Adatbányászat – Osztályozás
Ellenőrző kérdések

Mit értünk osztályozás alatt? Mi az osztályozás két fő célkitűzése?
Osztályozás kategorikus célváltozók értékének előrejelzését teszi lehetővé.

X: magyarázó független változó
Y: kimeneti, függő változó

Célunk tipikusan kettős.

(a) A D_N megfigyelések alapján szeretnénk egy jövőbeli X megfigyelésnél Y értékét megjósolni.

(b) A D_N megfigyelések alapján szeretnénk megérteni, hogy egy adott $\{X_{i_1}, \ldots, X_{i_K}\}$ bemeneti változó halmaznak mi a szerepe Y érték megjóslásában.

Ha a Y (kimeneti, függő változó) diszkrét akkor osztályozásról beszélünk

Mit nevezünk döntési fának (kimenet, bemenet, csomópont, címke, élek)?
Wikipedia: A **döntési fa** egy olyan, a döntéshozatalban használt grafikus modell, amit az optimális tevékenység határoz meg olyan esetekben, amikor több választási lehetőség is rendelkezésre áll, és a kimeneteik bizonytalanok.

doksiban: A döntési fák implicit módon egyetlen objektumra korlátozott állításokat fogalmaznak meg = ítélet logika.

döntési fa bemenete: egy tulajdonsághalmaz segítségével leírt objektum vagy szituáció
döntési fa kimenete: egy igen/nem döntés
belső csomópont: döntési helyzet, valamely tulajdonság tesztje
levél csomópont: az a logikai érték amelyet akkor kell kiadni, ha azt elértük
élek: a teszt lehetséges értékeivel címkézett alternatívák

Mi a logisztikus regresszió lényege? Mi az úgynevezett logisztikus szigmoid függvény?
A logisztikus regresszió egy eszköz az Y kimeneti változó függésének vizsgálatára az X (mátrix) bementi változóktól.

Logisztikus szigmoid függvény: $\sigma(x) = 1 / (1 + e^{-x})$ Ez egy jó cucc, mert deriválható.

$$P(y|\underline{x}) = \sigma(\sum_{i=0}^{n} \beta_i x_i),$$

Milyen alapvető feltételezésre épül a Naiv Bayes-háló? Adjon példát egy Naiv Bayes-hálóra, definiálja részeit!

Cél a P(Y,X1...Xn) eloszlás modellezése
feltételezzük hogy X1...Xn megfigyelések függetlenek

Y egy hipotézis, ezt keressük X1...Xn "fényében" (?jobb szó?)
P(Y|X1...Xn) számolható méghozzá a Bayes tétel alapján átforgatjuk:

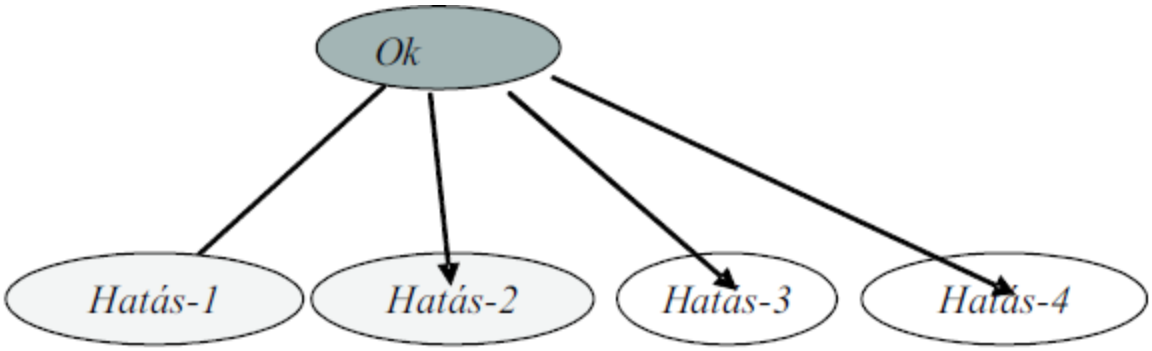
P(Y|X1...Xn)=P(X1...Xn|Y)*P(Y) / P(X1….Xn)

feltételezzük X1...Xn függetlenek:
P(Y|X1...Xn)=P(X1|Y)*...P(Xn|Y)*P(Y) / P(X1….Xn)

