni, hogy gyanús-e
- spam-szűrés

Általános módszer

- van egy csomó rekordunk, ahol minden érték (a célváltozó értéke is) ismert
- ez alapján egy modellt építünk, amit majd olyan új rekordokon használunk, amiknél a célváltozó értéke nem ismert, de minden más attribútumot tudunk
- az ismert rekordokat két részre osztjuk: training set és test set
- a training set-en betanítunk valami modellt
- a test set-en lemérjük, hogy mennyire jó
- ezután használhatjuk élesben

Illustrating Classification Task



Kérdések

- Milyen modellek vannak?
 - Döntési fák
 - Bayes-osztályozók
 - Mesterséges neurális hálózatok
 - (SVM: Support Vector Machine)
- hogyan állítunk elő egy konkrét modellt? Pl. ha már tudjuk, hogy döntési fát csinálunk, akkor hogyan csináljuk meg; vagy egy ANN-nél hogyan állítjuk be a paramétereket?
- Hogyan mérjük az előállított modell jóságát?
 - accuracy: eltalált címkék száma osztva az összes sor számával
 - error-rate: hibás predikciók száma osztva az összes sor számával

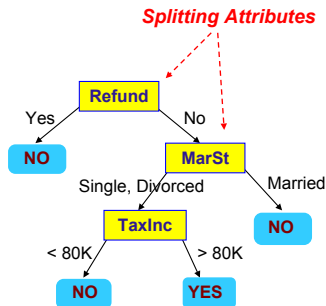
Döntési fa definíció

- gyökeres, lefelé irányított (legtöbbször) bináris fa
- belső csúcsok változókkal és ezekhez kapcsolódó feltételekkel címkézettek
- levelek a célváltozó valamely értékével címkézettek

Example of a Decision Tree

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

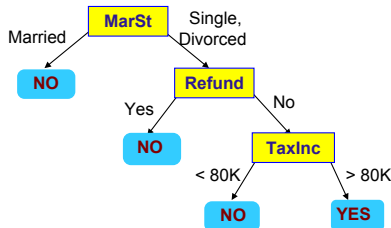
Training Data



Model: Decision Tree

Another Example of Decision Tree

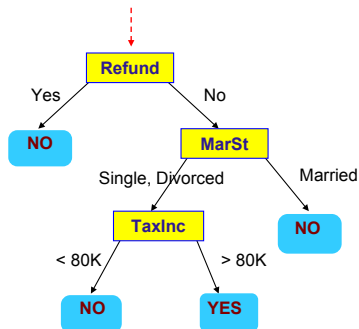
<i>Tid</i>	<i>Refund</i>	<i>Marital Status</i>	<i>Taxable Income</i>	<i>Cheat</i>
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



There could be more than one tree that fits the same data!

Apply Model to Test Data

Start from the root of tree.



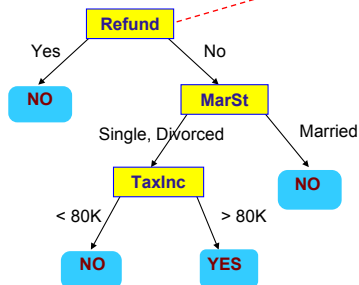
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Apply Model to Test Data

Test Data

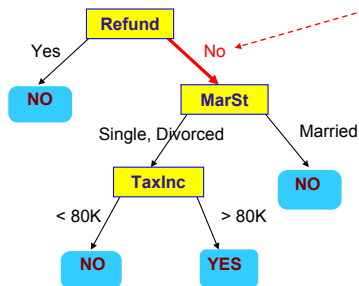
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

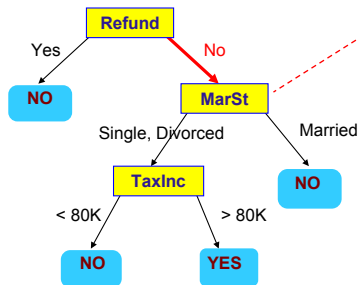
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

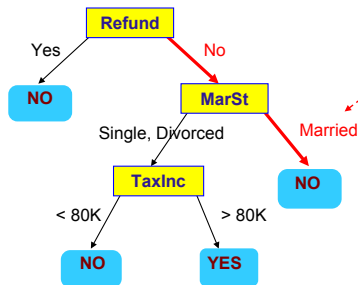
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

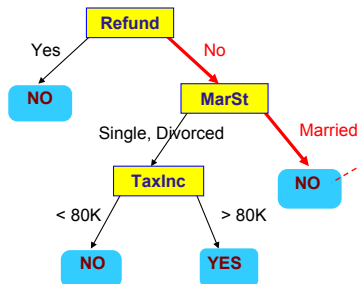
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Assign Cheat to "No"

Algoritmus döntési fa készítésére

- egy általános algot nézünk, ennek különböző verziói futnak különböző programokban
- nem a legjobb fát akarjuk megtalálni (ez amúgy is aluldefiniált fogalom), hanem egy elég jót
- mohó módon, lokális döntéseket hozva, gyorsan

