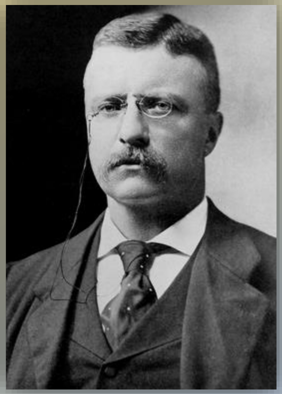
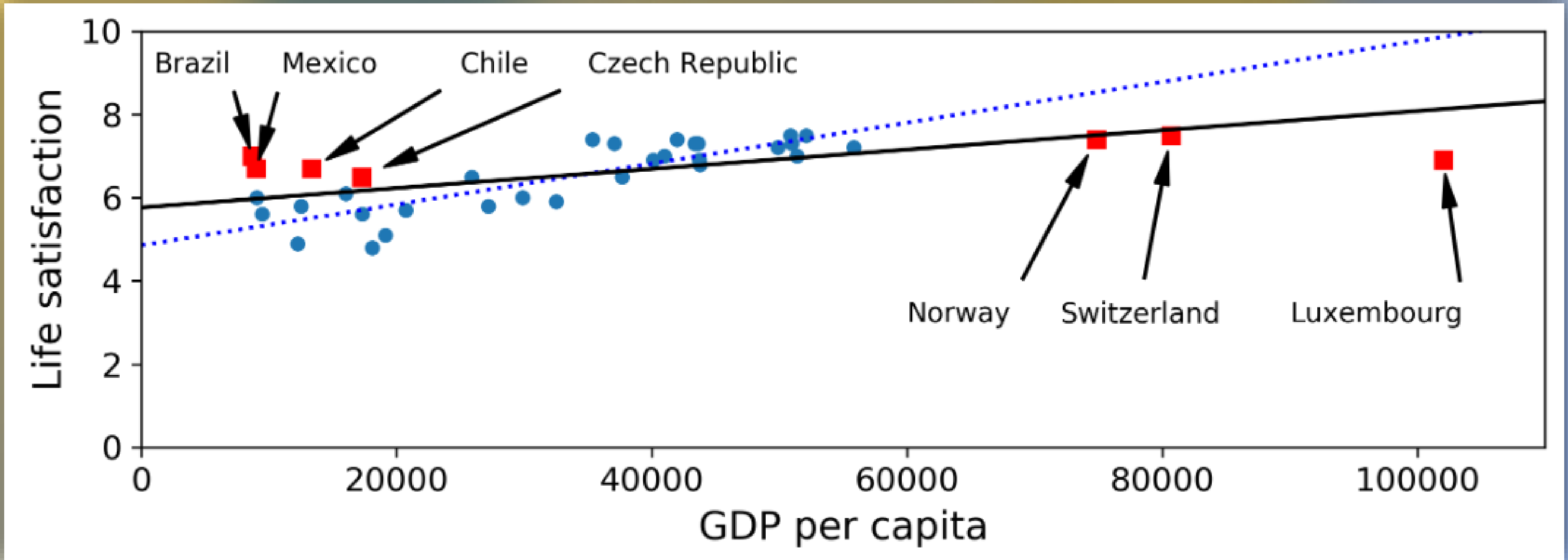
3. Előadás
Lasso, Ridge, Elasztikus háló
Early Stopping, Keresztvalidáció

A machine learning kihívásai

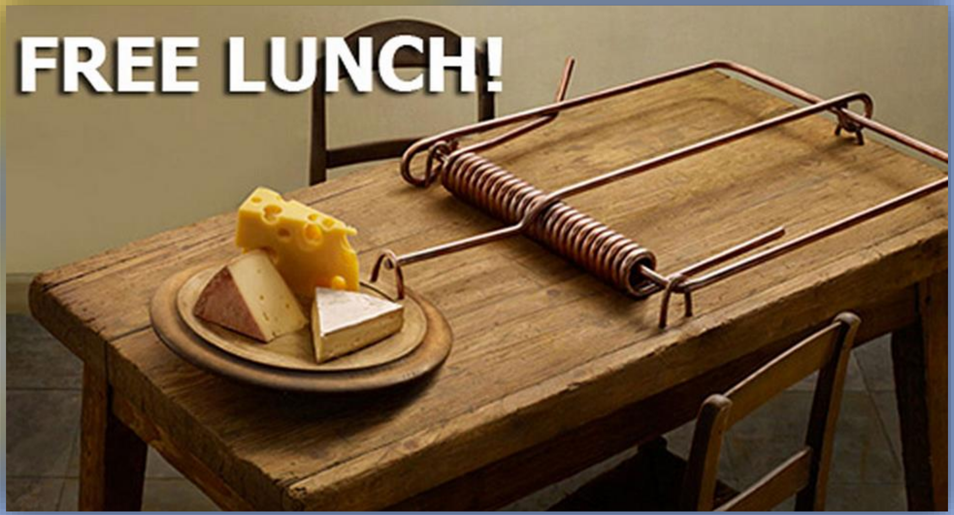


- Nem-reprezentatív vagy hiányos tanító adatok (Sampling Bias, Nonresponse Bias).
- Vegyünk egy példát: ha az alábbi országokra szeretnénk modellt állítani, de a pirossal jelölt országok hiányoznak, mert nincs adatunk a gazdagabb országokból, lényeges különbség lesz a létrejövő modellek között.



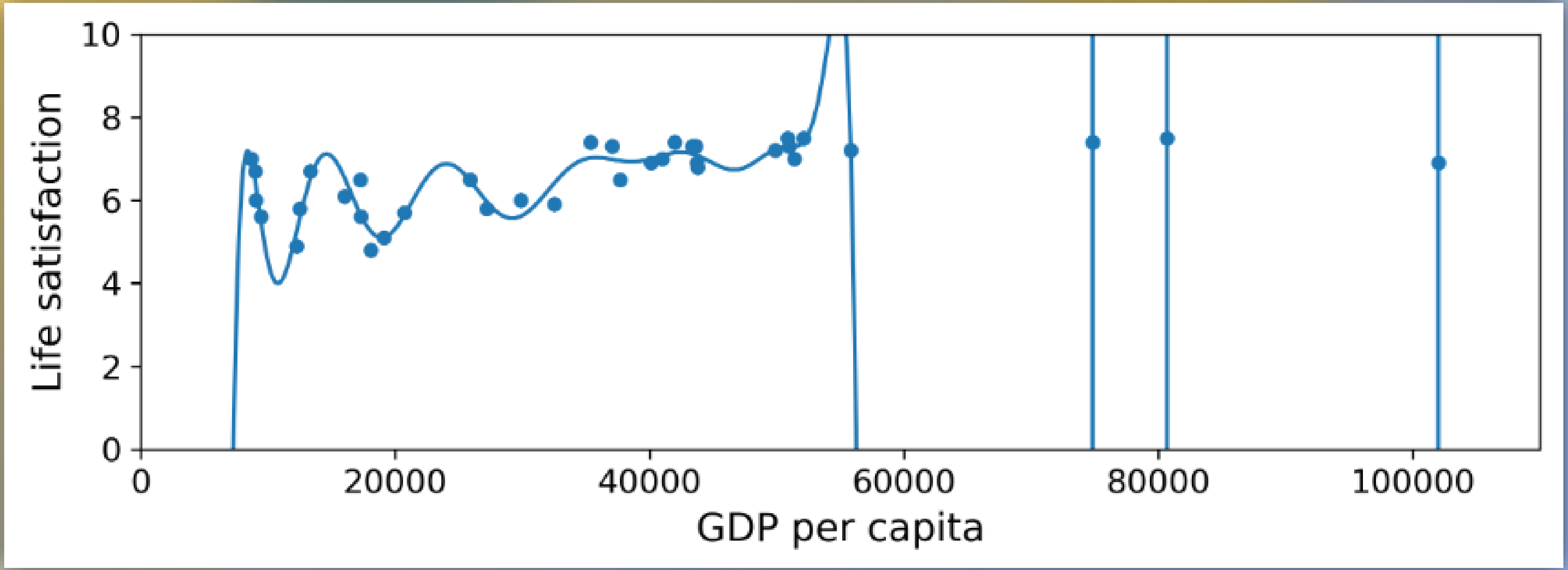
No Free Lunch elmélet

- Egy híres, [1996-os tanulmányban] David Wolpert demonstrálta, hogy ha nincs valamilyen állításunk, vagy elvárásunk az adatok irányába, akkor nincs okunk valamelyik modellt preferálni a többi helyett.
- Ez a No Free Lunch elmélet: valamelik adathalmaznál a lineáris regresszió, valamelyiknél pedig a neurális hálózat fog jobb predikciókhoz vezetni.
- Nincs olyan modell, amelyik lényegéből fakadóan *jobb* lenne mint a többi.
- Nem lehet megúszni a modellek összehasonlítását!



