Üzleti Elemzések Módszertana

3. Gyakorlat: Regularizált modellek

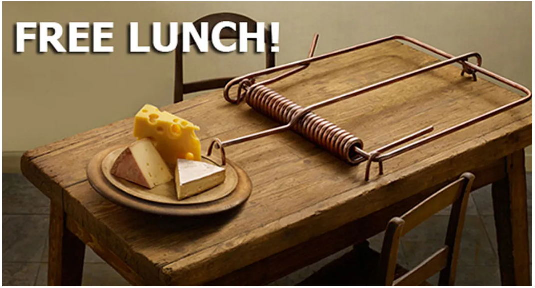
Kuknyó Dániel
Budapesti Gazdasági Egyetem

2023/24
2.félév

Nincs ingyen ebéd

Ez az elv azt sugallja, hogy egy algoritmus, amely egyes feladatokon jól teljesít, nem feltétlenül fog ugyanolyan jól működni más feladatokon. **Nincs egyetemesen optimális algoritmus minden lehetséges problémára**. Ennek következményei:

- Az algoritmusokat szükséges specializálni
- Az egyes feladatokon össze kell hasonlítani különböző modellek teljesítményét
- Nincs egyetemes megoldás egyetlen problémára

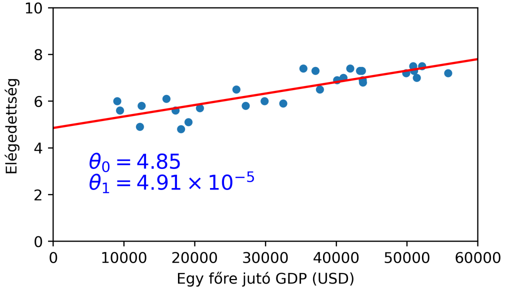


Nem reprezentatív vagy hiányos adatok

A diagramon az egy főre jutó GDP és az emberek életükkel való elégedettsége látható. A mintában nem szerepelnek a gazdagabb és szegényebb országok. Ebben az esetben a becsült modell:

$$\hat{y} = 4.85 + 4.91 \cdot 10^{-5} \cdot x$$

Tehát $\theta = [\theta_0, \theta_1] = \left[4.85, 4.91 \cdot 10^{-5}\right]$

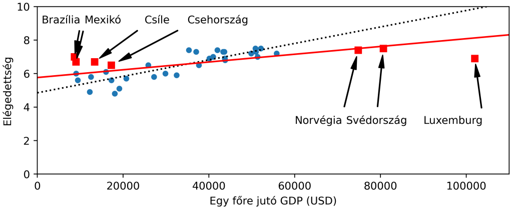


Nem reprezentatív vagy hiányos adatok

Érdemes megfigyelni mi történik, ha több ország is hozzáadódik a mintához, amik eddig nem szerepelnek benne. Az **illesztett modell egy szignifikáns változáson megy keresztül**:

$$\hat{y} = 5.8 + 2.31 \cdot 10^{-5}$$

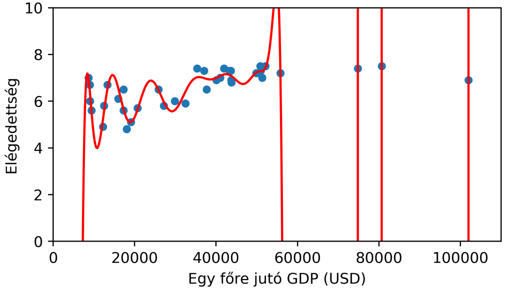
A modell változása annak a következménye, hogy az előző minta **nem volt megfelelően reprezentatív** az összes országra való tekintettel.



A túltanulás problémája

A túltanulás onnan ered, hogy az **illesztett modellnek túlságosan magas a szabadságfoka**, vagy túl sok iteráción keresztül folyik a tanítás.

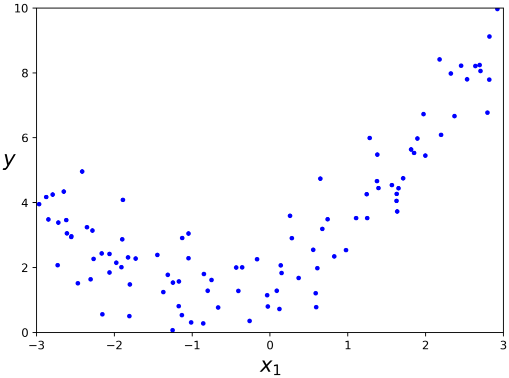
A túltanult modell **nagyon pontosan képes illeszkedni a tanító adatokra**, de nem fog jól teljesíteni a teszt adatokon. Emiatt rossz általánosító képességekkel fog rendelkezni.



Polinomiális adatok

A valóságban nagyon kevés olyan adathalmaz létezik, amik leírhatók egy lineáris modellel

Ennek megfelelően szükség van komplexebb, nagyobb paraméterszámmal rendelkező modellekre, mint a polinomiális regresszor modellek.



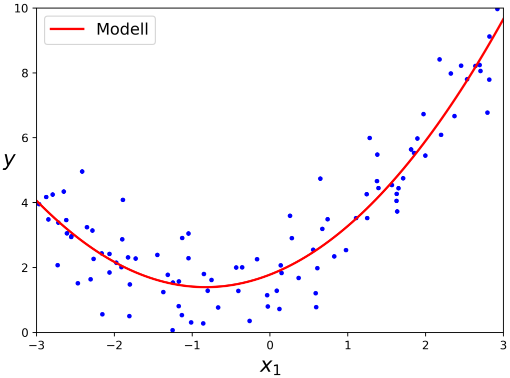
Polinomiális adatok

Polinomiális regresszor modell

Olyan típusú regressziós analízis, ahol a kapcsolat az x és y változók között **polinomok segítségével van modellezve**.

A modell magában foglalja **a független változók magasabb rendű hatványait** és azok interakcióit is. Általános formája:

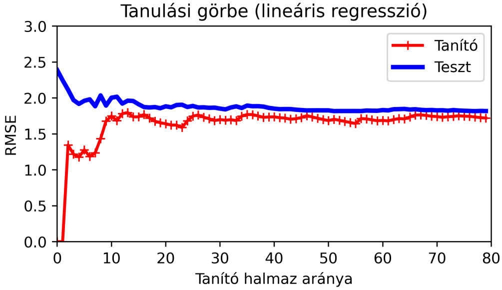
$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \cdots + \theta_n x^n$$



Tanulási görbe lineáris és polinomiális esetben

A tanulási görbe azt mutatja meg, hogy az adott modellnek mekkora a hibája a tanító és teszt adatokon, a tanító adatok mennyiségének függvényében.

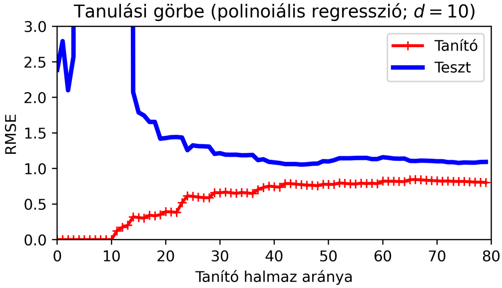
A polinomikus modellnek **magasabb a szabadságfoka, ezért alacsonyabb hibát képes elérni** ugyanazon az adathalmazon. Viszont ebben az esetben fennáll a túltanulás kockázata.



Tanulási görbe lineáris és polinomiális esetben

A tanulási görbe azt mutatja meg, hogy az adott modellnek mekkora a hibája a tanító és teszt adatokon, a tanító adatok mennyiségének függvényében.

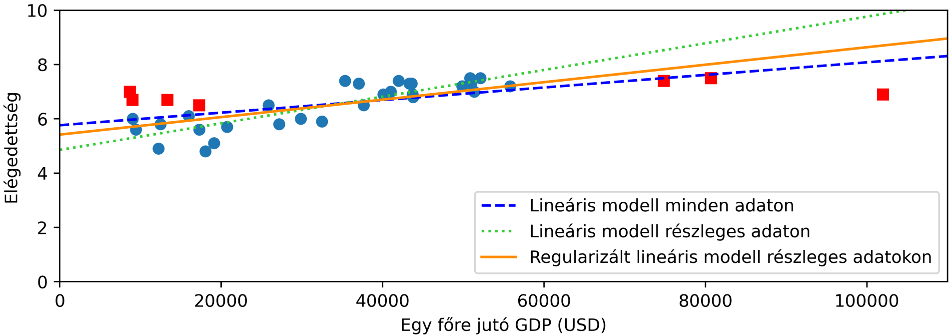
A polinomikus modellnek **magasabb a szabadságfoka, ezért alacsonyabb hibát képes elérni** ugyanazon az adathalmazon. Viszont ebben az esetben fennáll a túltanulás kockázata.



