Machine learning (ML) elmélet

Tartalomjegyzék

[Lesson 1 – PreProcessing data 3](#_Toc2363315)

[Lépések 3](#_Toc2363316)

[Dataset 3](#_Toc2363317)

[Missing data 3](#_Toc2363318)

[Categorical variable 3](#_Toc2363319)

[Training sets and testing sets 3](#_Toc2363320)

[Standardizáció és normalizálás 3](#_Toc2363321)

[Lesson 2 – Simple Linear Regression 4](#_Toc2363322)

[Képlet 4](#_Toc2363323)

[Közönséges legkisebb négyzetek: 4](#_Toc2363324)

[Lesson 3 – Multiple linear regression 5](#_Toc2363325)

[Képlet 5](#_Toc2363326)

[Feltételezések a lineáris regressziónál 5](#_Toc2363327)

[5 modell építő metódus 5](#_Toc2363328)

[All-in modell 5](#_Toc2363329)

[Backward elimination 5](#_Toc2363330)

[Forward selection 6](#_Toc2363331)

[Bidirectional Elimination 6](#_Toc2363332)

[Minden lehetséges modell 7](#_Toc2363333)

[Lesson 4 – Polynormial Linear regression 7](#_Toc2363334)

[Képlet 7](#_Toc2363335)

[Miért lineáris? 7](#_Toc2363336)

[lesson 5 - Support vector regression 8](#_Toc2363337)

[Fogalmak 8](#_Toc2363338)

[Miért SVR? Mi a fő különbség az SVR és az egyszerű regressziós modell között? 9](#_Toc2363339)

# Lesson 1 – PreProcessing data

## Lépések

Rövid leírások a témakörben található fogalmakról

### Dataset

Elemzendő adathalmaz

Tulajdonságai:

* tartalma releváns
* egy vagy több összefüggő változó (dependant variable DV)
* további adatok, amikről feltételezzük, hogy függőségben vannak a függő változókkal

Cél: Ezen feltételezéseket bizonyítani, majd ezen függőségek alapján megjósolni a függő változók alakulását

### Missing data

Előfeltétel, hogy minden elemi adatsor teljes legyen.

Pl: Medián, Átlag, Leggyakoribb érték

### Categorical variable

A dataset egy része mérhető adatokból áll, másik része viszont kategorizálható adatokból

Cél: olyan számszerű adatokat rendelünk az adott változókhoz, amelyet ennél fogva képes megemészteni az ML model.

* IGAZ/HAMIS
* elemi IGAZ/HAMIS: Annyi oszlopra kell felbontani, amennyi lehetséges értéket felvehet. 🡪 A legtöbb ML algoritmus egy számszerű skála esetén súlyozza a magasabb értékeket és így eltorzulhat a feltételezés.

Például:

|  |  |
| --- | --- |
| Magyarország | 1 |
| Németország | 6 |

|  |  |  |
| --- | --- | --- |
| Magyarország | 1 | 0 |
| Németország | 0 | 1 |

VAGY

### Training sets and testing sets

A dataset-et fel kell bontani egy tanuló (train) és egy tesztelési (test) darabra.

### Standardizáció és normalizálás

Releváns lehet az értékek közti távolság pl.: életkor ~ bér

* Euklédeszi távolság: P1 és P2 pont euklédeszi távolsága

#### Standardizáció:

#### Normalizáció:

# Lesson 2 – Simple Linear Regression

## Képlet

Dependant Constant együttható independant

variable (DV) (coefficient) variable (IV)

Salary ($)

+10k

Constant

30k

1 year

Experience

### Közönséges legkisebb négyzetek:

# Lesson 3 – Multiple linear regression

## Képlet

## Feltételezések a lineáris regressziónál

1. Lineáris
2. Homoscedaticity: Ha minden változó a sorozatban vagy vektorban ugyanazzal a véges variációval (szórásnégyzettel) rendelkezik
3. Multivariate normality
4. Independence of errors: független a hibáktól
5. Lack of multicollinearity: tényezőváltozók lineáris függetlenségének hiányát értik[[1]](#footnote-3)

## 5 modell építő metódus

* All-in
* Backward elimination: visszalépéses elimináció

Stepwise regression

(lépcsőzetes regresszió)

* Forward selection: előrelépéses kiválasztás
* Bidirectional elimination: Kétirányú elimináció
* Score comparison:

### All-in modell

Az összes független változót bevonja a modellbe, és ezek együttes hatását vizsgálja.

