

# Regresszió, klasszifikáció

Kormányos Andor

Komplex Rendszerek Fizikája Tanszék

2020 október 19.

# Mennyit muszály tanulni egy vizsgára?

Tegyük fel, hogy egy vizsgára készülünk. Azt szeretnénk tudni, hogy mennyi időt kell minimálisan eltölteni tanulással ahhoz, hogy jó eséllyel átmenjünk a vizsgán.

# Mennyit muszály tanulni egy vizsgára?

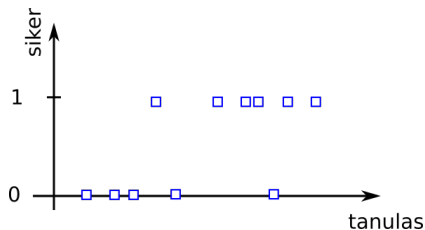
Tegyük fel, hogy egy vizsgára készülünk. Azt szeretnénk tudni, hogy mennyi időt kell minimálisan eltölteni tanulással ahhoz, hogy jó eséllyel átmenjünk a vizsgán.

Hogyan járhatunk el?

- megkérdezzük minél több felsőbb évest, hogy ők mennyit tanultak erre vizsgára
- feljegyezzük, hogy a  $i$ -k felső éves mennyit tanult :  $x^{(i)}$
- minden perc számít, ezért  $x^{(i)}$  folytonos változó
- feljegyezzük továbbá, hogy átment-e a vizsgán ( $y^{(i)} = 1$ ) vagy nem ( $y^{(i)} = 0$ )
- így kapunk egy adathalmazt:  $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}) \dots (x^{(N)}, y^{(N)})\}$

# Mennyit muszály tanulni egy vizsgára?

Mit csinálunk az adatokkal?



Stratégia:

- az adatokra próbálhatunk valamilyen függvényt illeszteni:  $h(x; \mathbf{a})$  (hipotézis),  $\mathbf{a} = (a_0, a_1 \dots a_j)$  paramétervektor
- az  $\mathbf{a}$  paramétereket valahogy meg kell határozni
- miután sikerül meghatározni  $\mathbf{a}$ -t: ha  $x_{sajat}$  időt szánunk tanulásra, akkor  $h(x_{sajat}; \mathbf{a})$  megmondja, hogy valószínűleg átmegyünk v nem a vizsgán

# További változók

A vizsgán való siker nem csak attól függhet, hogy mennyit tanultunk:

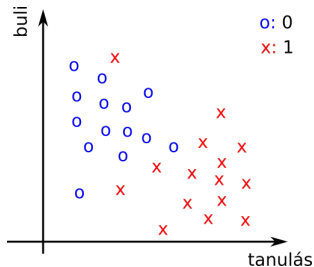
- hány órát aludtunk előző este,
- hány sört ittunk előző este, stb.

Finomíthajuk a hipotézist:

- bevezethetünk további változókat:  $x_j, j = 1, \dots, m$
- a kimenet továbbra is  $y = \{0, 1\}$
- az adathalmaz  $(x_1^{(i)}, x_2^{(i)}, \dots, x_m^{(i)}, y^{(i)}) = (\mathbf{x}^{(i)}, y^{(i)})$ ,  $i = 1 \dots N$  adatokból áll
- erre próbálhatunk illeszteni egy  $h(\mathbf{x}; \mathbf{a})$  hipotézist

# Mennyit muszály tanulni egy vizsgára?

Kétdimenziós példa:



- milyen  $h(\mathbf{x}; \mathbf{a})$  függvény írja le legjobban a két kimenet (0, 1) közötti határvonalat?
- hogyan határozhatjuk meg  $\mathbf{a}$  paramétervektort?

# Egy kis terminológia

## Regresszió vs klasszifikáció:

- ha  $y^{(i)}$  folytonos értékeket vehet fel: regressziós (illesztési) probléma (regression)
- ha  $y^{(i)} = \{0, 1\}$  csak két (általában: véges sok) diszkrét értéket vehet fel: klasszifikációs (classification) probléma (classification)

# Egy kis terminológia

## Regresszió vs klasszifikáció:

- ha  $y^{(i)}$  folytonos értékeket vehet fel: **regressziós (illesztési) probléma (regression)**
- ha  $y^{(i)} = \{0, 1\}$  csak két (általában: véges sok) diszkrét értéket vehet fel: **klasszifikációs (classification) probléma (classification)**

## Gépi tanulás

- a regressziós és klasszifikációs problémák jelentik az **gépi tanulás** egyik alterületét
- gépi tanuláson belül is az ún. **supervised learning** témakörébe tartoznak
- ez azt jelenti, hogy minden az  $x^{(i)}$ -hez adottak a  $y^{(i)}$  értékek, és a  $h(\mathbf{x}; \mathbf{a})$  hipotézisnek ezeket kell leírnia
- $x_j$ -ket **feature (jellemző)**-nek is szokták nevezni,  $y^{(i)}$ -t pedig **label**-nek
- van **unsupervised learning** is: pl adatok csoportosítása (clustering), “Coctail party algorithm”