Gépi tanulás – rövid bevezetés

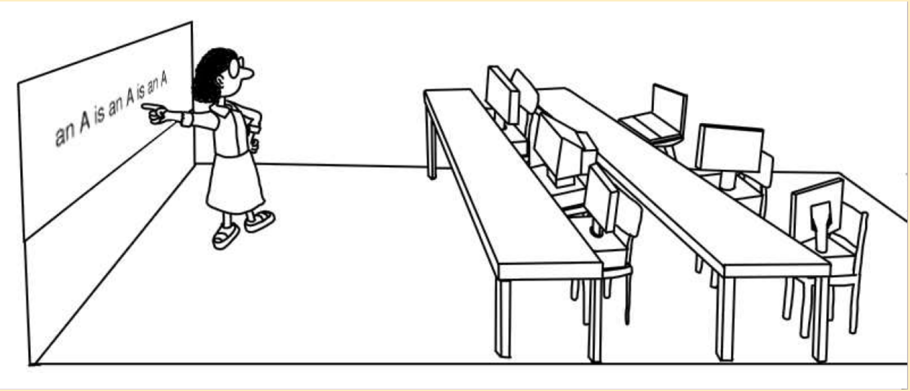
Szepesvári Csaba

MTA SZTAKI

2005 márc. 22

1. Mi a gépi tanulás?
2. Miért nehéz?
3. Ockham borotvája
4. Tanulás mint keresési probléma
5. Főbb probléma típusok
   - Klasszifikációs problémák
   - Tanuló algoritmusok
   - Regressziós problémák
   - Sűrűségfüggvény becslés
   - Altér keresés
   - Klaszterezés
6. Egy kis elmélet
7. Összefoglalás

Mi a gépi tanulás?

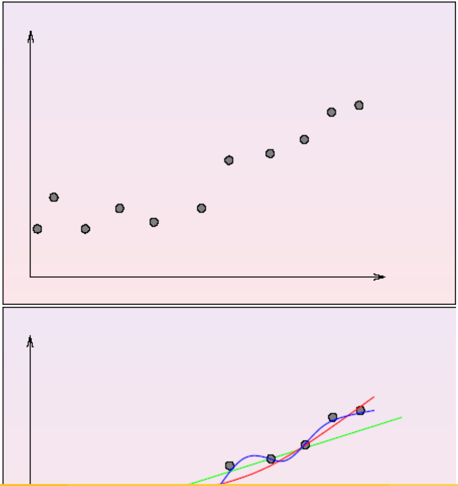


Mi a gépi tanulás?

- Tanulás: Fontos emberi tulajdonság
- ‘Jelentése’: Változás úgy, hogy a következő hasonló szituációbeli teljesítmény javuljon (valamiféle kritérium szerint).
- NEM magolás
- Magolás könnyű a számítógépnek; a nehézséget az új szituációkra való *általánosítás* jelenti

Miért nehéz?

- Véges sok tanítási minta segítségével meg kell adni egy függvényt ami kapcsolatot teremt az adatok között
- Végtelen sok lehetőségből kell választani



Ockham borotvája

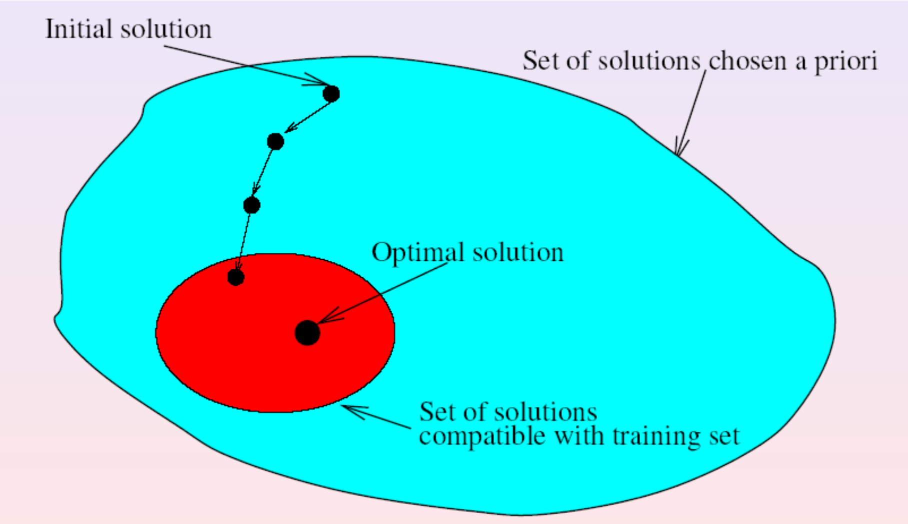
- William Ockham: 14. sz.-i szerzetes
- Egyszerűség Elve:

  „Pluralitas non est ponenda sine neccesitate”

  avagy

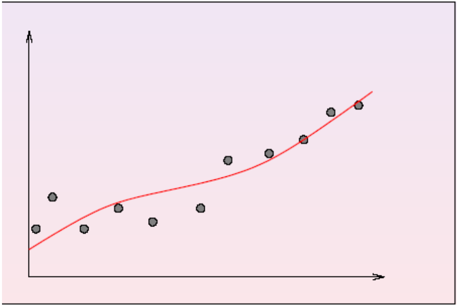
  „A többletet nem kell bevezetni szükségtelenül”
- Interpretáció: a sok lehetséges megoldásból válaszd a legegyszerűbbet
- Mi az hogy egyszerű?
- Egyszerűség: „prior” ismereteink határozzák meg, hogy mit tartunk egyszerűnek

Példa: simaság


Initial solution
Set of solutions chosen a priori
Optimal solution
Set of solutions
compatible with training set

Főbb probléma típusok

- Főbb típusok:
  - Regresszió , klasszifikáció , sűrűségfüggvény becslés
  - .. és még sokan mások:
    preferenciák tanulása, multi-osztályozás, klaszterezés, újdonság detektálás, tulajdonságkiemelés, ..