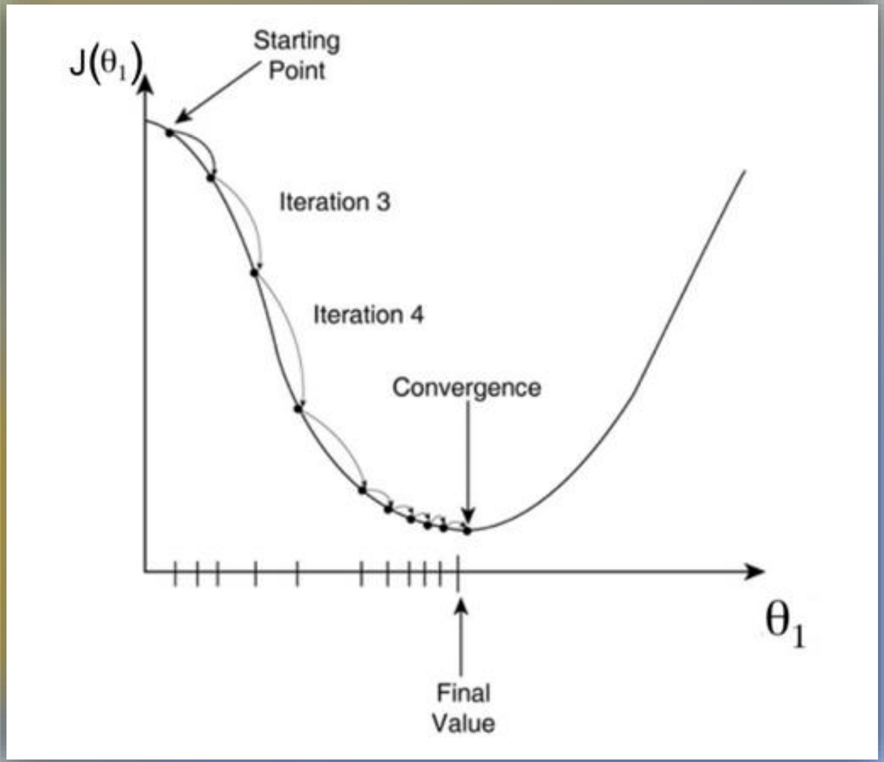
A túltanulás problémája

- Egy túltanult modell nagyon pontosan illeszkedik a tanító pontokra, de az általa reprezentált relációk a valóság kapcsolatait eltorzítják.
- A túltanulás onnan ered, hogy egy túlságosan nagy szabadságfokkal rendelkező függvényt illesztünk rá kevés tanító pontra.



Miért van szükség költség- és jóság függvényre?

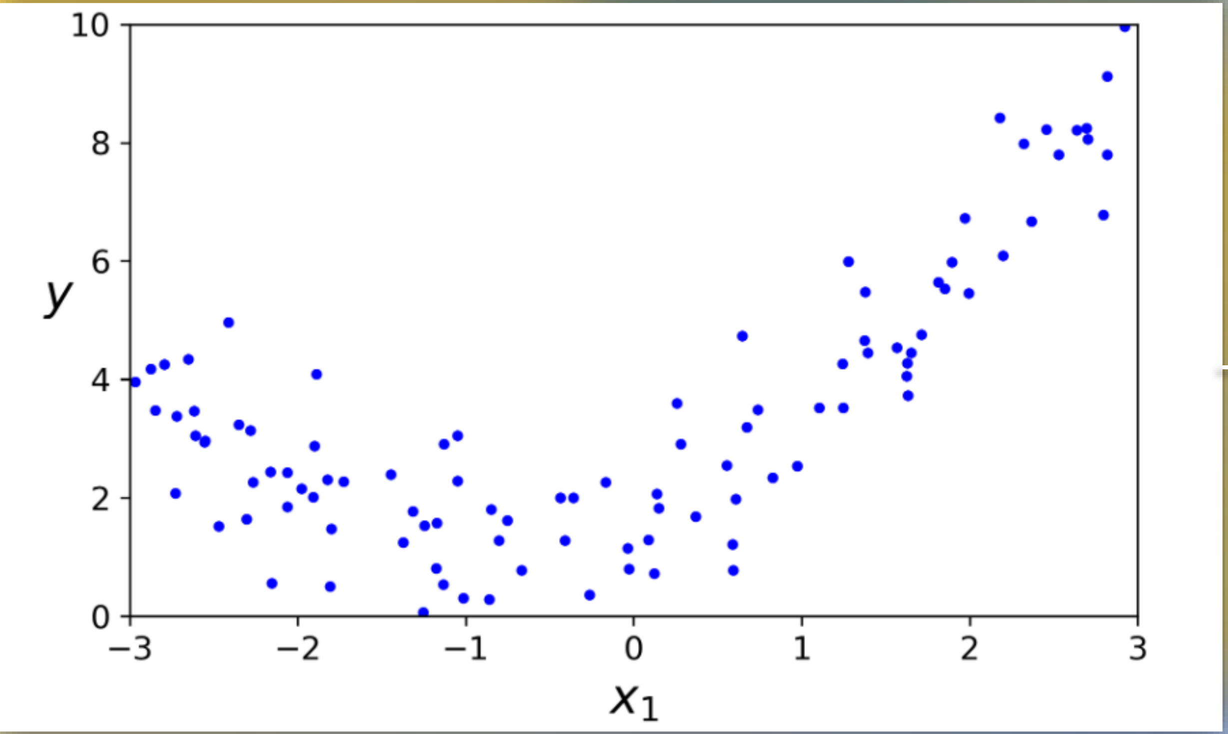
- Gyakori a különböző költség- és jóság függvény használata a tanítási és teszt fázis során. A regularizáción kívüli oka az, hogy a jól használható tanító költségfüggvénynek optimalizáció-barát deriváltjainak kell lennie, míg a teszt fázis során használt jóságfüggvénynek olyan közel kell lennie a célhoz, amennyire csak lehetséges.
- Például osztályozásnál: a **log loss-t** használjuk költségfüggvényként, a **precision-t** és **recall-t** jóságfüggvényként.

