

# Üzleti Elemzések Módszertana

## 1. Előadás: Regresszió

Kuknyó Dániel  
Budapesti Gazdasági Egyetem

2023/24  
2.félév

1 Bevezetés

2 Regresszió

3 Optimalizáció

4 Gradiens ereszkedés

## 1 Bevezetés

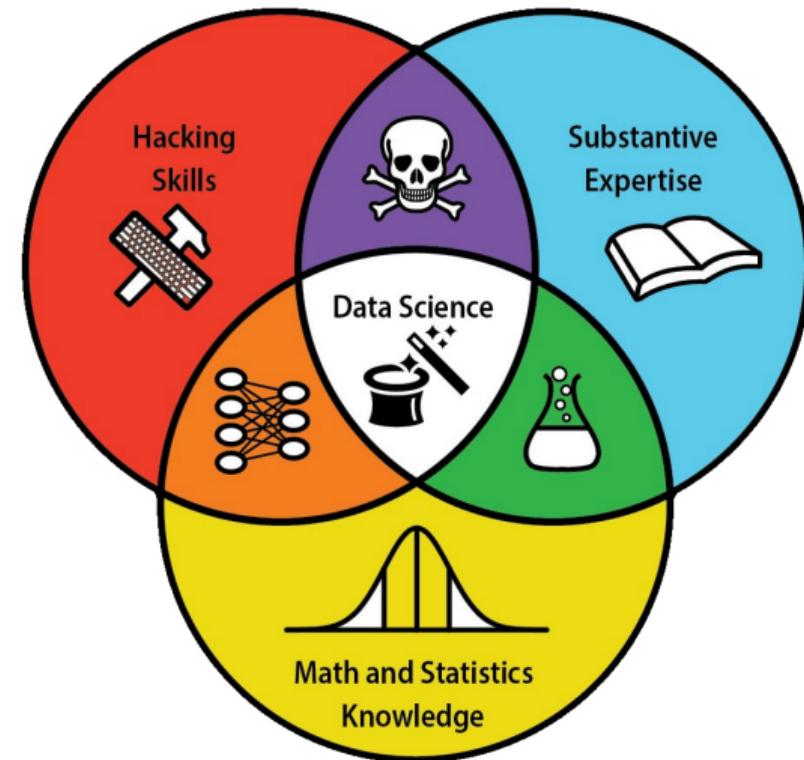
## 2 Regresszió

## 3 Optimalizáció

## 4 Gradiens ereszkedés

# Hol vagyunk?

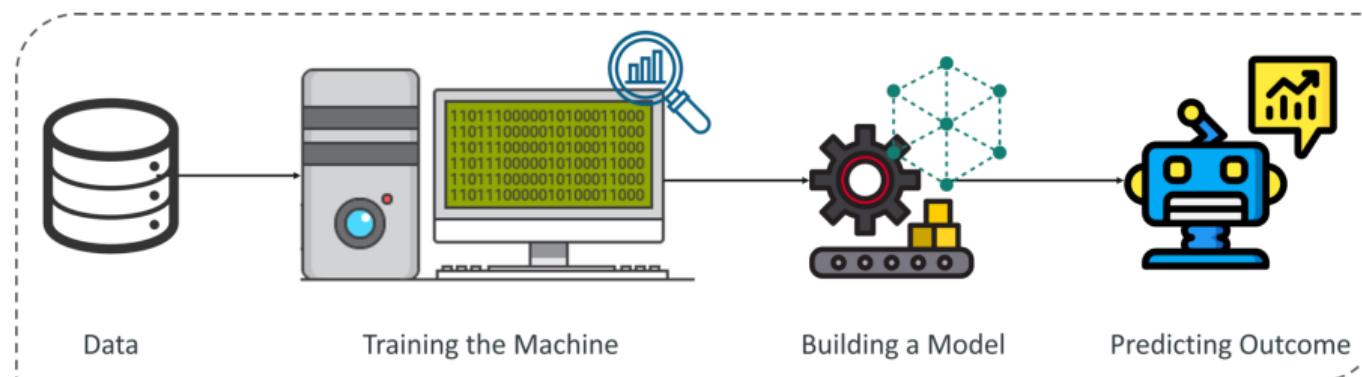
- **Programozási készségekre** van szükség nagy mennyiségű elektronikus adat kezeléséhez
- **A matematika és statisztika** ismerete lehetővé teszi a megfelelő módszerek és eszközök kiválasztását
- **A szaktudás** egy tudományos területen elengedhetetlen az eredmények értelmezéséhez



# A gépi tanulás mögötti megfontolás

A hagyományos szemléletmódban a programozó utasításokat írt egymás után a probléma megoldására.

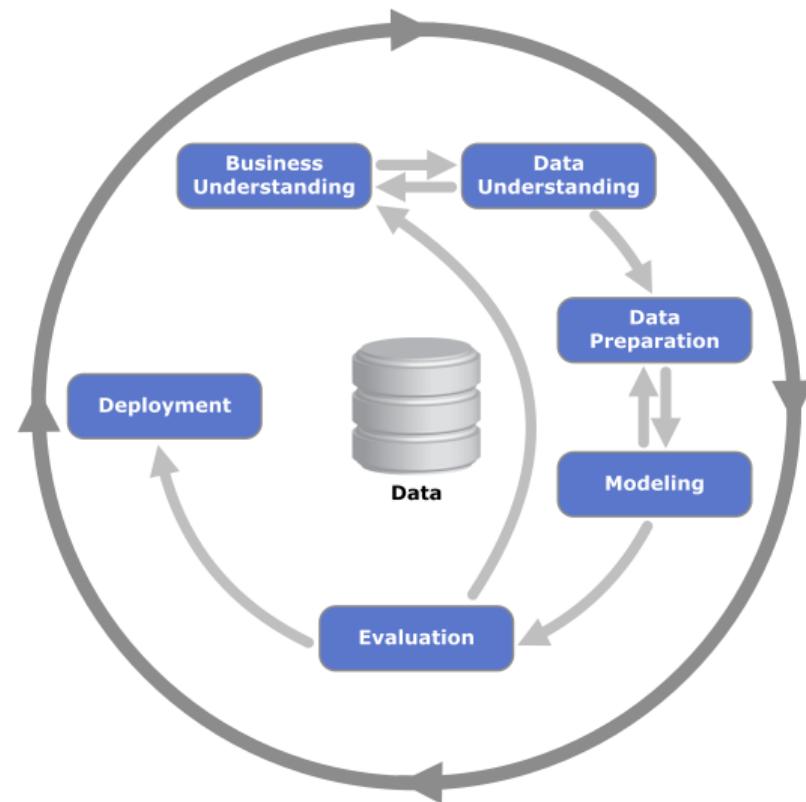
A gépi tanulás szemléletmódjában az algoritmus explicit programozás nélkül tanulja meg megoldani a problémát azáltal, hogy tapasztalat alapján tanul.



# A CRISP-DM módszertan

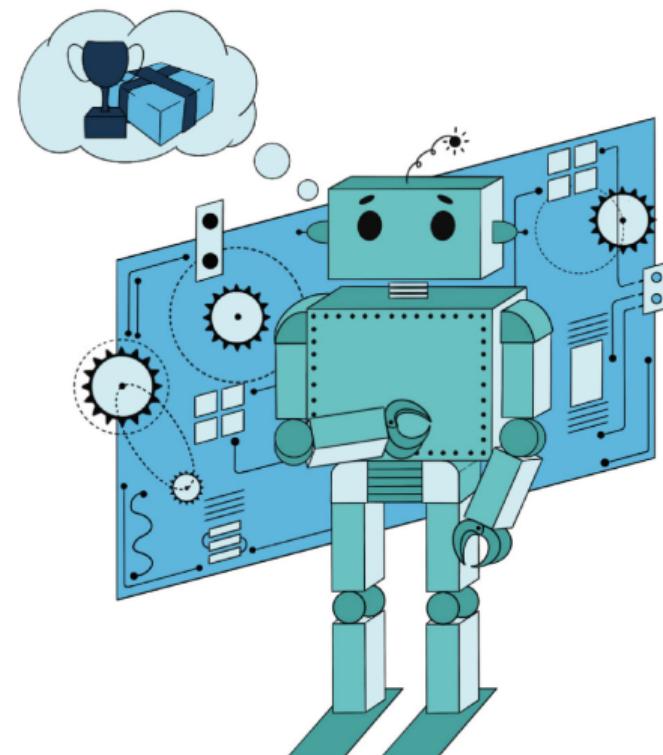
A Cross Industry Standard Process for Data Mining egy folyamatmodell ami az adatbányászati projektek folyamatát adja meg.

- **Business understanding:** Üzleti igények felmérése
- **Data understanding:** Rendelkezésre álló adatok gyűjtése
- **Modeling:** Modell adatokra illesztése
- **Evaluation:** Illesztett modell kiértékelése
- **Deployment:** Átadás a tulajdonosoknak és használatba vétel



# Kihívások a gépi tanulás területén

- Gyenge minőségű adatok
- Nem reprezentatív adatok
- Felesleges jellemzők
- Gép és ember kapcsolata
- Folyamatosan változó világ
- A problémák eredhetnek:
  - Rossz változókból
  - Rosszul általánosító algoritmusból
  - Nehezen értelmezhető eredményekből
  - Nagyon specifikus szakterületi specializációból
  - Elégtelen minőségű vagy mennyiségű adatból



1 Bevezetés

2 Regresszió

3 Optimalizáció

4 Gradiens ereszkedés

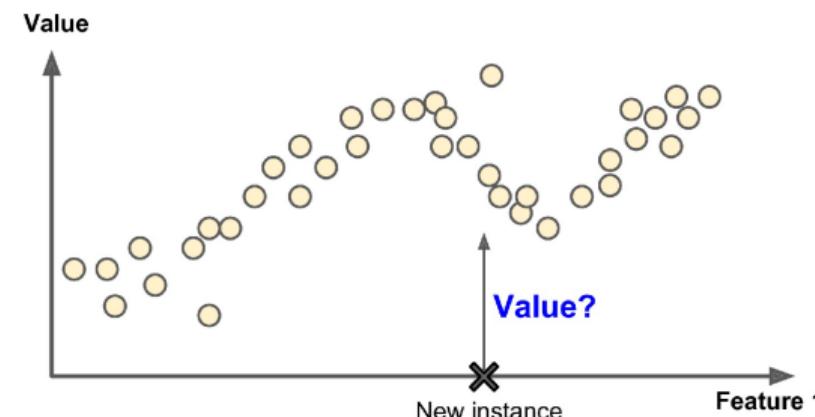
# Alapfogalmak

## Regresszió

Statisztikai elemző eljárás, amely változók közötti kapcsolatot modellez. Eredménye a **modell**.

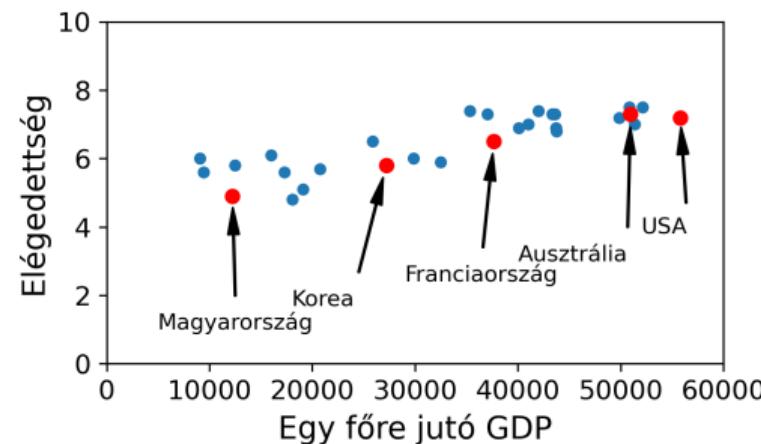
A regressziós elemzés **tárgya egy folytonos változó**.