Végül a P(X1...Xn) "csak" egy valamilyen "konstans" (ami az eredményt leosztja 0..1 be)

$$P(Y|X_{i_1},\ldots,X_{i_m}) \propto P(X_{i_1}|Y)\ldots P(X_{i_m}|Y)P(Y)$$

egy P(Xi1|Y) az a következő ábrán az "Ok" és a "Hatás-1" függésének valószínűsége:



Fent az ok (hipotézis), lenn az okozatok (hatások, szimptómák), az okozatok között nincsennek kapcsolatok. A nyilak által jelzett függéseket kell valószínűségek megadásával definiálni.

Milyen hibatípusok lehetségesek bináris döntések esetén?
Hamis pozitív (False Positive): azt mondja a doki hogy ki kell cserélni az agyad mer hibás és közben meg nem is

Hamis negatív (False Negative) : azt mondja a doki hogy nem ki kell cserélni az agyad mer tök jó, közben meg de is

(TP és TN nem hiba szerintem, de valaki erősítse meg) Sztem sem.

Miért szükséges egy döntési függvény „jóságát” vizsgálni?
Mert az olyan döntési függvények amelyek jellemzően hülyeséget adnak vissza, nem érnek egy fabatkát sem. Illetve a jóság érték alapján tudjuk összehasonlítani a különböző döntési függvényeket.

Hogy kapcsolódik ehhez a hasznosság/veszteség mátrix?
A tanulás folyamatában és az eredményének kiértékelésében is szükség van egy olyan minősítő/jellemző eljárásra, ami egy általános f függvény "jóságát" a célfüggvényhez f0-hoz megadja. Általános diszkrét, de nem bináris esetben egy hasznosság/veszteség mátrix megadása teszi ezt lehetővé.

Mit jelent a…

-Sensitivity (érzékenység)?
TP/(TP+FN) csak a “beteg” alpopulációkon belül definiált
~mennyit fedezett fel a valós betegek közül

-Specificity (specifikusság)?
TN/(TN+FP) csak a “nem beteg” alpopulációkon belül definiált
~mennyit fedezett fel a nem betegek közül

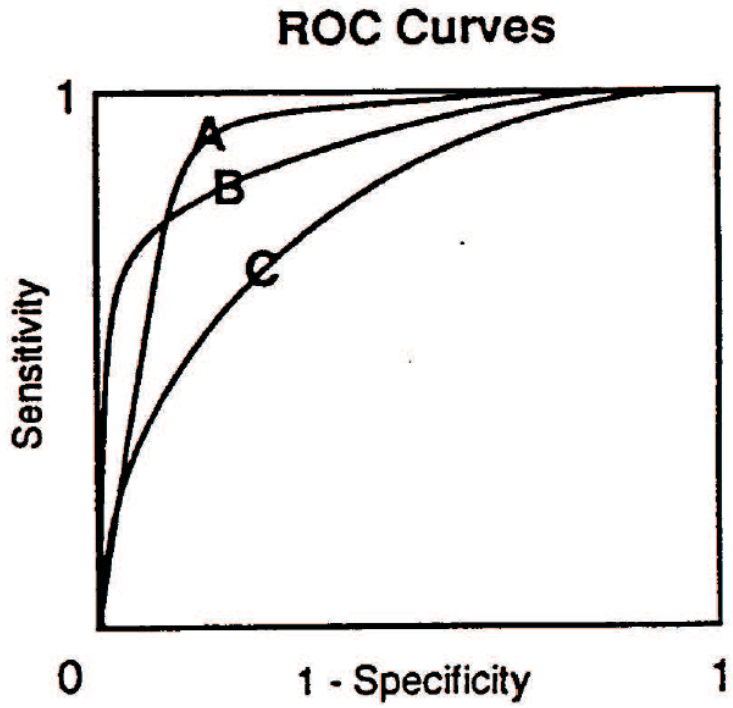
Sensitivity, Specificity nem függ a “beteg”/”nem beteg” aránytól, ezért ez a két mutató játszik fő szerepet egy döntési függvény jellemzésekor. Egyszerűbben ezek az értékek csak a veszteségfüggvény “jóságát” jellemzik.

-PPV – positive predictive value (pozitív prediktív érték)?
TP/(TP+FP)
~mennyire trafálhat bele az igazságba

-NPV – negatív predictive value (negatív prediktív érték)?
TN/(TN+FN)
~mennyire trafálhat bele a hamisságba

-Missclassification rate (osztályozási hiba)?
(FP+FN)/(TP+FP+FN+TN)
~mekkora az osztályozási hiba

Mit értünk ROC görbe alatt?



Ez más megfogalmazásban a lehetséges döntési küszöbökhöz tartozó érzékenységet (sensitivity) és (1-specifikusság) párokat ábrázolja.

Minél jobban besimul a görbe a bal felső sarokba annál nagyobb a ROC görbe alatti terület (AUC)

Ismétlés:
Sensitivity= TP/(TP+FN) -> mennyit fedezett fel a valós betegek közül
Specificity= TN/(TN+FP) -> mennyit fedezett fel a nem betegek közül
1-Specificity= 1-TN/(TN+FP) -> mennyit NEM fedezett fel a nem betegek közül

Tehát ha a görbe besimul a bal felső sarokba akkor az azt jelenti bármely döntési küszöbnél a “beteg” és “nem beteg” felismerés az tökéletes, a gauss görbék között nincs átfedés

Ha a küszöböt leviszem a lehető legkisebb értékre akkor nő a szenzittivitás mert mindenkit betegnek sorolok

http://www.date.hu/acta-agraria/2002-01/fazekasne.pdf

Egy döntési fa tanulása során mit nevezünk splittingnek?
A döntési fa tanulás minden egyes lépésénél egy csomópontot bontunk fel egy változó értékei mentén két részre (ez az ún. splitting).

Egy döntési fa tanulása során mit értünk purity alatt?
A cél az, hogy a létrejött partíciók a célváltozó értéke szempontjából minél egységesebbek legyenek, ennek a mértéke a tisztaság (purity).

Mit jelent a pruning egy döntési fa tanulása során?
Depth: Meghatározza a döntési fa maximális mélységét. A fa építése folyamán, ha eléri a csomópont a maximális mélységet, akkor felbontása (splitting) nem folytatódik tovább. E paraméterrel a fa nyesését (pruning) szabályozhatjuk.

Mire szolgál a költségmátrix (cost matrix)?
Abban az esetben, ha a célváltozó értékei (pl.: 0,1,2) nem egyformán gyakoriak, tehát vannak olyan értékek, melyeket gyakran (0,1), míg másokat ritkán vesz fel (2), akkor ezt egy osztályozó készítésekor figyelembe kell venni. Ha ezt nem tennénk meg, akkor olyan modellt kapnánk, ami félreosztályozza a ritka eseteket (2), de mivel ez összességében kevés elemet érintene, az osztályozási hiba alacsony maradna. Ennek kiküszöbölésére szolgál a Cost Matrix.

Mit tartalmaz a konfúziós mátrix (confusion matrix)?
Az osztályozás helyességét mutatja be. Négy alapcellát tartalmaz: igaz pozitív, igaz negatív, hamis pozitív, hamis negatív. Segítségével egyszerűen meghatározható számos az osztályozást jellemző jósági paraméter.

Mit nevezünk klaszterezésnek?
Az osztályozás abban hasonlít a korábban megismert klaszterezéshez, hogy a rekordokat csoportokba rendezzük a változók értékei alapján. Azonban egy lényegi különbség a két módszer közt, hogy az osztályozásnál a kialakítandó osztályok rögzítettek.