Regularizáció

Regularizáció

Gépi tanulásban és a statisztikai modellezésben egy technika, amelynek célja, hogy csökkentse a modell túlillesztését és javítsa a modell általánosító képességét új adatokra azáltal, hogy **segít a modell súlyait alacsonyan tartani**.



Ridge regresszió

A lineáris regresszió regularizált változata, más néven Tikhonov regularizáció. Az algoritmus a függvény pontos illesztése mellett **célja a súlyokat a lehető legalacsonyabban tartani**.

Ezt úgy éri el, hogy a tanítási fázisban bevezet egy **regularizációs büntetőkifejezést**, és hozzáadja a már meglévő költségfüggvényhez.

A Ridge költségfüggvénye

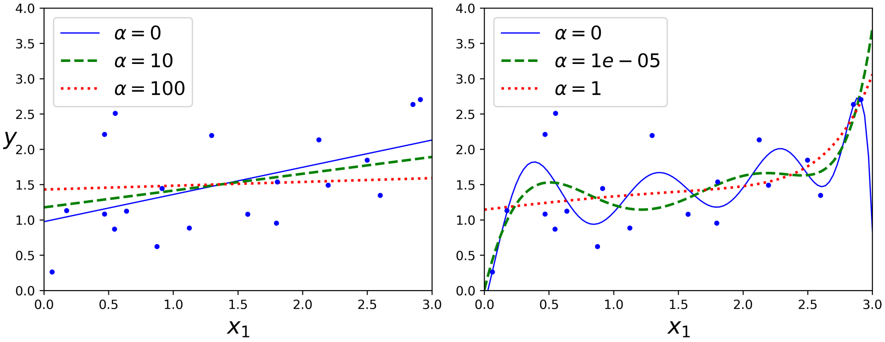
$$J(\theta) = MSE(\theta) + \alpha \sum \theta^2$$

Ahol:

- $MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ az átlagos eltérés-négyzet
- $\sum \theta^2$ az ℓ_2 norma büntetőkifejezés
- α a regularizáció mértékét adó együttható
- θ a paraméterek vektora

Ridge a gyakorlatban

A két ábrán különböző α paraméterrel tanított regularizált modellek láthatók. A bal oldali diagramon lineáris modellek, a jobb oldalin pedig polinomiális függvények illeszkednek az adathalmazra. Minél nagyobb az α, annál általánosabb az modell.



LASSO regresszió

A LASSO (**L**east **A**bsolute **S**hrinkage and **S**election **O**perator) regresszió a ridge-hez hasonlóan egy büntetőkifejezés segítségével regularizálja az illesztett modellt.

Ebben az esetben a büntetőkifejezés nem a paraméterek négyzetének összege, hanem az abszolút értékének.

A LASSO költségfüggvénye

$$J(\theta) = MSE(\theta) + \alpha \sum |\theta|$$

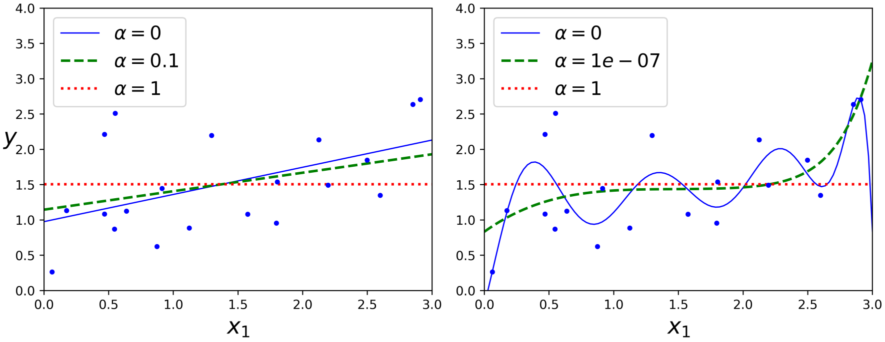
Ahol:

- $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ az átlagos eltérés-négyzet
- $\sum |\theta|$ az ℓ_1 norma büntetőkifejezés
- α a regularizáció mértékét adó együttható
- θ a paraméterek vektora

LASSO a gyakorlatban

A LASSO a regularizáláció mellett elvégzi a jellemzőkiválasztás műveletét is. A felesleges paramétereket eliminálja, tehát 0 értékeket ad nekik.

