Modellek visszamérésének hatékonysága

Dimenziócsökkentés

Hogyan lehet csökkenteni dimenziókat, úgy, hogy ne veszítsünk sok információt?

Eddig feature selection-t csináltunk, vagyis dimenziókat választottunk ki, de a dimenziócsökkentés nem ezt célozza, hanem az összes szám csökkentését.

PCA

Principal Component Analysis, Főkomponens analízis

- Két dimenzióban fogjuk elemezni: x_1, x_2.
- Ezekhez meghatározzuk a varianciákat: $\omega_{x_1}, \omega_{x_2}$.

Cél

1. Az új változók lehetőleg minél jobban leírják az adathalmazunk varianciáját.
2. Az új dimenziók ortogonálisak ('merőlegesek') legyenek egymással.
3. Ugyanannyi dimenziót szeretnék létrehozni, mint amennyi volt.

Dimenziók

1. Egy jó dimenziónak (PC_1) a lehető legnagyobb varianciája van (kb. mint egy átló): $\omega_{PC_1} > \omega_{x_1} > \omega_{x_2}$
2. A következő dimenzió merőleges az elsőre.

Az új dimenziók az eredetiek lineáris transzormációi lesznek:

1. dimenzió: $a_1x_1 + b_1x_2$
2. dimenzió: $a_2x_1 + b_2x_2$

Az újabb komponenesekkel egyre kevesebb varianciát tudunk befogni. Addig adjuk hozzá

az új dimenziókat, amíg meg nem magyarázzuk a teljes varianciát, vagy annak egy általunk meghatározott részét.

Kombinálhatóság más modellekkel

PCA segítségével át lehet formázni az adathalmazt, úgy, hogy jobban kedvezzen egy modelhez.

1. Vizualizációhoz tudjuk használni amit pl klaszterezésnél segíthet minket.
2. A PCA dimenzióhoz rendelem hozzá a korábban már definiált klasztereket (?).

Probléma a PCA-va

Nem tudjuk az új változók jelentését értelmezni.

Visszamérési módszerek

Most csak a felügyelt tanulómodellek visszamérési módszereit nézzük (ezek könnyebben számíthatóak).

Predikció típusai

1. **Döntés** osztályozás sikeressége: hány darabot sikerült eltalálni?
2. **Sorbarendezés** osztályozási konfidenciák alapján.
3. **Becslés**, regresszió sikeressége: mennyire sikerült jól eltalálni?

Döntés pontossága

Logisztikus regresszió: három új változót ad az adathalmazhoz

1. Prediktált érték (melyik kategóriába esik)
2. Adott döntéshez tartozó konfidencia (mennyire valószínű, hogy oda fog esni, ahova ő jósolta)
3. Többi osztály (?)

Logisztikus regresszió teljesítménymérése

- MSE: négyzetes hiba átlaga

$$\frac{\sum_i^n (y_i - f_i)^2}{n}$$

- RMSE: MSE négyzetgyöke
- MAE: abszolút hiba

$$\frac{\sum_i^n |y_i - f_i|}{n}$$

Becslés pontossága

R^2

$SS_{ERR} = \sum(y_i - f_i)^2$, pontoknak az egyenstől vett távolságnégyzeteinek összege.

$SS_{TOT} = \sum(y_i - \bar{y}_i)^2$, a pontoknak az átlaguktól való eltéréseik négyzetösszege.

$$R^2 = 1 - \frac{SS_{ERR}}{SS_{TOT}}$$

Az egyenesnek az átlaghoz képes viszonyított magyarázási jóságát mutatja meg.

Ezt azért is szeretik használni, mivel egy fix [0, 1] értékkészleten számolják mindig.

2. Döntés pontossága

Konfidencia alapú sorbarendezhető jelentősége:

1. Csak a nagyon valószínű előrejelzéseket szeretnénk használni
2. Máshogy viselkedünk a bizonytalanabb esetekkel

Ezt a ROC görbével fogjuk elemezni.

- A legbonyolultabb döntési model
- Bináris osztályozási probléma

Első lépés

Meghatározzuk, hogy melyik értéket próbáljuk elsősorban meghatározni

- **Primary outcome (P_0)**: az egyik értékre ez a fő kimenet (default: 1).

- **Secondary outcome (P_1)**: a másodlagos kimenet.

Tévesztési mátrix

Dimenziók

- Valós értékek
- Előrejelzési értékek

Esetek

1. True positive
2. True negative
3. False positive: Elsőfajú hiba
4. False negative: Másodfajú hiba

A fontos belátás, hogy nem egyenrangúak a hibák egymással (leginkább alkalmazási szempontból). Például, nem mindegy, hogy hogyan értelmezzük, hogy adjunk-e hitelt, vagy tényleg beteg-e valaki, stb.

A ROC görbe felépítése

Tréning adathalmaz:
Teszt adathalmaz: Az eredeti célváltozó mellé teszi a következőket:

1. Becsült érték
2. Confidenciát [0,1] a primary outcome-hoz (P_0)

Sorbarendezi a sorokat, az alapján, hogy mennyire biztosan tudjuk az **1-es célváltozót** megbecsülni. Ez akkor jó, ha a célváltozóban is előre kerülnek az 1-es célváltozó értékű esetek.

ROC görbe ábrázolása

Kumulatív függvényt csinálunk

Első lépés

Konfidencia értékcsoport: [0, 1]

1. Ezek közül kijelöljük a P_0-t 1 és 0.9 közötti konfidenciával megmagyarázó eseteket.

2. Ábra felrajzolása
    - X tengely: az összes **valós** P_1 közül hányan kerültek bele a fenti kritériumba
    - Y tengely: a **valós** P_0 hány százaléka került bele a fenti kritériumba

Várhatóan a 'bal felső' sarokban lesznek a függvények.

Következő lépés

A 0.9 és 0.8 PC1 konfidenciájú esetekkel is megcsináljuk. Majd a következő kategóriával, és így tovább.

Nem szigorúan, de monoton a görbe.

Speciális Esetek

A ROC görbe ezek alapján

- **Tökéletes** model: függőleges egyenes 0-ban
- **Véletlen** model: egyenes átló
- **'Lassan növekedő'** ROC görbe: van ilyen, ilyenkor kell invertálni a modelt, mert akkor P_1-re (0-ra) már működik.

AUC

AUC érték összefoglalja, hogy mennyire jó :

- tökéletes: 1
- véletlen: 0.5

Amikor a döntési modell rosz, de a sorbarendezés jó

- Logisztikus regresszió egy valószínűség felett rendeli hozzá a becsült változót az esethez.
- Mekkora valószínűséggel jelzi előre a P_0-t -et [de itt van valmi kavar, azzal kapcsolatban, hogy hogyan értelmezzük az egy dimenzióra vonatkozó előrejelzési koefficienst]

Nem kategóriákkal, hanem lépegetve vizalizálva a ROC görbét

1. ha 1-est jelez előre: fel lép
2. ha 0-ást jelez előre: jobbra lép

[vagy pont fordítva?]

Egy lépés nagysága pedig megyezik 1-esek, 0-ások viszonylagos mennyiségével.

GINI érték

GINI = 2 (AUC - 0,5)