A **felügyelt tanulás** kategóriájába tartozik, tehát a modellezés során ismertek a kívánt outputok.



# A regresszió komponensei

- **Célváltozó:** Egy jelenség, amely az elemzés tárgyát képezi. Példában: élet elégedettség.
- **Független változó:** Azok a megfigyelések, amelyek alapján a célváltozó megbecsülhető.
- **Modell:** A változók közötti feltételezett kapcsolat, ami a valóságot leírja.

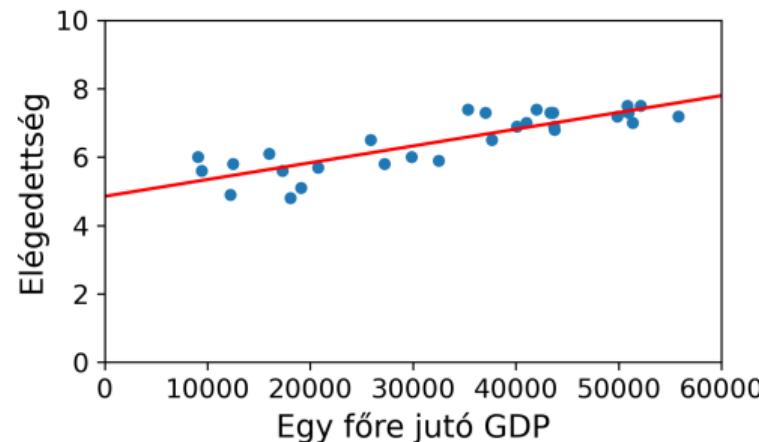


# A regresszió alkalmazása

Tipikusan akkor használatosak a regressziós eljárások, amikor a vizsgálat tárgya, hogy **egyes jelenségek hogyan befolyásolnak másokat**. Ezzel lehetséges annak a feltárása, hogy milyen **kapcsolati rendszer szerint hozhatók összefüggésbe**.

Regresszió segítségével lehetséges **valamilyen választ előre jelezni**.

Például meg lehet jósolni az egy országban élők saját életükkel való elégedettséget az ország egy főre jutó GDP mutatója alapján.

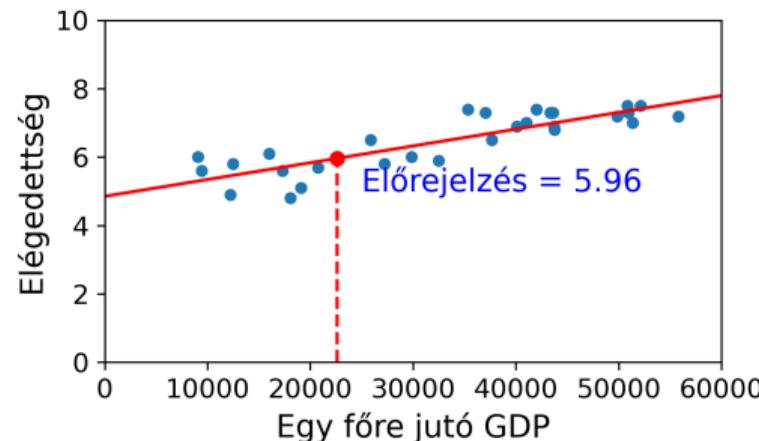


# A regresszió alkalmazása

Tipikusan akkor használatosak a regressziós eljárások, amikor a vizsgálat tárgya, hogy **egyes jelenségek hogyan befolyásolnak másokat**. Ezzel lehetséges annak a feltárása, hogy milyen **kapcsolati rendszer szerint hozhatók összefüggésbe**.

Regresszió segítségével lehetséges **valamilyen választ előre jelezni**.

Például meg lehet jósolni az egy országban élők saját életükkel való elégedettséget az ország egy főre jutó GDP mutatója alapján.



# A regresszió felírása

A lineáris regresszió célja, hogy valamely  $y$  változót megbecsülje adott

$x = (x_1, x_2, \dots, x_r)$  magyarázó változók alapján.

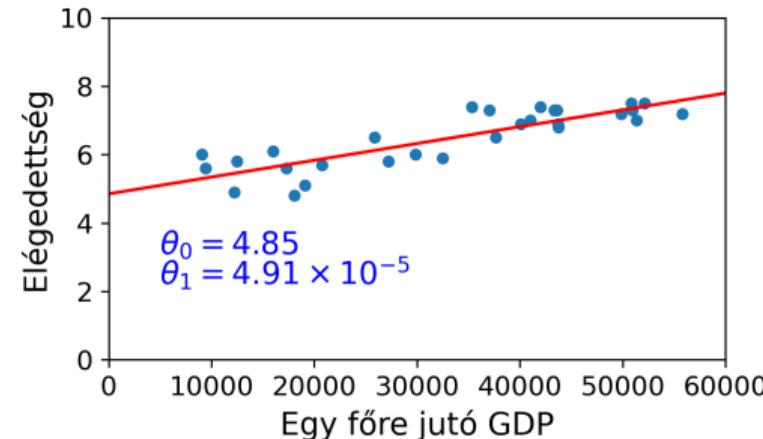
## Lineáris regresszió

A regresszió egyenlete:

$$\hat{y} = \theta_0 + \theta_1 \cdot x + \varepsilon$$

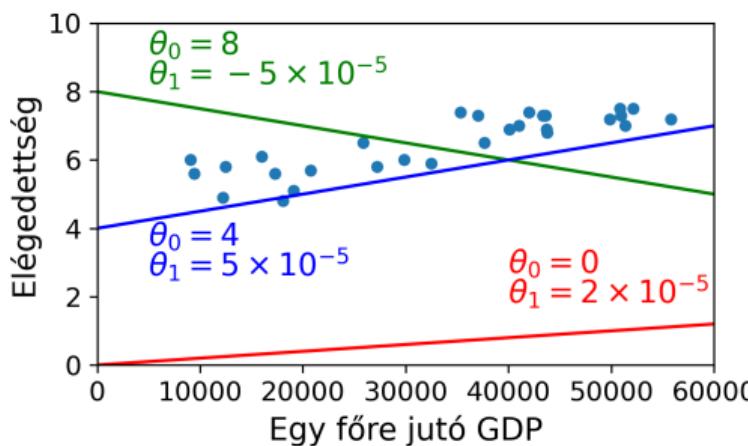
Ahol  $\hat{y}$  a becsült érték,  $\theta$  értékek a regressziós együtthatók vagy **paraméterek**,  $\varepsilon$  pedig a véletlen hiba.

Ebben az esetben  $\theta_0$  az egyenes eltolása,  $\theta_1$  pedig a meredeksége.



# A regresszió teljesítménye

A gépi tanulás esetén szükség van arra, hogy meglehessen mondani, mennyire jó a modell. Regresszió esetén a feladat a legkisebb hibához tartozó modell megkeresése.



## Reziduum

Az  $y_i$  valós érték és  $\hat{y}_i$  becsült érték távolsága adott  $d$  távolságfüggvény szerint:

$$d(y_i, \hat{y}_i)$$

## Költségfüggvény

A rezidumok összege az összes minta adatpontra:

$$L(y, \hat{y}) = \sum_{i=1}^n L(y_i, \hat{y}_i)$$

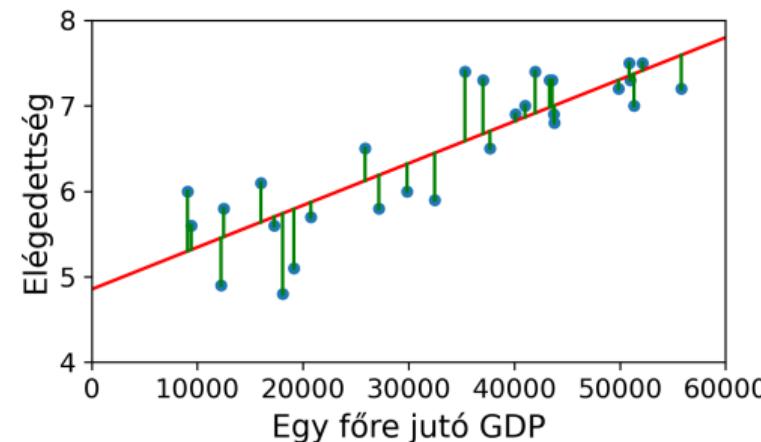
# Hiba kiszámítása MSE módszerrel

Az egyik legismertebb költségfüggvény az átlagos négyzetes hiba.

## MSE (Mean Squared Error)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Ahol  $y_i$  az aktuális minta adatpont és  $\hat{y}_i$  a regressziós modell által adott becsült érték.



1 Bevezetés

2 Regresszió

3 Optimalizáció

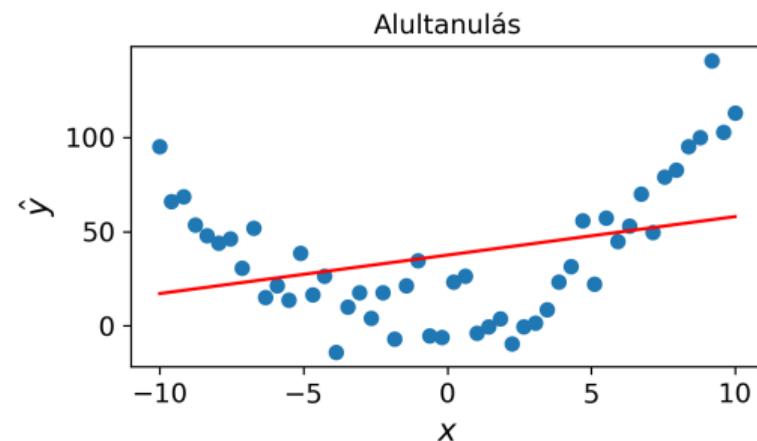
4 Gradiens ereszkedés

# Alultanulás és túltanulás

A gépi tanulásban két egymással ellentétes célra kell optimalizálni a modelleket. Ez a jó általánosító képesség és a pontos becslés.

## Alultanulás

Alultanulás esetén az illesztett modell **túlságosan általános**, nem képes hitelesen leképezni a valóságban rejlő komplex kapcsolati rendszert. Az alultanult modell **egyformán rosszul teljesít mind a tanító és előre nem látott adatokon**.

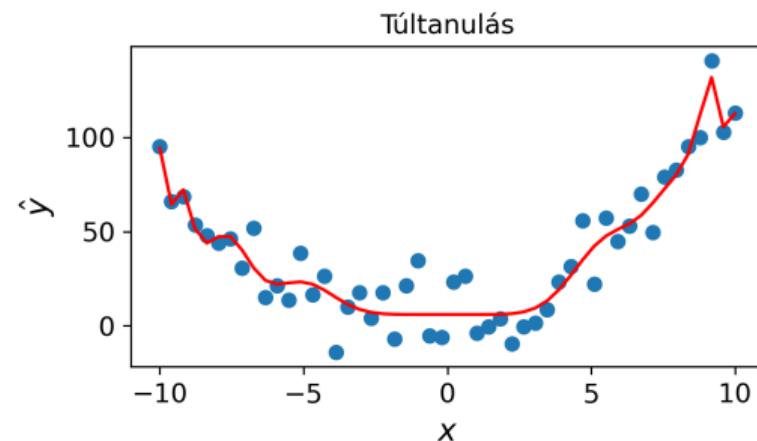


# Alultanulás és túltanulás

A gépi tanulásban két egymással ellentétes célra kell optimalizálni a modelleket. Ez a jó általánosító képesség és a pontos becslés.

## Túltanulás

Túltanulás esetén a modell nagyon pontosan, akár hiba nélkül illeszkedik a tanító adatpontokra, de **komplexitása miatt elveszíti a jó általánosító képességét**, és nem fog jól teljesíteni előre nem látott adatokon.

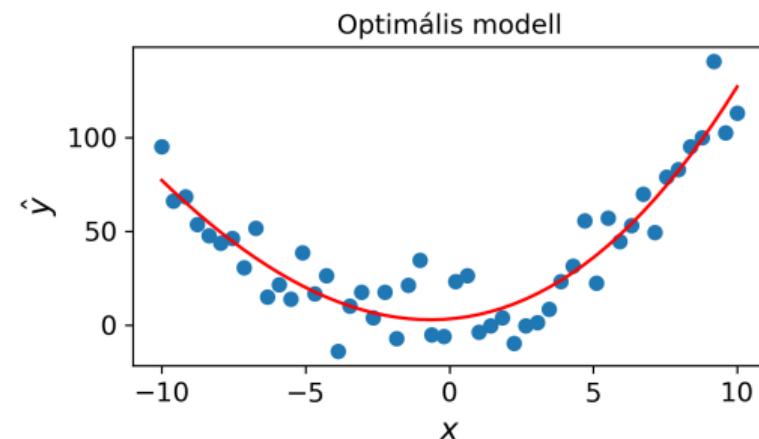


# Alultanulás és túltanulás

A gépi tanulásban két egymással ellentétes célra kell optimalizálni a modelleket. Ez a jó általánosító képesség és a pontos becslés.

## Optimális modell

Az optimális modell **egyszerre** képezi le hűen a valóság kapcsolatait és általánosít olyan módon, hogy az előre nem látott adatokon is jó előrejelzéseket legyen képes adni.



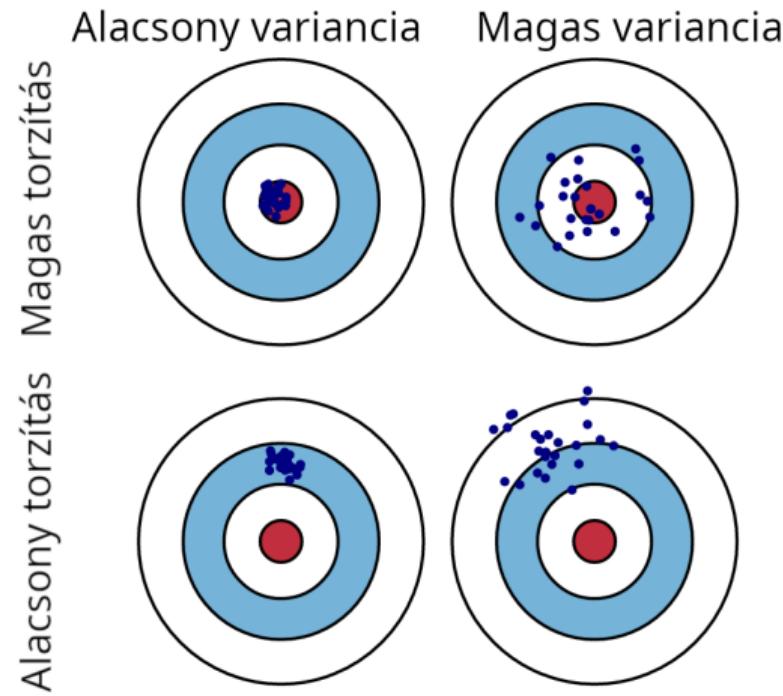
# Torzítás és variancia

A modellezés két szélsőértéke a torzítás és variancia.

## Torzítás

Ez arra utal, hogy **mennyire jól képes a modell leképezni a valós kapcsolatokat a tanító adatokon**.

Magas torzítás azt jelenti, hogy a modell túlságosan egyszerű, és nem képes megragadni az adatok összefüggéseit.



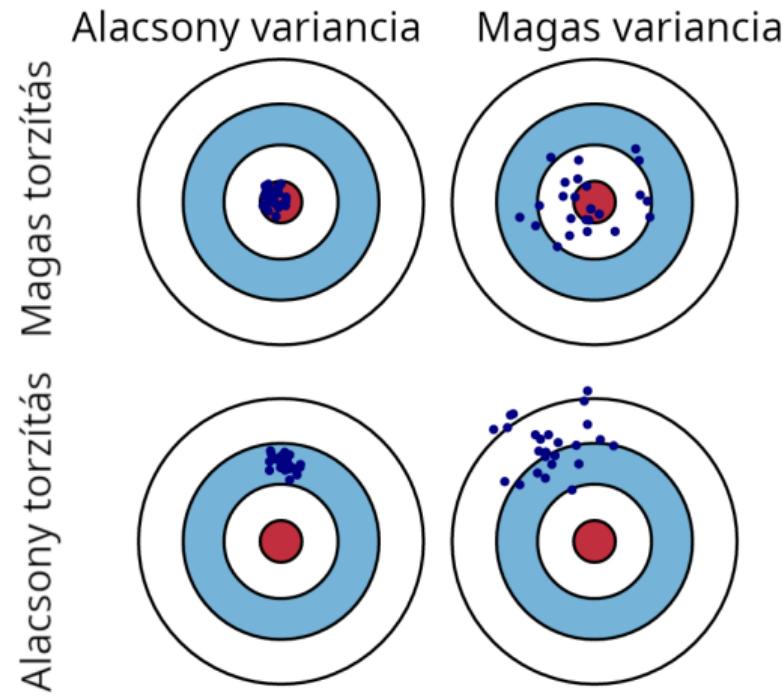
# Torzítás és variancia

A modellezés két szélsőértéke a torzítás és variancia.

## Variancia

A variancia azt mutatja meg, hogy a modell **hogyan reagál a különböző tanító adatkészletekre**.

Magas variancia esetén a modell túlzottan érzékeny a tanító adatok kis változásaira, tehát túlilleszti magát a tanító adatokra, és nem képes jól generalizálni új, látatlan adatokra.

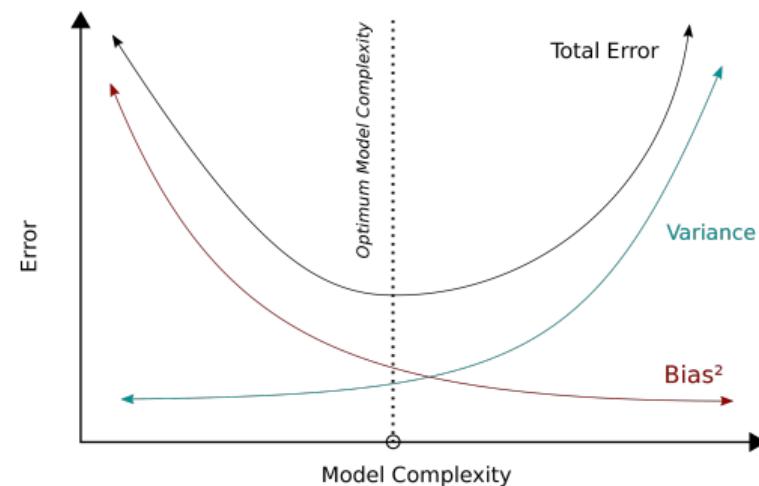


# A modell optimalitása

A modellezés során egyszerre kell úgy illeszteni a függvényt, hogy jól általánosítson, ugyanakkor a valóság kapcsolatait ne hamisítsa meg.

Amikor a tanítás elején a modell nagyon egyszerű, jól képes általánosítani, viszont a hibája magas lesz a tanító adatpontokon.

Ahogy egyre tanultabb lesz a modell a hiba csökkenni fog, viszont annál specializáltabb lesz és veszíti el az általánosító képességét.



1 Bevezetés

2 Regresszió

3 Optimalizáció

4 Gradiens ereszkedés

# Gradiens

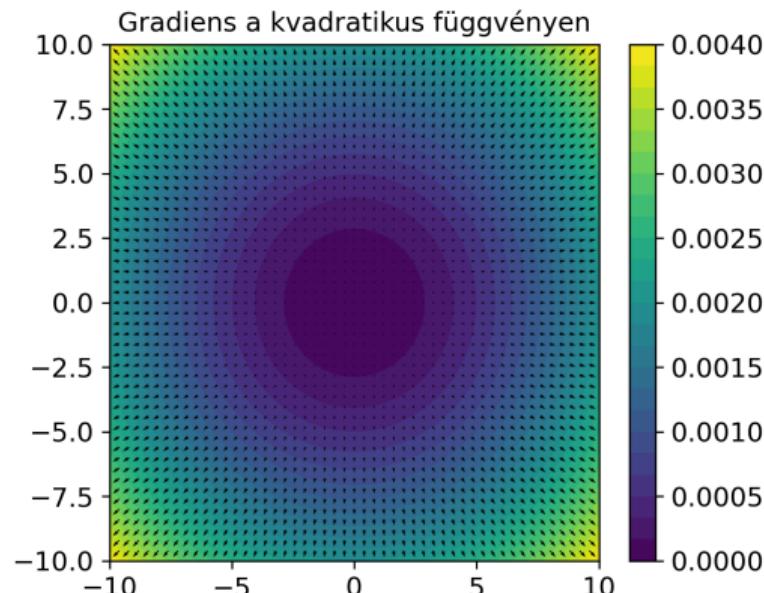
## Gradiens

Egy  $f(x, y, z)$  függvény esetén gradiens egy olyan vektor, amely arra az irányra mutat, ahol az  $f$  függvény a legmeredekebben emelkedik.

Ha  $f$  3D térben van definiálva,  $(x, y, z)$  koordinátákkal, a gradiens:

$$\nabla f = \left( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right)$$

A gradiens tetszőleges dimenziószámra általánosítható.

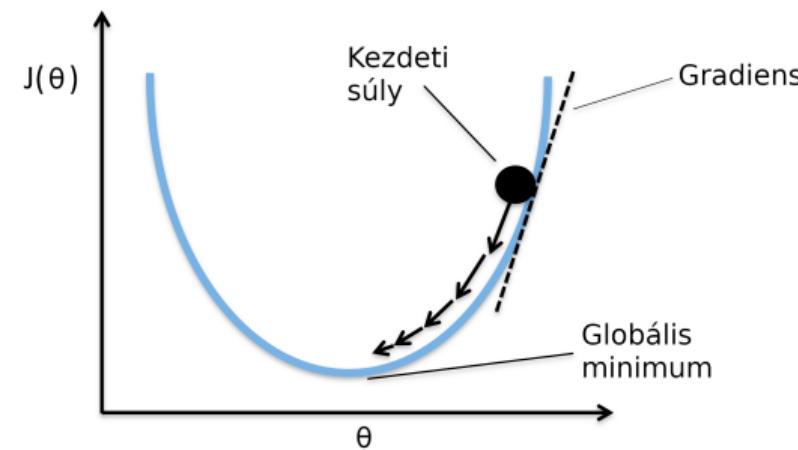


# Gradiens ereszkedés

## Gradiens ereszkedés

Iteratív optimalizálási módszer **egy célfüggvény minimum helyének megtalálására**, amely a célfüggvény gradiensét használja a keresési irány meghatározására.

Az eljárás alapvető elgondolása, hogy a függvény gradiensének ellentétes irányában haladva eljut a legkisebb értékhez, mivel a gradiens a függvény növekedésének legnagyobb irányát mutatja.



# Gradiens ereszkedés

Egy alapvető algoritmus gradiens ereszkedésre:

---

## Algoritmus 1: Gradiens ereszkedés

---

**Input:**  $\alpha, f(\theta)$

$\theta_0 \leftarrow 0;$

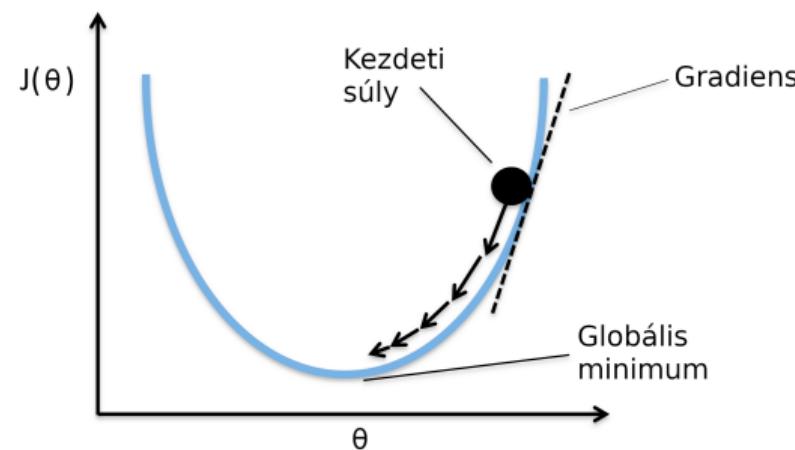
**for**  $t = 0 \rightarrow max_t$  **do**

$\nabla f(\theta_t)$  gradiens kiszámítása;  
     $\theta_{t+1} = \theta_t - \alpha \cdot \nabla f(\theta_t);$

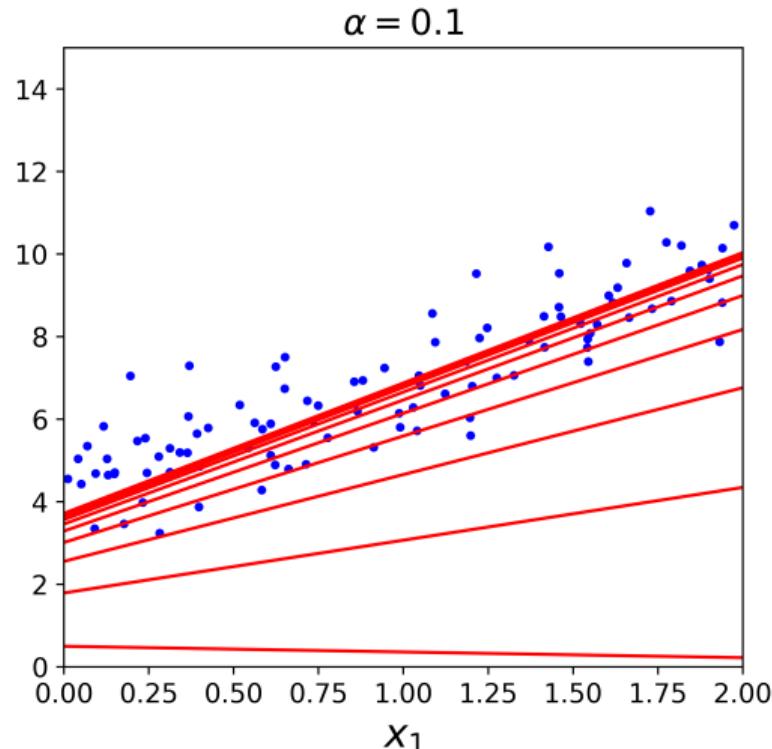
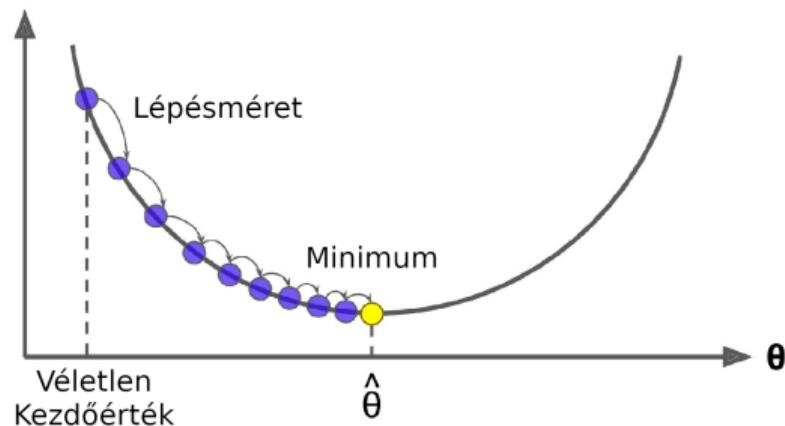
**end**

---

Ahol  $\theta$  a célfüggvényt meghatározó paraméterek vektora,  $\nabla f(\theta_t)$  a célfüggvény gradiense,  $\alpha \in [0, 1]$  a tanulási sebesség.

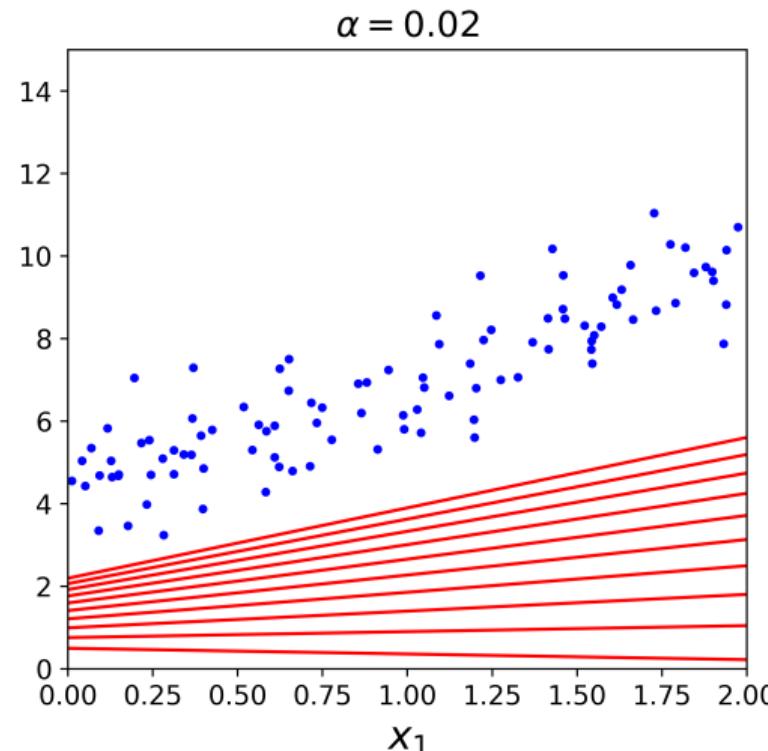
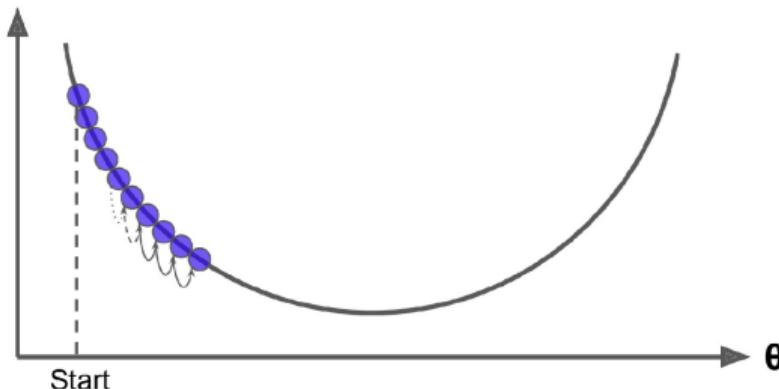


# Gradiens ereszkedés optimális esetben



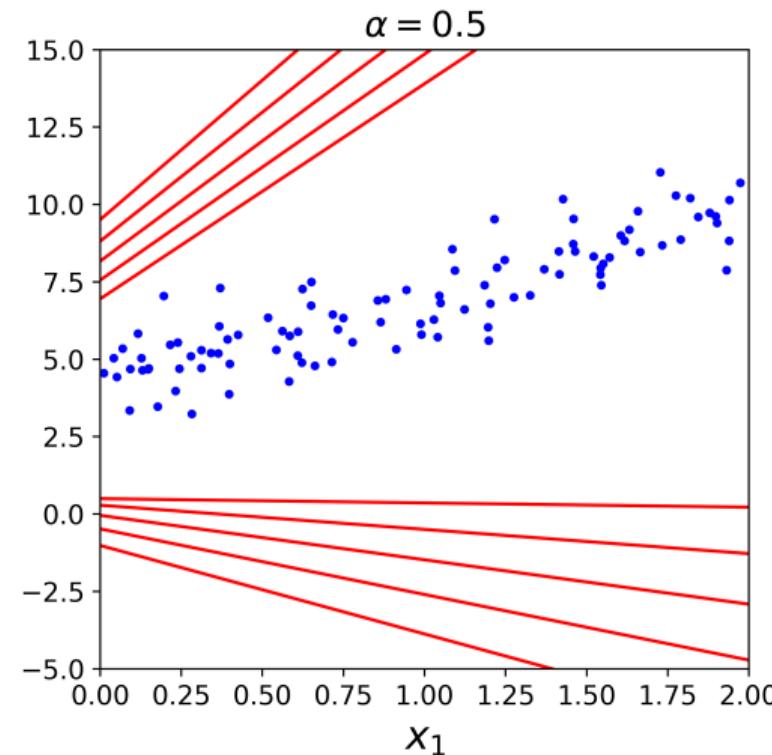
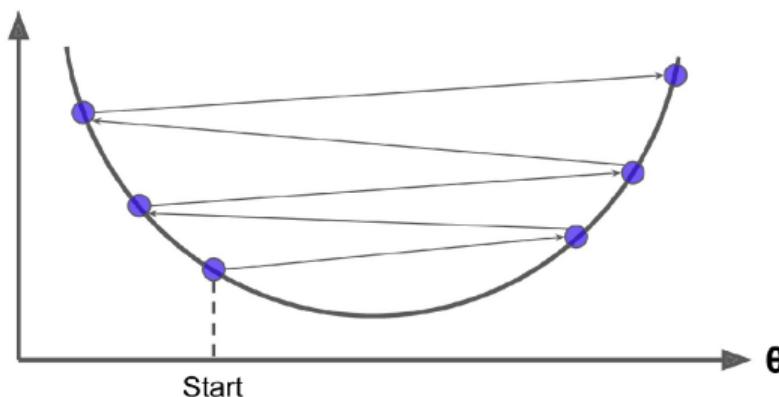
# Túl alacsony tanulási sebesség

Ha a tanulási sebesség túlságosan alacsony, az optimalizáció lehet sosem éri el a minimumot.

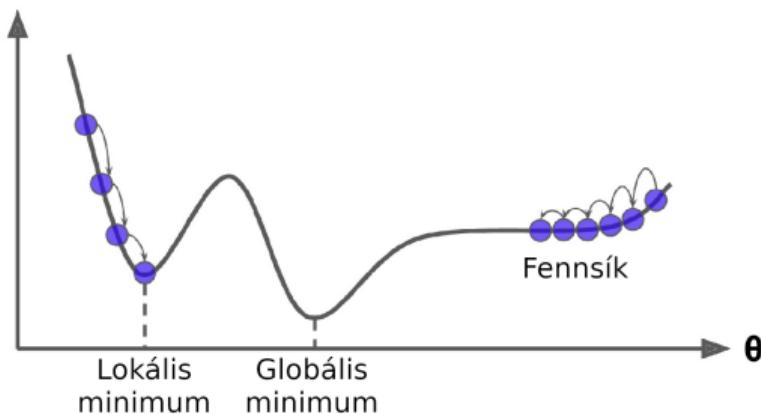


# Túl magas tanulási sebesség

Túlságosan magas tanulási sebesség esetén az algoritmus divergálhat vagy nagyon lassú lesz az optimalizáció folyamata.



# Egyéb problémák



- Az optimalizáció folyamata megakadhat egy lokális minimum helyen.
- Ha az algoritmus elér egy fennsíkot, az alacsony gradiens érték miatt instabillá válhát.