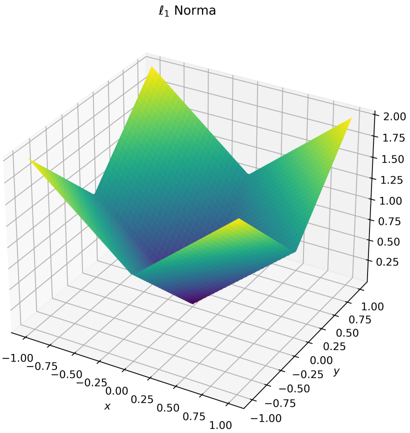
Az így létrejövő modellben kevés lesz a nem nulla súly.



ℓ_1 és ℓ_2 norma

A LASSO büntetőkifejezése az ℓ_1 normát, a ridge pedig az ℓ_2 normát használja.

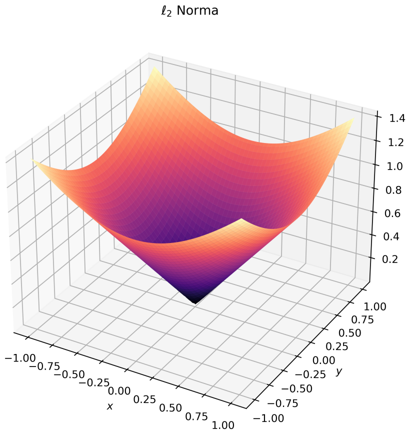
- A LASSO jobban teljesít kevés független változó esetén
- A LASSO végez jellemzőkiválasztást
- A LASSO az egymással korreláló változók közül egyet hagy meg
- A ridge az egymással korreláló változókat együtt kezeli
- A ridge jobban teljesít, amikor minden független változó befolyásolja az outputot



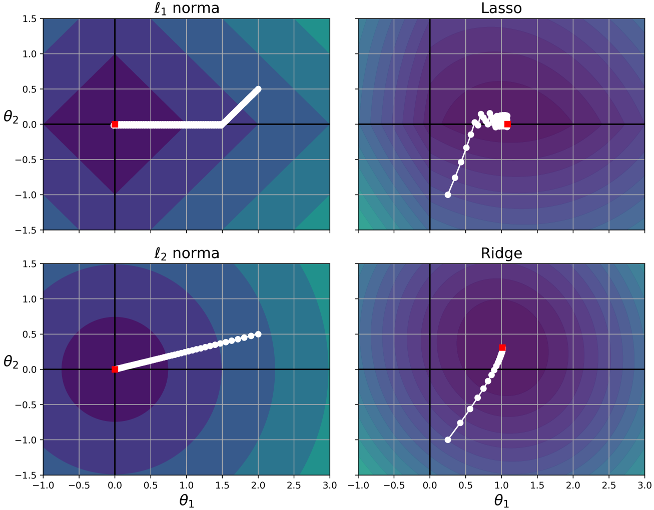
ℓ_1 és ℓ_2 norma

A LASSO büntetőkifejezése az ℓ_1 normát, a ridge pedig az ℓ_2 normát használja.

- A LASSO jobban teljesít kevés független változó esetén
- A LASSO végez jellemzőkiválasztást
- A LASSO az egymással korreláló változók közül egyet hagy meg
- A ridge az egymással korreláló változókat együtt kezeli
- A ridge jobban teljesít, amikor minden független változó befolyásolja az outputot



Gradiens ereszkedés LASSO és ridge regularizált függvényeken



Elasztikus hálók [demo]

Az elasztikus háló egy **arányosságot definiál a LASSO és ridge modellek között**. A büntetőkifejezés egy keveréke a LASSO és ridge költségfüggvényeinek egy adott r keverési arány szerint.

Ha $r = 0$ az elasztikus háló megegyezik a ridge modellel, ha $r = 1$ akkor megegyezik a LASSO függvénnyel.

Az elasztikus háló költségfüggvénye

$$J(\theta) = MSE(\theta) + r \cdot \alpha \cdot \sum |\theta| + (1 - r) \cdot \alpha \cdot \sum \theta^2 =$$

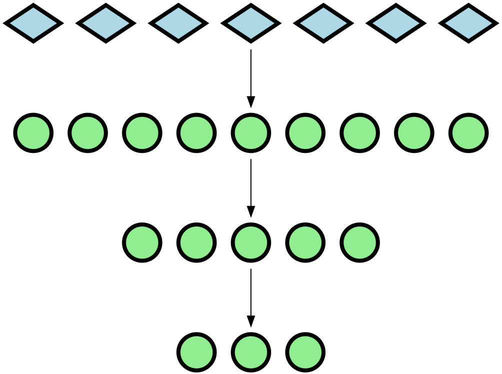
$$= MSE(\theta) + r \cdot \alpha \cdot \ell_1 + (1 - r) \cdot \alpha \cdot \ell_2$$