### Backward elimination

Első körben minden változót beépít a modellbe, majd azokat eliminálja, melyek kivételével nem csökken szignifikánsan a modell magyarázó ereje. A nem szignifikáns hatásúakat kizárja a modellből.

Röviden a teljes attribútum halmazból indul, minden lépésnél a halmaz legrosszabb elemét hagyja el.

#### Lépések

1. Válasszon egy szignifikáns szintet a modellben való tartózkodáshoz. Pl.: SL =0,05
2. Illeszkedjen a modellre az összes lehetséges előre jelzővel
3. A legmagasabb P – értékű predikátumot kell figyelembe venni. Ha P > SL, akkor STEP

4, ha nem akkor FIN

1. Eltávolítjuk a predikátumot
2. Illeszkedjen a változó nélküli modellbe FIN: A modelled elkészült

### Forward selection

A legnagyobb standard regressziós koefficienssel[[2]](#footnote-4) bíró, legerősebb hatású változót építi be először, mielőtt a következőt beépítené, F-teszttel megvizsgálja a program, melyik az a következő változó, mely szignifikánsan növelhetné az R2–értékét, tehát a modell magyarázó erejét. Addig építi be az újabb változókat, amíg tovább már nem növelhető a megmagyarázott hányad. Csak a szignifikáns változókat építi be a modellbe. Az újabb és újabb változók egymás elől „happolják el” a magyarázó erőt.

Röviden az előrelépéses kiválasztás az üres attribútum halmazzal indul, majd minden lépésnél a legjobb attribútummal bővíti a részhalmazt.

#### Lépések

1. Válasszon egy szignifikáns szintet a modellben való tartózkodáshoz. Pl.: SL =0,05
2. Illeszkedjen minden egyszerű regressziós modellhez **y ~ xn**. Válassza ki a legalacsonyabb P- értéket.
3. Tartsa meg ezt a változót és illessze be az összes lehetséges modellt egy további előrejelző hozzáadásával
4. Tekintsük a prediktort a legalacsonyabb P-értéknek. Ha a P <SL, akkor STEP3-ra megy,

egyébként menjen a FIN

FIN: Tartsd meg az előző modellt

### Bidirectional Elimination

Az előző két módszert kombinálja. Elsőként a legnagyobb standard regressziós együtthatóval bíró változót viszi be a modellbe, majd fokozatosan a többi kisebb erejűt, miután bevitt egy változót, megvizsgálja, mely változó vehető ki az alakulóban lévő modellből úgy, hogy az R2 értékét szignifikánsan csökkentené a kivétel. A legjobb magyarázó változók kiválasztását célzó automatikus modellépítési módszer. Használata javasolt epidemiológiai vizsgálatokra és feltáró vizsgálatokra, adott változókra építhető alternatív hipotézisek kialakítására.

Röviden úgy kombináljuk össze a fentieket, hogy az aktuális halmazunkból elhagyjuk a legrosszabbat, és hozzávesszük a maradék attribútumok közül a legjobbat.

#### Lépések

1. Válasszon egy szignifikáns szintet a modellbe való belépéshez és tartózkodáshoz

Pl. SL\_ENTER = 0,05, SL\_STAY = 0,05

1. Végezze el a FORWARD SELECTION következő lépését:

(az új változónak meg kell adnia a P < SL\_ENTER értéket)

1. Végezze el a visszafelé történő eltávolítás összes lépését (a régi változónak P < SL\_STAY

maradnia kell)

1. Nincs új változó, és a régi változó nem léphet ki 🡪 FIN: A modell kész van

### Minden lehetséges modell

#### Lépések

1. Válasszon egy alkalmassági kritériumot
2. Minden lehetséges regressziós modell: 2n-1 összes kombináció
3. Válaszd ki a az egyiket a legjobb kritériummal.

Példa:

10 oszlop 1023 modellt jelent.