Hunt algoritmus, vázlat

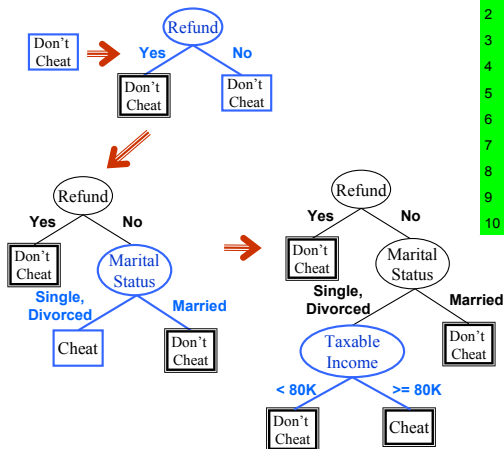
- elején egy csúcs, ide tartozik minden rekord, címke a többségi címke
- később: választunk egy csúcsot, amit érdemes lenne szétvágni és valami attribútum mentén szétvágjuk egy vagy több részre
- vége: ha már nem érdemes vágni sehol

Hunt algo, kérdések

- mikor érdemes vágni?
- melyik csúcsot vágjuk, ha több lehetőség is van?
- hogyan vágunk?
- az új csúcsokat hogyan címkézzük?
- mikor van vége?

Hunt's Algorithm

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Hunt algo, kérdések

- Mikor van vége?
 - Ha már nincs olyan csúcs, amit vágni érdemes.
- Mikor nem érdemes vágni?
 - Ha nem akarunk tovább osztani: minden rekord azonos cél-címkéjű
 - Ha nem tudunk tovább osztani: olyan sorok vannak különböző címkével, amiknek minden más attribútuma megegyezik
- Melyik csúcst vágjuk, ha több lehetőség is van?
 - Valami bejárás szerint, pl. szélességi, mélységi.
- Hogyan vágunk?
 - Erről mindárt, ez érdekes.
- Az új csúcsokat hogyan címkézzük?
 - Többségi címkével.

Milyen attribútum mentén és hogyan vágunk?

Fő elv: sokféle vágást kipróbálunk és a legjobb szerint vágunk.

- Mik a lehetséges vágások? (Függ a szóba jövő attribútumok típusától.)
- Mi egy vágás jóságának a mértéke? Hogyan mérjük ezt?

Lehetséges vágások az attribútum fajtája szerint

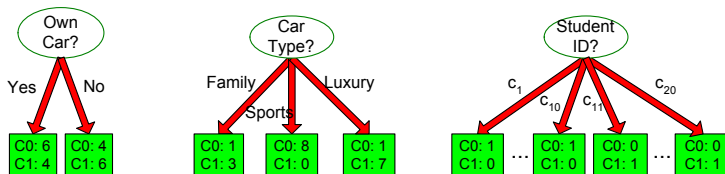
- bináris attribútum: igen vagy nem, két gyerek csúcs lesz
- kategória típusú attribútum:
 - multiway split: minden lehetséges értékhez egy gyerek, az üres csúcsok címkéje a szülő többségi címkéje lesz
 - bináris split részalmaz szerint: ebből van $2^t - 1$, ahol a t a lehetséges kimenetek száma
 - bináris vágás egy érték szerint: ebből van t darab (mint marital status az előző példában)
- folytonos attribútum
 - bináris vágás, az attribútum értéke kisebb-e egy adott küszöbnél (mint income az előző példában)
 - többes vágás: melyik sávba esik az érték

Vágás jósága, alapelvek

- többféle mérőszám van, mindegyik egy számértéket rendel a vágáshoz
- így a különböző vágások összehasonlíthatóak
- nagyrészt konzisztensek egymással a különféle mérőszámok
- mindegyik azt méri, hogy mennyire lesz homogén a létrejövő gyerek populáció a célváltozó címkézése szerint

How to determine the Best Split

Before Splitting: 10 records of class 0,
10 records of class 1



Which test condition is the best?

Vágás jósága

- 3 fő mérőszámot nézünk
- fő elv ugyanaz mindegyiknél:
 - egy rekordhalmazra definiálunk egy mérőszámot, ami az adott rekordhalmaz diverzitását mutatja (ebből lesz három féle mérőszám)
 - egy vágás jóságát azzal mérjük, hogy a szülőcsúcs diverzitása és a létrejövő gyerekcsúcsok diverzitása mennyire tér (mennyit nyerünk azon, ha szétvágjuk a szülőt egy adott módon, mennyivel lesznek homogénebbek a gyerekek)

először azt nézzük, hogy hogyan lehet mérni egy rekordhalmaz, azaz a döntési fa egy csúcsának imhomogenitását

Inhomogenitás mérése: Gini-index, entrópia, classification error

- van egy t csúcsunk (egy rekordhalmaz), aminek egy c darab lehetséges értéket felvevő célváltozó szerinti homogenitását akarjuk mérni
- p_i jelöli a rekordhalmazban előforduló i értékű rekordok relatív gyakoriságát
- $$\text{Gini} = 1 - \sum_{i=1}^c p_i^2$$
- $$\text{entrópia} = - \sum_{i=1}^c p_i \log p_i$$
- $$\text{classification error} = 1 - \max_{i \in \{1, \dots, c\}} p_i$$

Comparison among Splitting Criteria

For a 2-class problem:

