Osztályozás, regresszió

Nagyméretű adathalmazok kezelése – Tatai Márton

Osztályozási algoritmusok

- Osztályozás
  - Diszkrét értékkészletű, ismeretlen attribútumok értékének meghatározása ismert attribútumok értéke alapján
  - Egy megfigyelt entitás egy osztályba sorolása előző megfigyelések alapján
  - Általában két fázisban épül fel az algoritmus:
    - Modell készítése tanító pontok felhasználásával
    - Modell alkalmazása új adatokra (ismeretlen attribútumokkal)
  - Más szóval felügyelt tanulás (supervised learning)
    - A felügyelet nélküli változata a Klaszterezés

Alkalmazási területek

- Adatbányászat ☺
- Hitel alkalmasság elbírálása (vagy veszélyességi szint meghatározása)
- Viselkedés előrejelzése (megvesz / nem vesz meg)
- Szöveg elemzése
  - E-mail forgalom (spam felismerés)
- …

Osztályozó algoritmusok osztályozása

- Eager
  - Folyamatosan építi a modellt rendelkezésére álló adatpontok alapján
- Lazy
  - Csak azután dolgozik, miután megkapta az osztályozandó adatot
- Hogyan mérjük a pontosságot?
  - A tanító adatok mellett készítsünk elő egy teszt adatsort is, amin ellenőrizhetjük a modellt

Eager (buzgó) algoritmusok

- Szabály alapú osztályozás
- Döntési fa
  - Egy fa, csúcsaiban szabályokkal, levelei osztályokat reprezentálnak
- Naív Bayes módszer
- Bayes-hálózatok
  - Statisztikai módszerek, megadják, hogy mi az esélye annak, hogy egy megfigyelt entitás az adott osztályba tartozik
- Support Vector Machine
  - Az adatokat elhelyezi az n-dimenziós térben, majd megkeresi az osztályokat határoló hipersíkot

k legközelebbi szomszéd

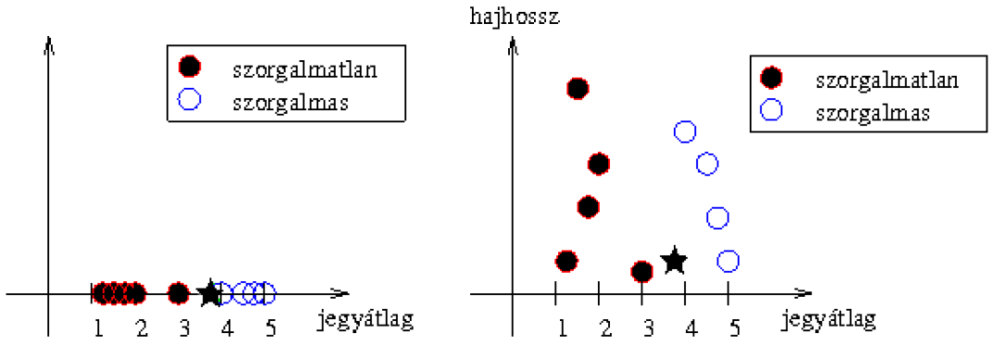
- Lusta algoritmus
- Egy osztályozandó entitáshoz megkeresi a k leghasonlóbb ismert entitást
- Mi alapján?
  - Távolság. Pl.: Euklideszi, Manhattan
- Az új entitás abba az osztályba fog tartozni, ami leggyakoribb a közeli szomszédjai között
- Fontos az adatok előfeldolgozása!
  - A forintban vett jövedelem jóval nagyobb, mint a magasság
  - Vagy súlyozhatjuk az egyes tulajdonságokat

k legközelebbi szomszéd (folyt.)

- Mi a helyzet a kategorikus attribútumokkal?
  - Ha egyezik, legyen 1, egébként 0
- Sok ismert entitás esetén tetszőleges ponthoz közel lesznek szomszédai, így egyre jobb becslést kapunk.
- Dimenzióátok!
  - Emlékezz: A dimenzió növelésével drasztikusan csökken a pontok sűrűsége
- Ráadásul n ismert entitás mellett O(n) az algoritmus, tehát több ismerettel egyre lassabb
  - A feladat jól párhuzamosítható
  - A keresési tér felosztása (KD-fa)

k legközelebbi szomszéd (folyt.)

- Érzékeny az irreleváns attribútumokra



- Megoldást jelenthet
  - Területi szakértelem
  - Statisztikai tesztek

k legközelebbi szomszéd (folyt.)

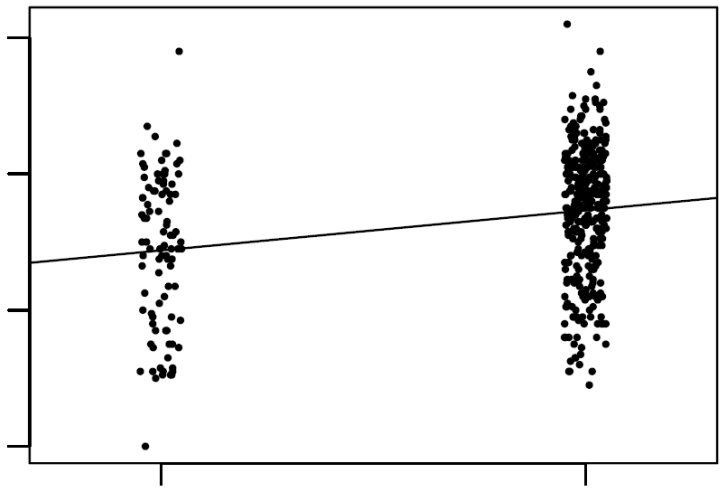
- Osztályozás helyett használhatjuk egy ismeretlen, folytonos attribútum értékének megbecslésére is
  - Nem a leggyakoribb attribútum értéket rendeljük hozzá, hanem a k legközelebbi szomszéd attribútumainak átlagát

Regressziós analízis

- Statisztikai módszer ismert és ismeretlen változók kapcsolatának megismerésére, előzetes megfigyelések alapján.
- Alkalmas folytonos változók értékének megjóslására, de használható osztályozásra is.
- Általában nem illeszkedik pontosan a megfigyelt adatokra, hanem egy olyan modellt próbál felállítani, ami valamilyen értelemben a legjobb a megfigyelések alapján.

Lineáris regresszió – egyszerű lineáris modell

- Gyermekek teszteredménye, az édesanyjuk iskolai végzettségének függvényében



$$\text{kid.score} = 78 + 12 \cdot \text{mom.hs} + \text{error},$$

- A mom.hs egy indikátor változó

Lineáris regresszió – egyszerű lineáris modell

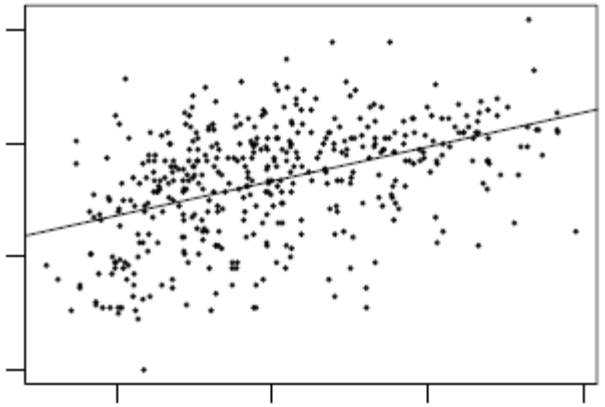
- A lineáris regresszió más megközelítésben:
  - Hogyan változik a kimeneti változó átlaga a bemeneti változó függvényében
  - A bemeneti változó a megfigyelt pontok egy halmazát jelöli ki, akikhez a kimeneti változónak egy értéke tartozik
- A regressziós egyenes átmegy mind a két populáció (gyerekek teszteredményei, akiknek édesanyja végzett középiskolát, és akiknek nem) átlagán.
- Ezt a modellt becslésre használva azt kapjuk, hogy egy gyermek teszteredménye átlagosan
  - 78, ha édesanyjuk nem végzett középiskolát
  - 91, ha elvégezte a középiskolát

Lineáris regresszió – egyszerű lineáris modell

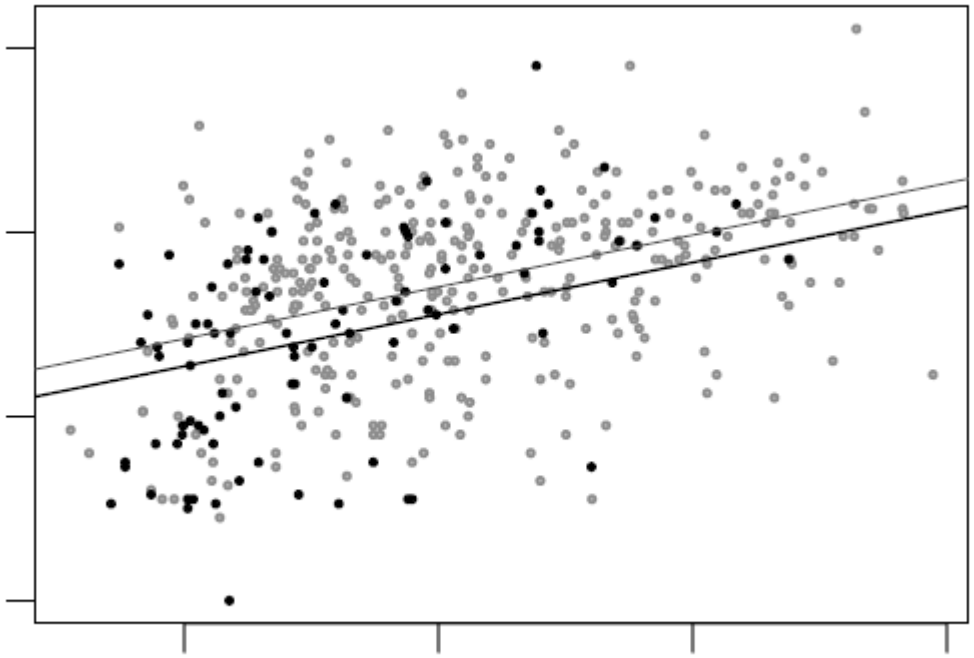
- Folytonos bemeneti változóval

$$\text{kid.score} = 26 + 0.6 \cdot \text{mom.iq} + \text{error},$$

- Azon gyermekek teszteredményének átlaga, akiknek szülői IQ-ja 1 ponttal különbözik 0.6 ponttal tér el
- A modell megadja azon teszteredményeket is, akikhez 0,vagy negatív IQ-jú szülő tartozik – ez nem túl hasznos



Lineáris regresszió – több bemeneti változó



- A sötét foltok azon gyermekekhez tartoznak, akiknek anyja nem végzett középiskolát

Lineáris regresszió – több bemeneti változó

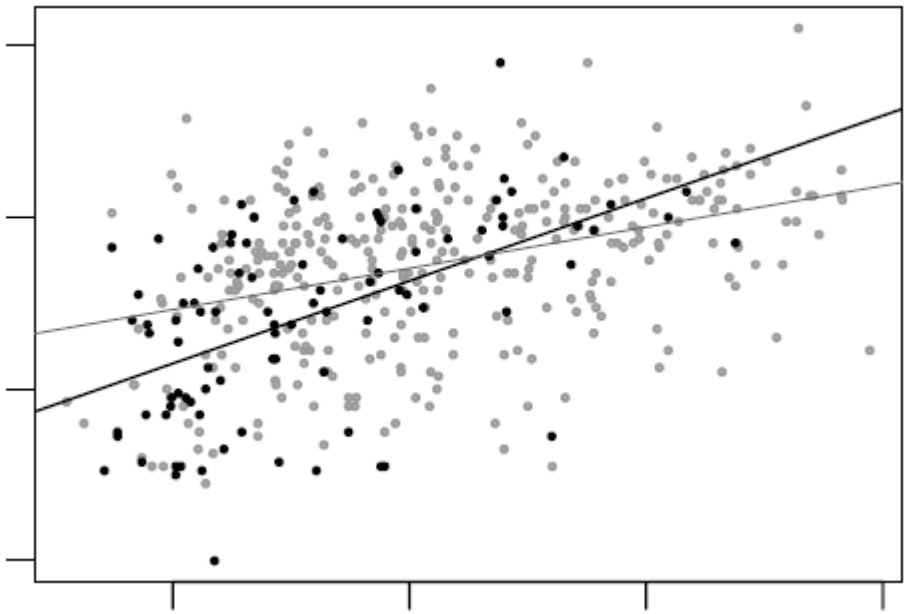
$$\text{kid.score} = 26 + 6 \cdot \text{mom.hs} + 0.6 \cdot \text{mom.iq} + \text{error},$$