# Lesson 4 – Polynormial Linear regression

## Képlet

### Miért lineáris?

Mert csak a változókban nemlineáris.

**Előny:**

Az eddigi nemlineáris függvények közül a polinom adja a legkényelmesebben használható függvényformát

**Hátrány:**

Nehéz tárgyi értelmet adni a nemlineáris tagoknak és együtthatóiknak (négyzetes, köbös stb.).

Fokszám növelés!!!!!!!!

**Excel példa:**

Sorrend = 2 🡪 x2

|  |  |  |
| --- | --- | --- |
| Position | Level | Salary |
| Business Analyst | 1 | 45000 |
| Junior Consultant | 2 | 50000 |
| Senior Consultant | 3 | 60000 |
| Manager | 4 | 80000 |
| Country Manager | 5 | 110000 |
| Region Manager | 6 | 150000 |
| Partner | 7 | 200000 |
| Senior Partner | 8 | 300000 |
| C-level | 9 | 500000 |
| CEO | 10 | 1000000 |

Sorrend = 4 🡪 x4 (Python: degree, R: ^)

# lesson 5 - Support vector regression

Az SVR egy regressziós algoritmus, így az SVR-nél folytonos értékekkel dolgozhatunk osztályozás helyett, ami SVM.[[3]](#footnote-5)

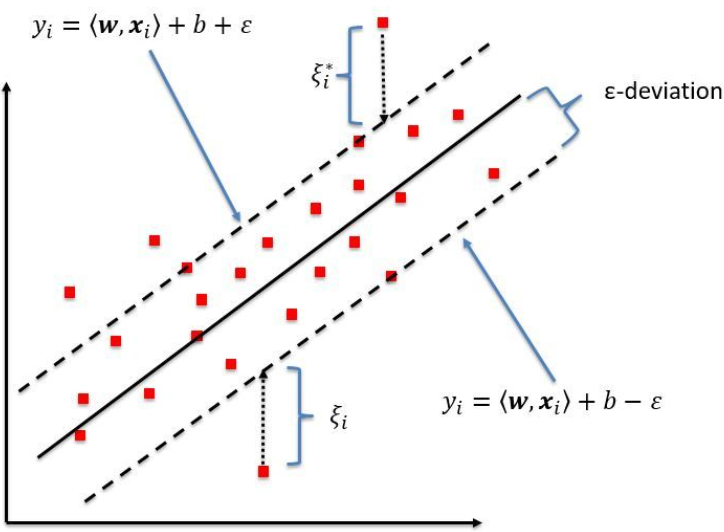
## Fogalmak

* Kernel: Az alsó dimenziós adatok magasabb dimenziós adatokba történő leképezésére használt funkció.
* Hyper Plane: Az SVM-ben ez alapvetően az adatosztályok közötti elválasztó vonal. Bár az SVR-ben úgy fogjuk definiálni, mint azt a sort, amely segít megjósolni a folyamatos értéket vagy a célértéket
* Határvonal (Boundary line): Az SVM-ben a Hyper Plane-n kívül két vonal van, amely margót hoz létre. A támogató vektorok lehetnek a határvonalakon vagy azon kívül. Ez a határvonal elválasztja a két osztályt. Az SVR-ben a koncepció ugyanaz.
* Támogató vektorok (Support vectors): Ezek az adatpontok, amelyek a legközelebb vannak a határhoz. A pontok távolsága legalább minimális.

## Miért SVR? Mi a fő különbség az SVR és az egyszerű regressziós modell között?

Egyszerű regresszióban igyekszünk minimalizálni a hibaarányt. Miközben az SVR-ben egy hiba küszöbén belül próbálunk illeszkedni.

Példa: Az egydimenziós támogató vektor regressziós (SVR) modell vázlata. Csak a „tube” (csövön) kívül eső pontokat használják előrejelzések készítéséhez.



1. Valamelyik tényezőváltozó kifejezhető a többi tényezőváltozó nem triviális lineáris kombinációjaként. [↑](#footnote-ref-3)
2. Együttható [↑](#footnote-ref-4)
3. http://www.saedsayad.com/support\_vector\_machine\_reg.htm [↑](#footnote-ref-5)