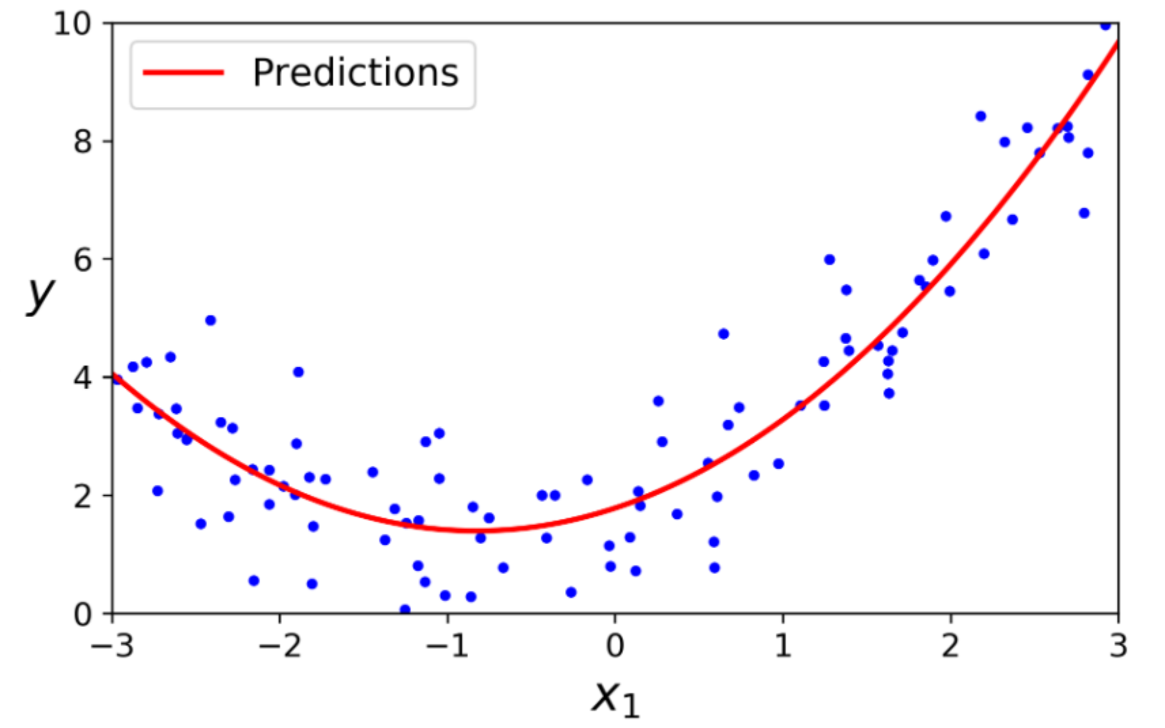


- Θ_1: adott paraméter értéke, pl. meredekség
- $J(\Theta_1)$: költség Θ_1 szerint

Polinomikus regresszió

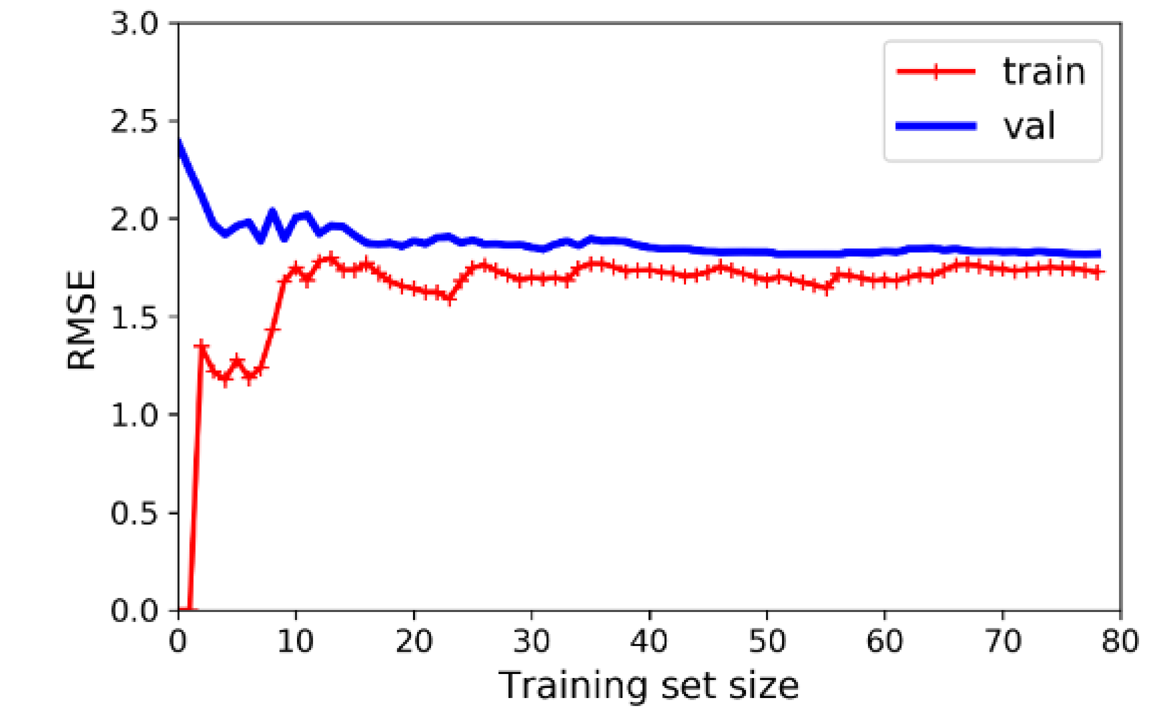
- Mi van, ha az adatok komplexebbek egy egyenes vonalnál?
- Egy lineáris modellt lehetséges nemlineáris adatokra illeszteni.
- Az egyik módja, hogy a paramétereket hatványra emeljük, és egy kiterjesztett lineáris modellt tanítunk az új jellemzőkkel.

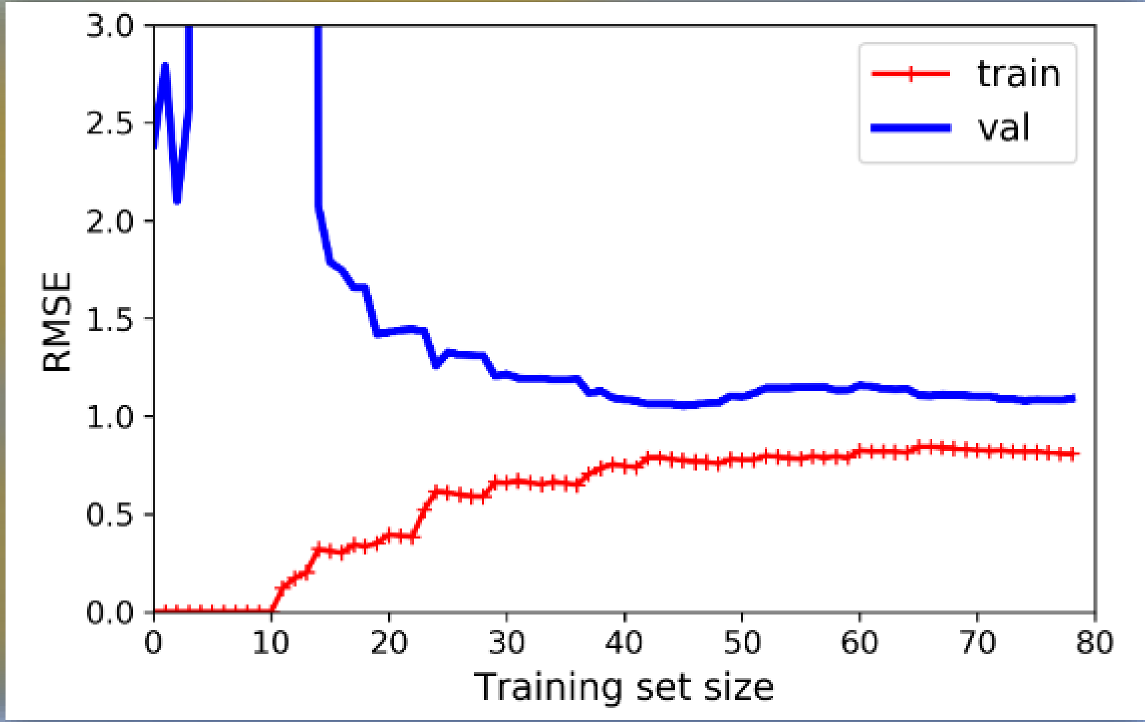




Lineáris vs. Polinomikus regresszió

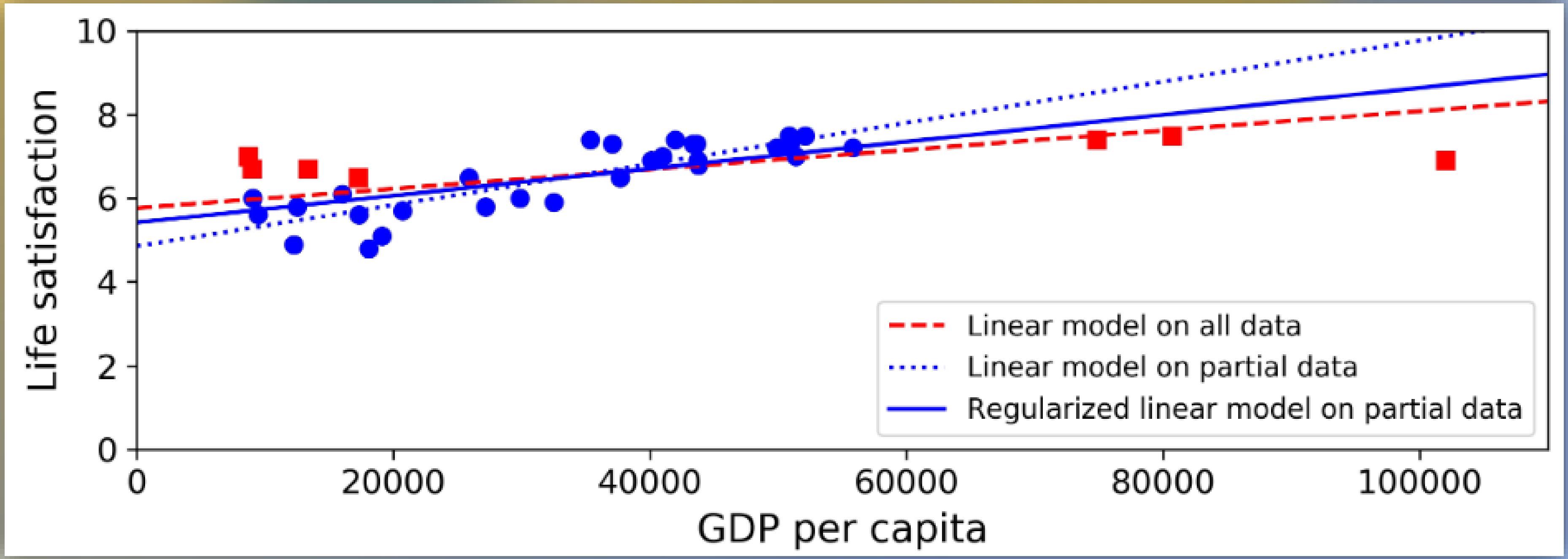
- Polinomikus: az együtthatók magasabb rendűek is lehetnek.
- A tanulási görbe azt mutatja meg, hogy az adott modellnek mekkora a hibája a tanító és teszt adatokon, a tanító adatok mennyiségének függvényében.
- Lássunk két tanulási görbét lineáris és polinomikus regresszióra:





A regularizáció

- Látjuk tehát, hogy minél kevesebb az illesztett függvény **szabadságfoka**, annál könnyebben elkerülhető a túltanulás.
- Egy lineáris modell esetében a regularizáció tipikusan úgy érhető el, hogy a modell súlyai felé megkötésekkel élünk.
- A lineáris modellnek két súlya van: θ_0 és θ_1, a metszéspontot és a meredekséget szabályozzák. Ezek adják modell szabadságfokát.



Ridge regresszió

- A lineáris regresszió regularizált változata, más néven Tikhonov regularizáció. Az algoritmus a függvény pontos illesztése mellett segít a **súlyokat a lehető legalacsonyabban tartani**. Ez a regularizáció.
- Ezt úgy éri el, hogy a tanítási fázisban bevezet egy regularizációs kifejezést, és hozzáadja a már meglévő költségfüggvényhez. A regularizáció mértékét α hiperparaméter szabályozza.
- Hiperparaméternek nevezzük azokat a változókat, amelyek a tanítást szabályozzák, és közben végig állandóak.
- A ridge regresszió költségfüggvénye:

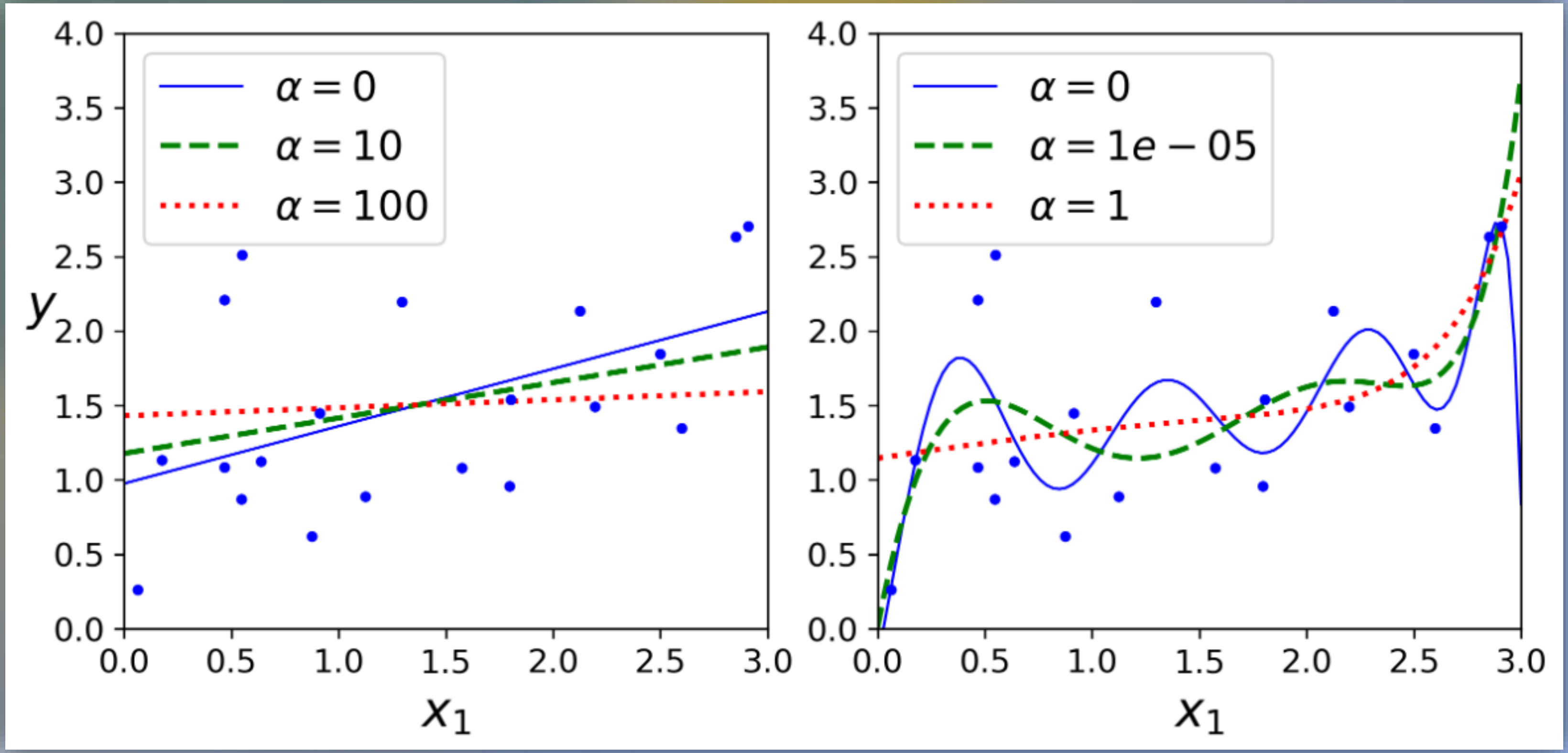
$$J(\boldsymbol{\theta}) = \mathrm{MSE}(\boldsymbol{\theta}) + \alpha\frac{1}{2}\Sigma_{i=1}^{n}\theta_i^2$$

Átlagos eltérés-négyzet

Regularizációs büntetés minden paraméter után: ℓ2 norma

Ridge működés közben

- Mindkét ábrán Ridge modelleket látunk különböző α hiperparaméterekkel tanítva, lineáris adatokon. A bal oldali diagramon lineáris modellek tanítottunk. A jobb oldalin pedig legfeljebb 10 szabadságfokkal rendelkező polinomikus függvényeket illesztettünk.
- Az adatok normalizálása szükséges az eljárás használatához.
- Milyen viszonyban van egymással α és a létrejövő függvény?