- Próbáljuk meg az előző két számot kombinálni
- Ha az anyuka középiskolát végzett, várhatóan 6 ponttal magasabb lesz az átlagos teszteredmény
- Probléma: a két regressziós egyenes meredeksége ugyanolyan, pedig lehet, hogy a megfigyelésünk nem ezt sugallja

Lineáris regresszió – bemenetek közötti kölcsönhatások

- Úgy gondoljuk, hogy az édesanya végzettsége és IQ-ja összefüggésben van, igazítsuk ehhez a modellt. Vegyük fel a következő változót: mom.hs * mom.iq

$$\text{kid.score} = -11 + 51 \cdot \text{mom.hs} + 1.1 \cdot \text{mom.iq} - 0.5 \cdot \text{mom.hs} \cdot \text{mom.iq} + \text{error}$$

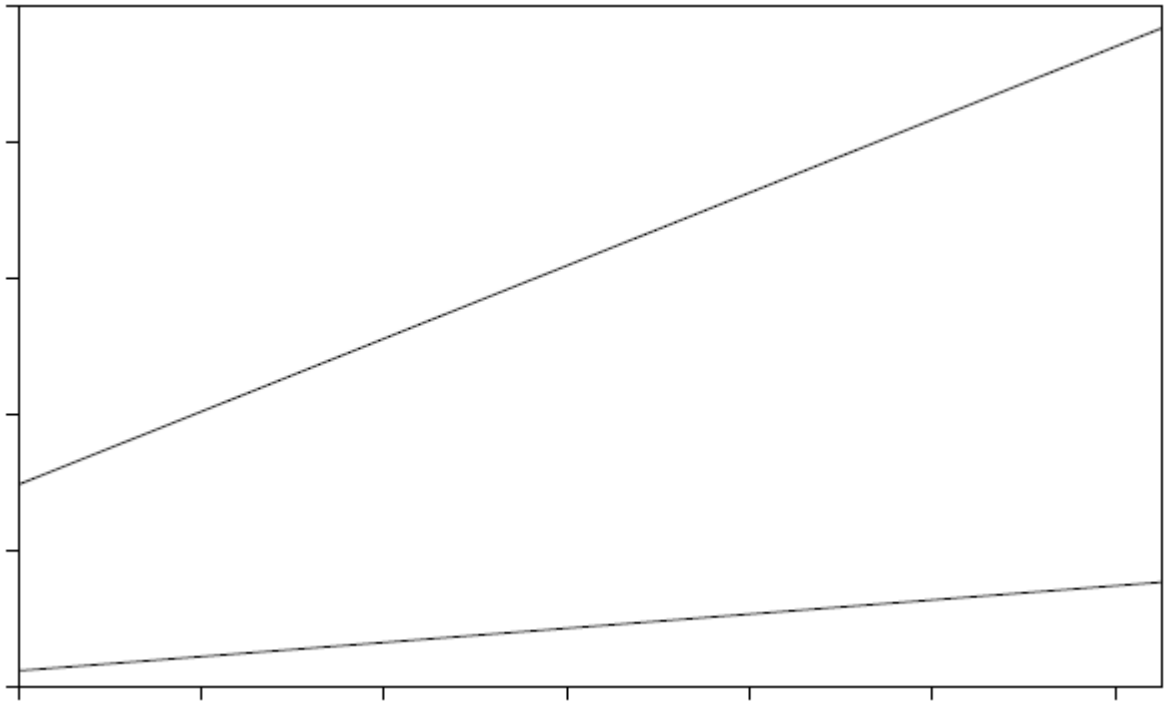


Lineáris regresszió – bemenetek közötti kölcsönhatások

- A modell továbbra is lineáris. Három különböző változónk van
  - mom.hs
  - mom.iq
  - mom.hs * mom.iq
    - Ez a változó megváltoztatja az egyenes meredekségét attól függően, hogy az édesanya végzett-e középiskolát
  - (Igazából négy változónk van, a konstans taghoz is tartozik egy, aminek értéke 1)
- A kölcsönhatások nagyon fontosak tudnak lenni