- $MSE(\theta)$: Átlagos eltérés-négyzet
- $r \in [0, 1]$: Keverési arány
- α: Regularizációs együttható

Jellemző összevonás és kiválasztás

Jellemző összevonás

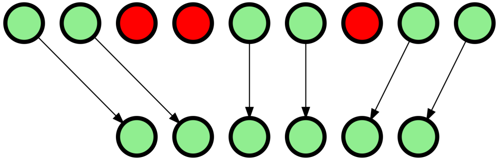
Jellemzők összevonása során meglévő változók alapján új változók kerülnek létrehozásra valamilyen aggregációs függvény vagy más eljárás alapján pl. főkomponenselemzés.



Jellemző összevonás és kiválasztás

Jellemző kiválasztás

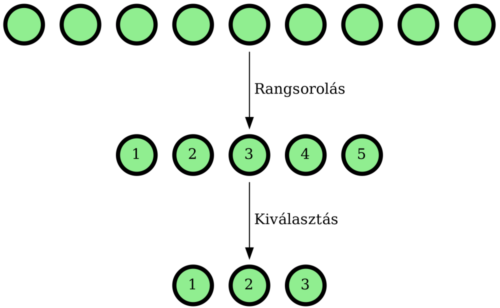
A jellemzők kiválasztása során változók közül eldobásra kerülnek azok, amelyekre nincs szükség a modellezés szempontjából valamilyen metrika szerint.



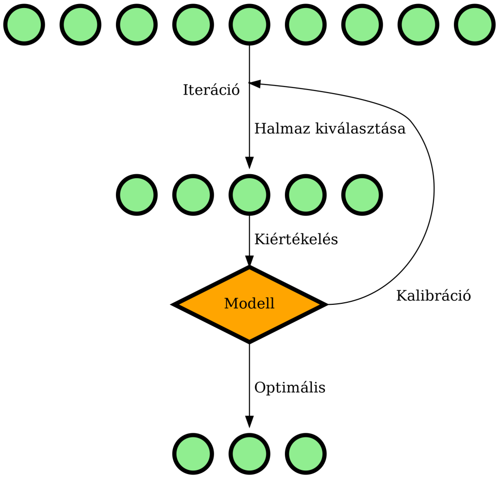
Jellemzők kiválasztásának lehetséges módjai

Szűrés

Nem tesztel adott algoritmust, csak egy módszertan szerint fontossági sorrendet definiál a változók között, és az egy küszöbérték alattiakat elveti.



Jellemzők kiválasztásának lehetséges módjai



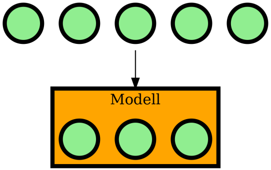
Iteratív

Specifikus modelleket kiértékel a jellemzők különböző részhalmazai szerint, majd azt választja ki amelyik a legjobb eredményt adja.

Jellemzők kiválasztásának lehetséges módjai

Beágyazott

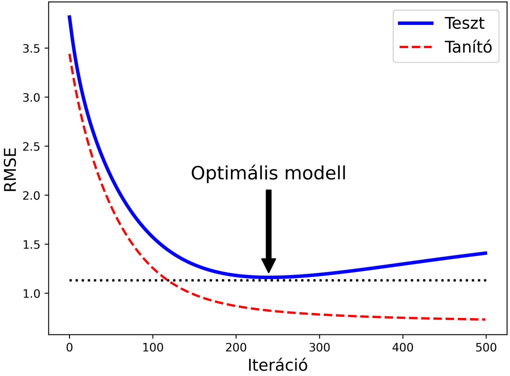
Minden technika ide tartozik, ami a tanítási fázisban jellemzőkiválasztást végez. Pl.: LASSO.



Korai leállás

A korai leállás egy olyan technika gépi tanulásban amely **segít megakadályozni a túltanulást egy tanítási folyamat során**.

Ezáltal **leállítja a tanítást, amint a hálózat teljesítménye elkezd romlani**, vagy nem javul tovább a validációs adatkészleten.



k-hajtásos keresztvalidáció

A technika az eredeti adatkészletet k egyenlő részre osztja, **majd minden iterációban egy különböző részhalmazt használ tesztelésre, a többit pedig tanításra**.

Ez lehetővé teszi **a modell általánosító képességének pontosabb becslését**.