Lasso regresszió

- Least Absolute Shrinkage and Selection Operator Regression.
- Hasonlóan a Ridge-hez, egy büntető kifejezést ad a költségfüggvényhez, ezzel nagyobb költségeket rendelve a túltanultabb modellekhez.
- A büntető kifejezés nem a négyzetes, hanem az abszolút hibákat adja hozzá az átlagos négyzetes eltéréshez.
- A Lasso költségfüggvénye:

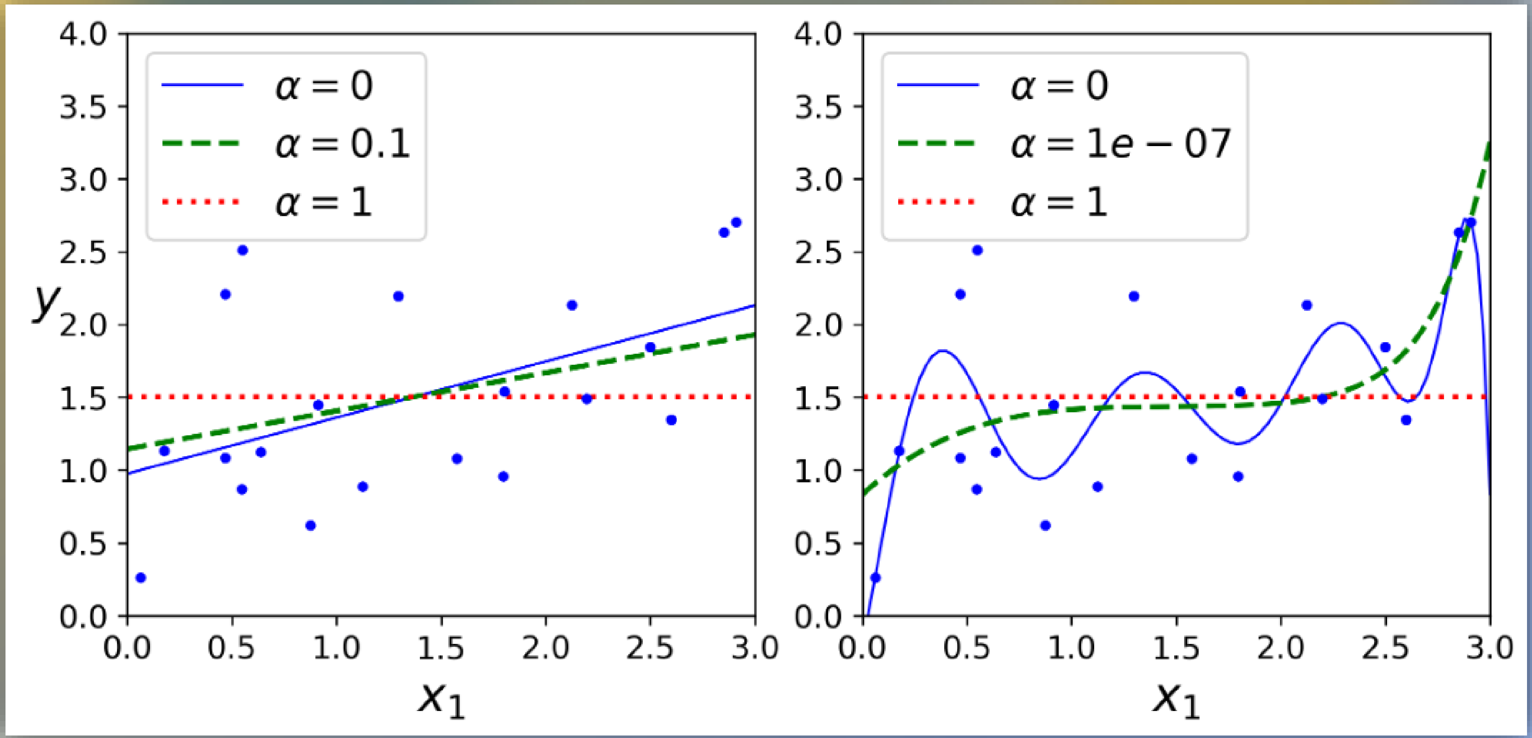
$$J(\boldsymbol{\theta}) = \mathrm{MSE}(\boldsymbol{\theta}) + \alpha \Sigma_{i=1}^{n} \left|\theta_i\right|$$

Átlagos eltérés-négyzet

Regularizációs büntetés minden paraméter után: ℓ1 norma

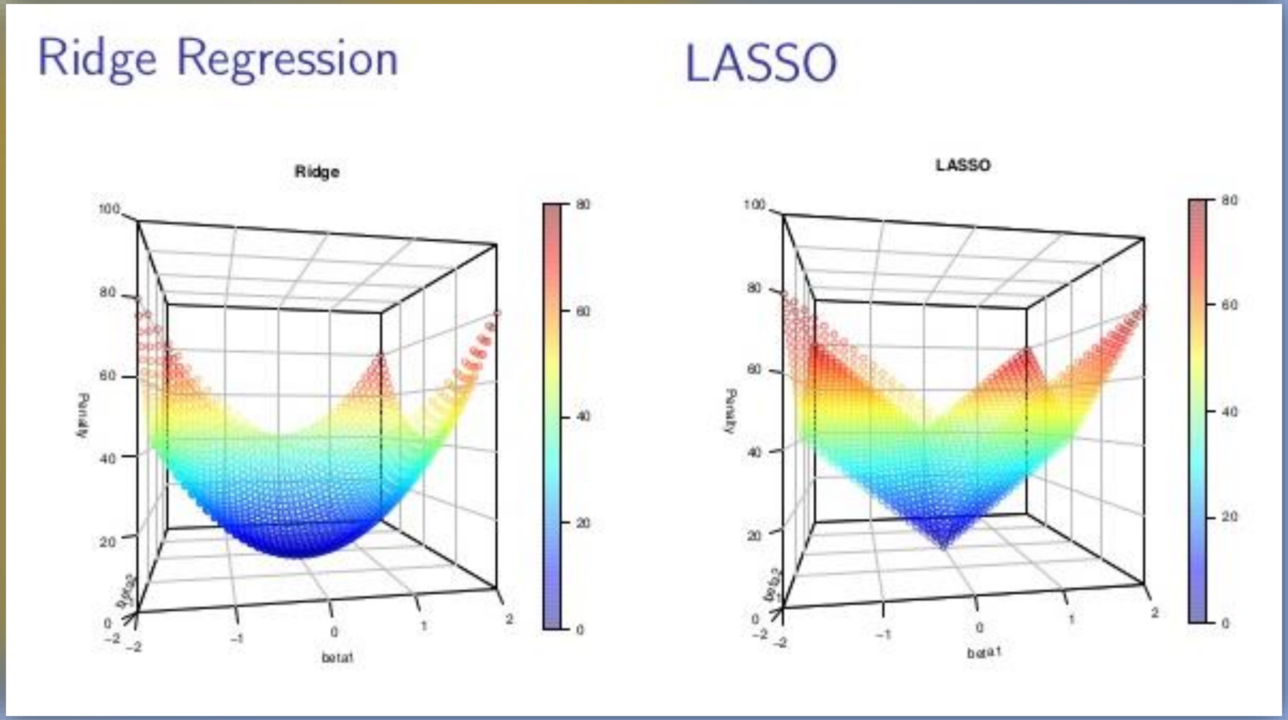
A Lasso jellemzői

- A legkevésbé fontos paraméterek értékeit általában eliminálja ($= 0$).
- Például: a jobb oldali diagram zöld függvénye szinte négyzetesnek néz ki, vagy már majdnem lineárisnak.
- Más szóval: a lasso automatikusan elvégzi a **jellemzőkiválasztás** műveletét: olyan modellt ad eredményül, ahol kevés a nem-nulla súly.



Ridge vs. Lasso

- A Lasso jobban teljesít, amikor kevés független változó van.
- A Ridge jobban teljesít, amikor minden prediktor befolyásolja az outputot.
- A valóságban nem tudjuk, hogy hány változó befolyásolja az outputot. Keresztvalidációval meg lehet állapítani.
- A Lasso végez jellemzőkiválasztást.
- A multikollinearitás problémája:
  - A Ridge-ben az egymással korreláló változókat együtt kezeli.
  - A Lasso az egymással korreláló változók közül egyet hagy meg.

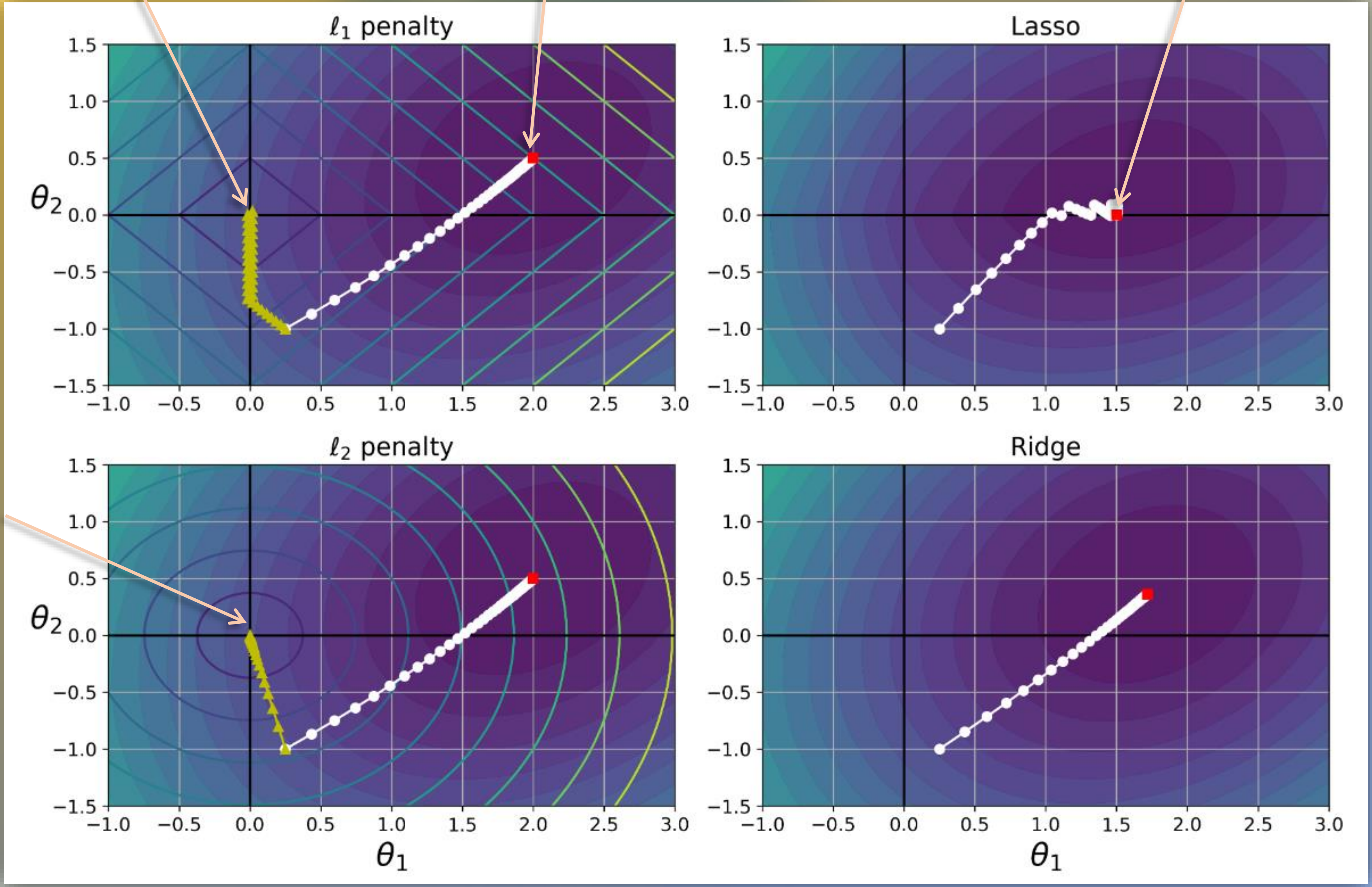


Az ℓ1 büntetés minimuma, gradiens ereszkedéssel, α = 0.5

Regularizálatlan MSE minimum, Gradiens ereszkedéssel

A regularizált + a regularizálatlan függvény együttes optimuma

Az ℓ2 büntetés minimuma, gradiens ereszkedéssel, α = 0.5



Elasztikus hálók

- Az elasztikus háló egy középút a Ridge és Lasso között.
- A regularizációs kifejezés egy egyszerű keveréke a Ridge és Lasso büntető kifejezéseinek, adott arány (r) szerint.
- Ha $r = 0$ az elasztikus háló megegyezik a Ridge-el, ha az $r = 1$ akkor Lasso-ról beszélünk.
- Az elasztikus háló költségfüggvénye:

$$J(\boldsymbol{\theta}) = \mathrm{MSE}(\boldsymbol{\theta}) + r\alpha\sum_{i=1}^{n}\left|\theta_i\right| + \frac{1-r}{2}\alpha\sum_{i=1}^{n}\theta_i^2$$

ℓ2 norma

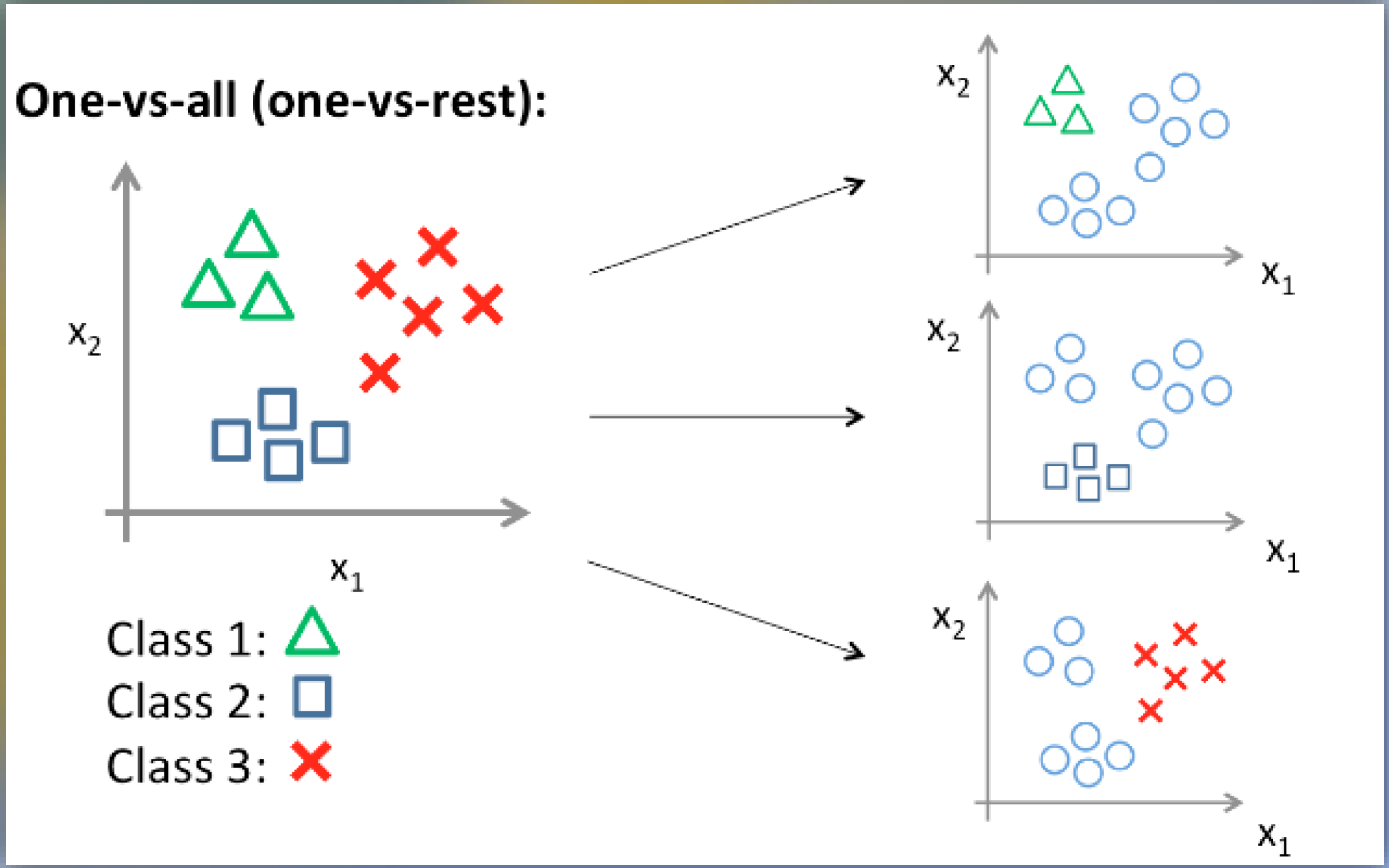
Elasztikus háló
Ridge arány

ℓ1 norma

Elasztikus háló
Lasso arány

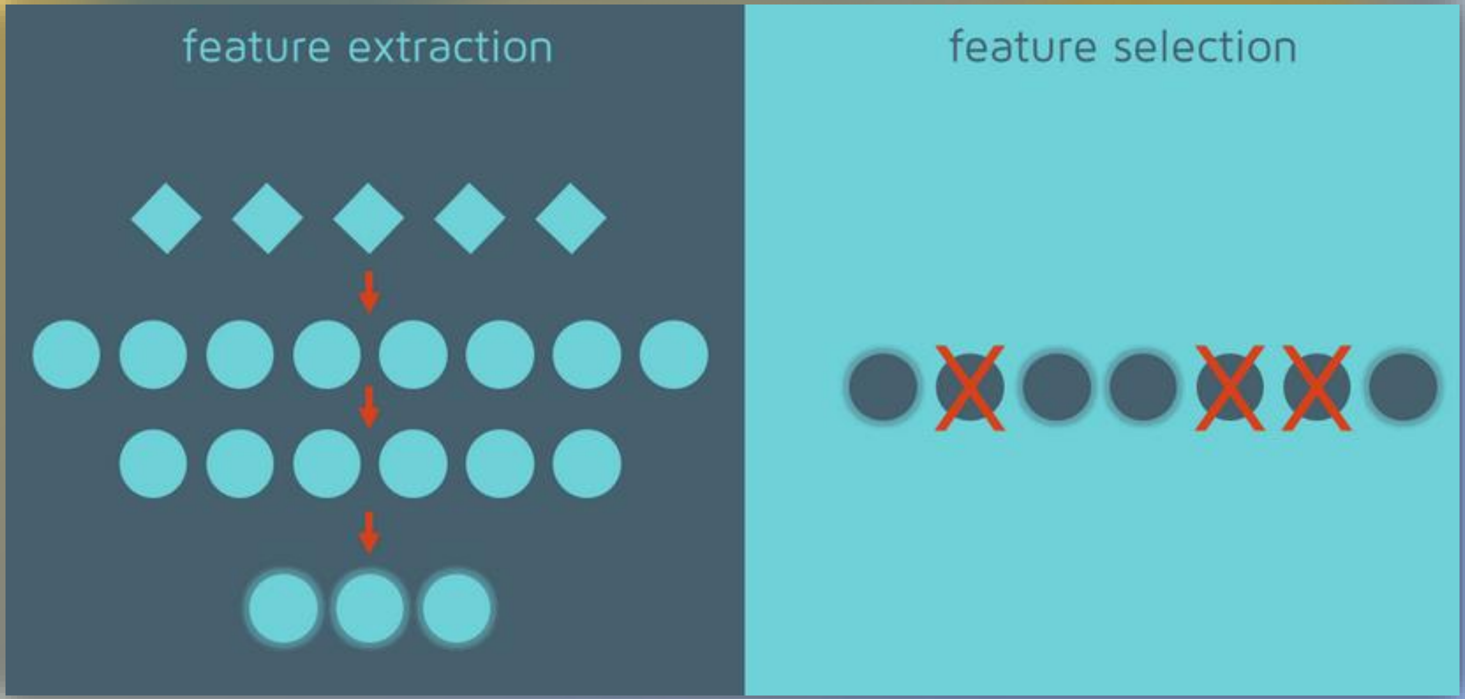
Ridge, Lasso osztályozás

- Egy osztályozási problémát vissza lehet vezetni regresszióra.
- Ekkor a $[0,1]$ osztályokat átalakítja $[-1,1]$ címkékké, és a regresszió eredménye megegyezik a predikció előjelével:
  - Negatív predikció → $\hat{y} = -1$
  - Pozitív predikció → $\hat{y} = 1$
- Multiclass osztályozás esetén **One-vs-All** típusú osztályozás történik.
- A modell visszavezeti a multiclass problémát binárisra: azt vizsgálja, hogy egy mintaegyed inkább egy adott osztályba tartozik-e, vagy az összes többibe.



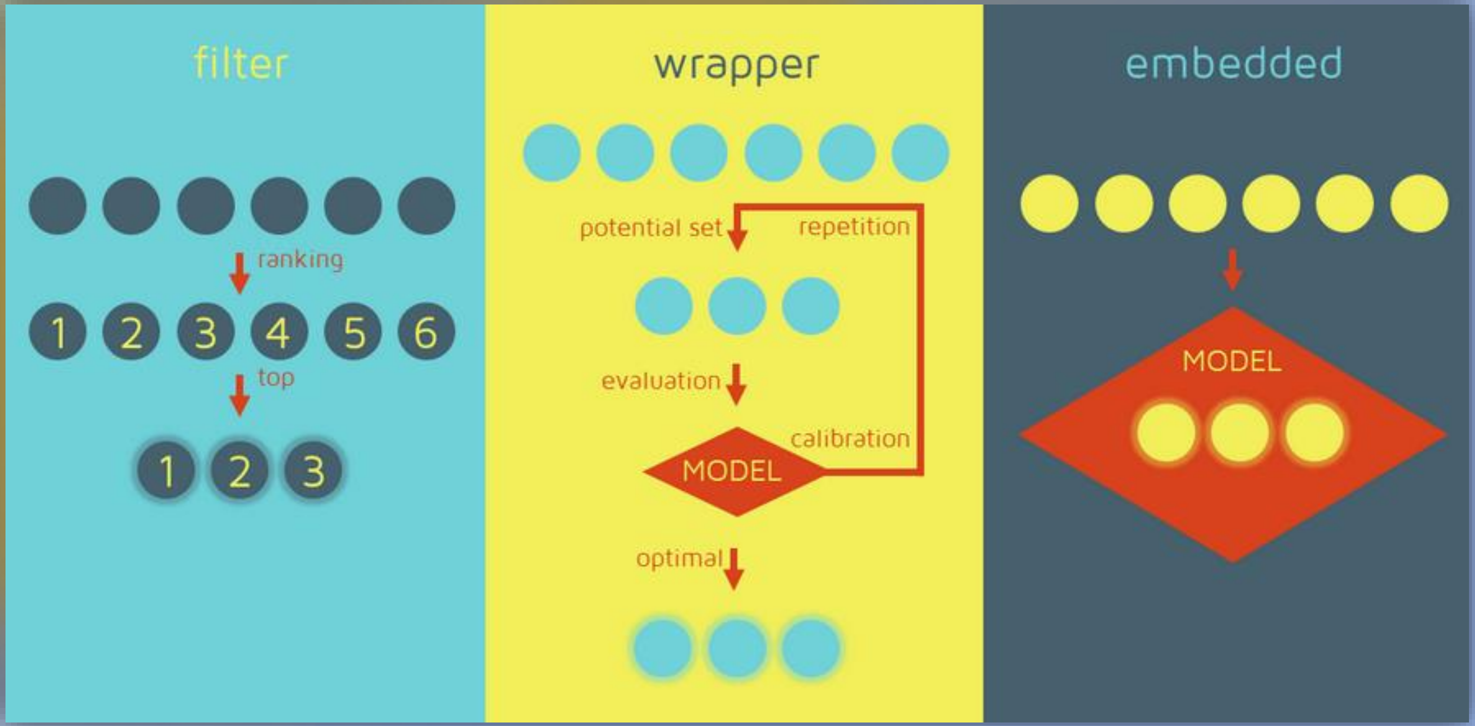
Jellemzösszevonás vs. Jellemzőkiválasztás

- Jellemzőkösszevonás során a meglévő változókat aggregálva tárunk fel az adathalmazban rejlő látens változókat. Pl: Főkomponenselemzés
- Jellemzőkiválasztás során a meglévő változók közül eldobjuk azokat amelyek a predikció szempontjából irrelevánsak. Pl: Korreláció alapján



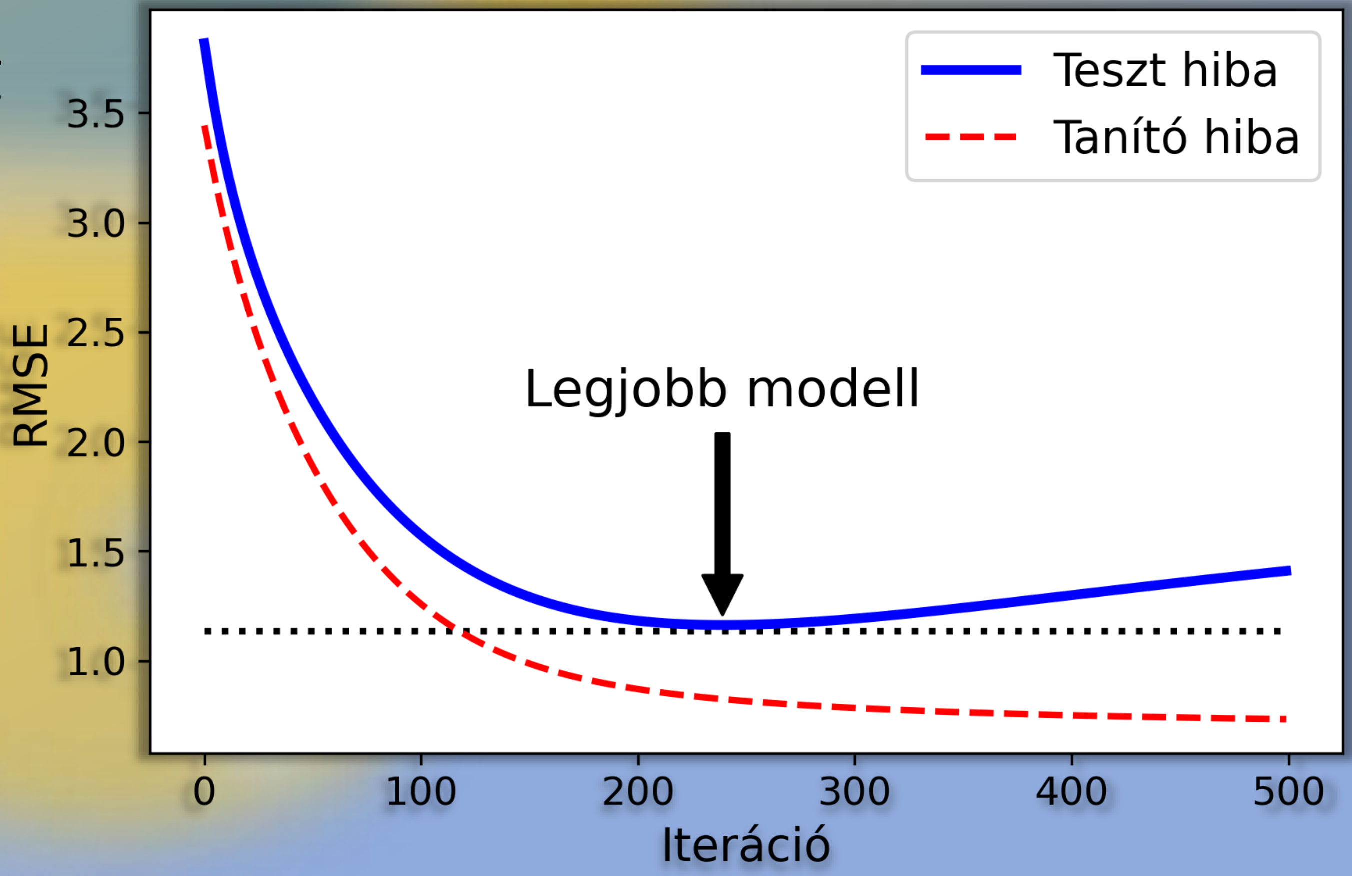
A jellemzőkiválasztás lehetséges módjai

- **Szűrés:** nem tesztel adott algoritmust, csak egy módszertan szerint fontossági sorrendet definiál a változók között, és egy küszöbérték alattiakat elveti.
- **Wrapper:** Specifikus modelleket kiértékel a jellemzők különböző részhalmazai szerint, majd azt választja ki amelyik a legjobb eredményt adja. Nagyon költséges, és túltanulás-gyanús, de ha sikerül, nagyon jó modelleket ad.
- **Beágyazott:** Minden technika ide tartozik, ami a tanítási fázisban jellemzőkiválasztást végez. Pl. Lasso



Early Stopping

- Egy másik, egészen az eddigiektől eltérő módja az iteratív tanuló algoritmusok, mint pl. a gradiens ereszkedés regularizálására, hogy abbahagyjuk a tanítást akkor, amikor a tesztadatok hibája elér egy minimumot.
- A tanítási iterációk előre haladtával a tesztadatokon mért hiba egy ideig csökken, majd amikor a modell túltanulttá válik, elkezd emelkedni.
- Sztochasztikus és mini-batch gradiens ereszkedésnél a görbék nem ennyire simák, és akkor lehet kiszállni, amikor már egy ideje minimumon van a hiba. Pl. 5 iteráció óta.



K-fold Keresztvalidáció

- A tanító adathalmazt k darab, fold-nak nevezett részhalmazba különítjük el.
- Ezután k különböző modellt tanítunk és értékelünk a k részhalmazon.
- Mindezt úgy, hogy minden tanításra és kiértékelésre a tanító halmaznak más és más részét használjuk fel.
- Az eredmény egy k elemből álló értéksor, ami tartalmazza az egyes modellek hibáit.
- Ezzel megkapjuk nemcsak az átlagos hibát, de a hiba szórását is.
- Mi lehet az eljárás hátulütője?