Lineáris regresszió – bemenetek közötti kölcsönhatások

- Az otthon közelében található radon források hatása tüdőrák esélyére dohányzók, illetve nemdohányzók esetében



Jelölések

$$\begin{aligned} y_i &= X_i\beta + \epsilon_i \\ &= \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad \text{for } i = 1, \ldots, n, \end{aligned}$$

- y:kimeneti változó X: bemeneti változó, B: paraméter
- k különböző bemeneti változónk van, i a megfigyeléseket indexeli
- ϵ_i: Az i-ik ponthoz adódó véletlen hiba
  - Feltételezzük, hogy a hiba eloszlása normális, 0 várható értékkel és σ szórással
- $\hat{\beta}$: A becslés (regressziós egyenes) együtthatói

Regressziós egyenes számítása

- Célunk az eltérések(residuals) négyzetösszegét minimalizálni

$$\sum_{i=1}^{n}(y_i - x_i^T b)^2 = (y - Xb)^T(y - Xb),$$

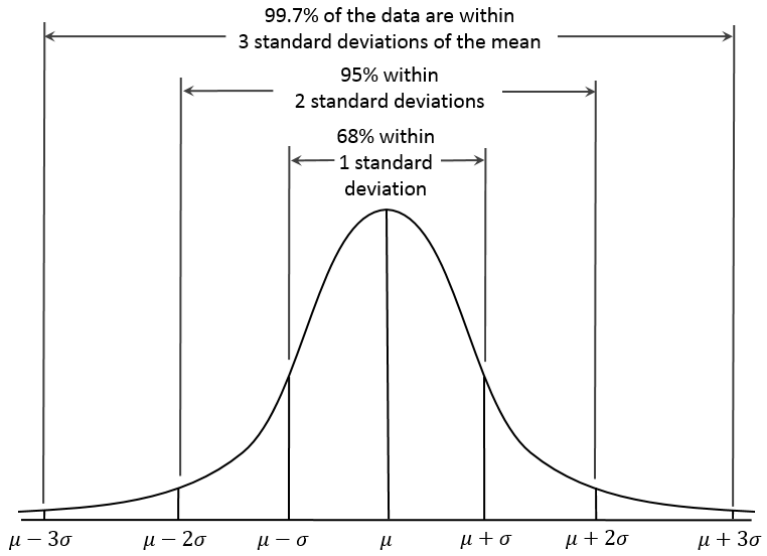
- Ezt b szerint deriválva, majd 0-val egyenlővé téve megkapjuk, hogy

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- Ez az „algoritmikus” megközelítés. Mivel a modellünk lineáris, és a feltételezett hiba standard eloszlású ($0, \sigma$ paraméterekkel). Ezt a becslést le lehet vezetni a maximum likelihood összefüggésből is.
- Feltételezzük, hogy több megfigyelésünk van, mint bemeneti paraméterünk

A lineáris regresszió jellemző számai

- A modellben található véletlen hiba miatt a becslés is bizonytalan
- Minden paraméterhez tartozik egy standard hiba. Az mondható, hogy ennek kétszeresén belül levő értékek konzisztensek a megfigyelésekkel



A lineáris regresszió jellemző számai

- Eltérések szórása

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{n} r_i^2/(n-k)},$$

- A gyermekek teszteredményei esetén ez például 18, ami azt jelenti, hogy nagyjából 18 pontnyi pontossággal tudjuk előre jelezni az eredményeket az adatok alapján. Minél kisebb ez, annál jobban illeszkedik modellünk a megfigyelésekre. Ezt hívhatjuk a modellünk által „megmagyarázott” szórásnak.
- A megfigyeléseinkben található összes szórás - s. $R^2 = 1 - \hat{\sigma}^2/s_y^2$
- R^2 tehát a modellünk által megmagyarázott és az összes szórás aránya. Látható, hogy ez annál jobb, minél nagyobb, mivel a nagyobb érték azt jelenti, hogy a szórás egy nagyobb részét sikerült megmagyaráznunk.

A lineáris regresszió jellemző számai

- A teszteredményes példában $R^2 = 22\%$
- Azonban nagyobb R^2 nem mindig eredményez jobb modellt.
- Demo:
  - http://www.arachnoid.com/polysolve/

A regressziós modell feltételezései

- Helyesség
  - Legfontosabb, hogy az adat helyes legyen, a szóban forgó kutatás keretén belül reprezentatívnak kell lennie.
  - Az összes releváns bemenetet érdemes felhasználni modellben, és ellenőrizni kell, hogy megfelel-e az elvárásoknak
  - Helyes következtetések levonása: Gyermekek teszteredményei nem feltétlenül tükrözik az intelligenciájukat
- Linearitás
  - A modell a bemeneti változók egy lineáris függvénye
  - Amennyiben a linearitás sérül, érdemes megpróbálkozni a változók transzformációjával

A regressziós modell feltételezései

- A hibák szórása
  - Egymástól független
  - Egyenlő nagyságú
  - Ezt meg lehet vizsgálni, ha az eltéréseket ábrázoljuk a bemeneti változó függvényében
- A legjobb ellenőrzés, ha egy területi szakértő ellenőrzi a modellünket, vagy további megfigyelésekkel validálni tudjuk.

Bemenetek transzformációja

- Amennyiben az adatok nem illeszkednek az előbb említett feltételek mellett a megfigyeléseinkre, megoldást nyújthat a változók transzformációja, új változók felvétele
- Felvehetjük x mellé/helyett x^2-et, ezáltal „U” alakra illeszkedő adatokat is modellezhetünk
- Ha a bemenet kis változása a kimenet egyre nagyobb változását eredményezi, lehet, hogy kifejezőbb modellt kapunk, ha vesszük a kimenet logaritmusát

$$\log y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i$$

$$y_i = e^{b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i}$$

Logisztikus regresszió

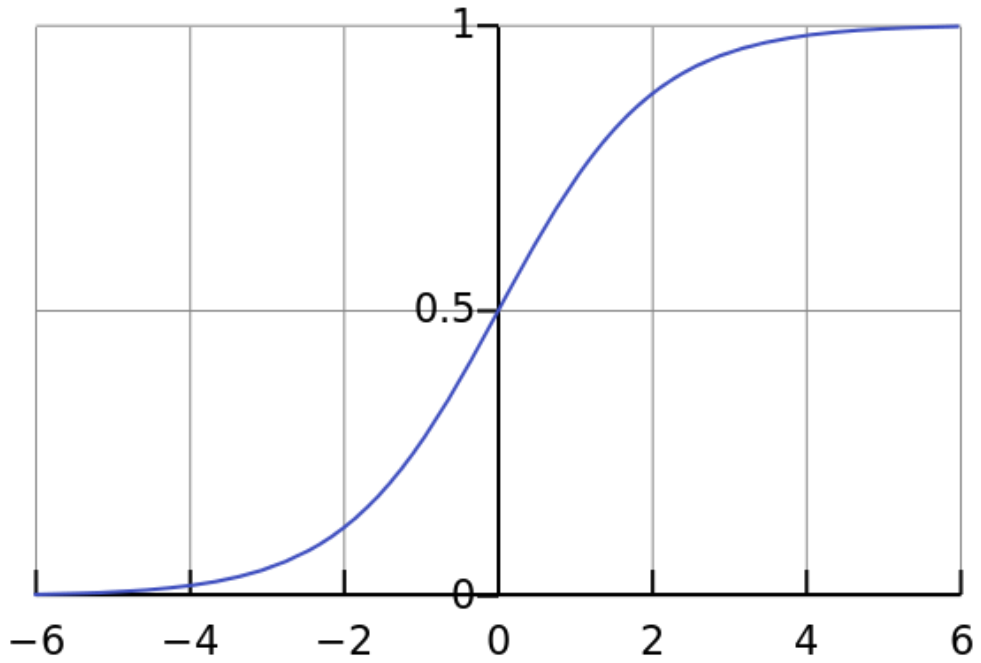
- Bináris adatok modellezésére általában a regressziónak egy másik formáját használják, a logisztikus regressziót
- A logisztikus regresszióval tehát lényegében valószínűségeket jelzünk előre

$$\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta),$$

- A logit függvény a $(0,1)$ intervallumot képzi le a $(-\infty, \infty)$ intervallumra, inverze pedig folytonos értékeket képez a másik irányba, ezáltal egy valószínűséget kapunk.

$$\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$$

Logisztikus regresszió



Logisztikus regresszió

- Látható, hogy a függvény görbe, tehát fix értékhez változó növekedés tartozik
- A magasabb valószínűségeknél egyre nagyobb „befektetés” szükséges változás eléréséhez
- Lényegében a skála elején és végén a változások egyre mérsékeltebbek, így lehet az adott intervallumban tartani az értékeket

Logisztikus regresszió

- Példa:
- 1992-es választások az USA-ban. A kimeneti változó 1, ha a választó Bush-t (republikánus) preferálta, 0, ha Clinton-t (demokrata)
- A hipotézis, hogy a gazdagabb emberek nagyobb valószínűséggel szavaznak Bush-ra
- A választókat 5 bevételi kategóriára osztották (átlagosan 3.1)
- A kiszámított logisztikus regressziós modell

$$\Pr(y_i = 1) = \text{logit}^{-1}(-1.40 + 0.33 \cdot \text{income}).$$

Logisztikus regresszió

- A bemeneti változó értelmezése:
  - Egy kereseti kategória növekedése mekkora növekedést hoz a Bush-ra való szavazás esélyében
  - Függ attól, hogy melyik kategóriából indultunk
- A változókat a lineáris regresszióhoz hasonlóan kombinálhatjuk

One versus all

- A logisztikus regressziót fel lehet használni kategória típusú kimeneti változó modellezésére is
- Vesz egy osztályt, és „szembe állítja” az összes többivel egy logisztikus regresszió erejéig (adott bemeneti értékek mellett)
- Amelyik osztálynál legnagyobb a valószínűség, azt adja eredményül
- Ez a módszer nem csak a logisztikus regresszióval használható, hanem minden bináris kimenetű osztályozó algoritmussal

Általánosított lineáris regressziós modell

- A két bemutatott regressziós modell az általánosított modell speciális esetei
- Az általános modell elemei
  - Megfigyelések
  - Bemeneti változók és paraméterek, amelyek egy lineáris komponenst állítanak elő
  - Egy kapcsolati függvény(link function), amely a lineáris komponenst képezi le valamilyen módon

$$\hat{y} = g^{-1}(X\beta)$$

  - Véletlen komponens
    - Meghatározza, hogy hogyan kapjuk a hiba értékét
  - Egyéb paraméterek (szórások, intervallumok, …)

Általánosított lineáris regressziós modell

- A paraméterek becslésének egyik módja a maximum likelihood módszere
- A lineáris regressziós modell az általános egy speciális esete
  - A kapcsolati függvény az identitás függvény
  - A véletlen komponens a normális eloszlást követi
- A logisztikus regressziós modell is egy speciális eset
  - Kapcsolati függvény: logit
  - Véletlen komponens: binomiális

Poisson regressziós modell

- Az általánosított lineáris modell leszármazottja
- Számláló (gyakoriság) jellegű kimeneti változók becslésére használják
- A kapcsolati függvény a logaritmus. Ez leképzi a lineáris komponenst a pozitív valós számok halmazára
- A véletlen komponens poisson

Áttekintett módszerek/algoritmusok

- Osztályozás
  - k legközelebbi szomszéd módszere
- Regresszió
  - Egyszerű lineáris
  - Többváltozós lineáris
  - Bemenetek egymásra hatása
  - Bemenetek transzformálása
  - Általánosított lineáris modell
  - Logisztikus regresszió
    - One vs. All