Probléma csoportok

- Felügyelt tanítás („tanár” jel):
  - regresszió
  - klasszifikáció
  - preferenciák tanulása
- Felügyelet nélküli:
  - sűrűségfüggvény becslés
  - klaszterezés
  - újdonság detektálás
  - tulajdonságkiemelés
  - altér keresés, faktor analízis, független komponens analízis
  - ...
- Megerősítéses tanulás
- On-line tanulás vs. off-line tanulás
- Rekurzív (inkrementális) vs. one-shot
- ..

Klasszifikációs problémák

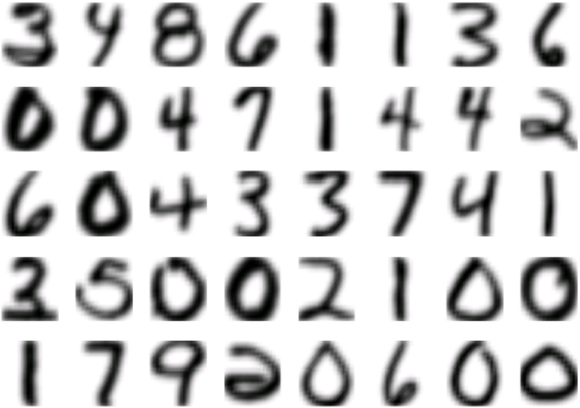
- Adatok: $(X_1, Y_1), \ldots, (X_n, Y_n)$, független, azonos eloszlásúak
- $X_i \in \mathbb{R}^d$
- $Y_i \in \{0, 1\}$ (esetleg $Y_i \in \{1, 2, \ldots, m\}$)

Cél

Adott új X példára legvalószínűbb $Y =?$ (ill. $P(Y|X) =?$).

Arcfelismerés

Karakterfelismerés



Klasszifikációs szabályok, Bayes szabály

- Klasszifikációs szabály: $g : \mathbb{R}^d \to \{0, 1\}$
- Hiba: $L(g) = P(g(X) \neq Y)$
- Bayes szabály
  - Posterior: $\eta(x) = P(Y = 1|X = x) = \mathbb{E}[Y|X = x]$
  - Bayes szabály: $g^*(x) = \mathbf{1}(\eta(x) > 1/2)$

Bayes szabály, Bayes hiba

Állítás:

Nincs más szabály, amelyik a Bayes szabálynál kisebb hibát érne el:

$$P(g^*(X) \neq Y) \leq P(g(X) \neq Y)$$

Bizonyítás.

$$L(g) = 1 - \mathbb{E}[\mathbf{1}(g(X) = 1)\eta(X) + \mathbf{1}(g(X) = 0)(1 - \eta(X))]$$

$\square$

Definíció: Bayes hiba

A Bayes szabály által elért hiba (a legkisebb elérhető hiba):

$$L^* = L(g^*).$$

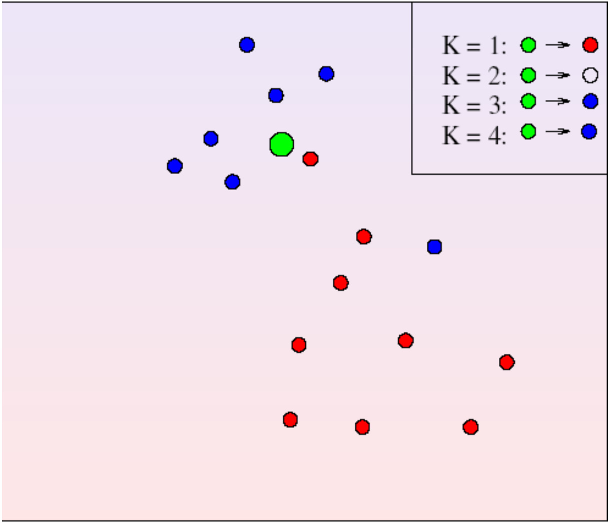
Tanuló algoritmusok

- $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.
- g_n; függ a múltbeli adatoktól: $g_n(x; D_n)$

$$L(g_n) = P(g_n(X) \neq Y | D_n) \text{ val.valtozo!}$$

- Legközelebbi szomszéd (k-NN) algoritmusok:
  - $g_n(x) = \mathbf{1}(\sum_i w_{ni}(x)\mathbf{1}(Y_i = 1) > \sum_j w_{nj}(x)\mathbf{1}(Y_j = 0))$
  - $w_{ni}(x) = 1/k$, ha i-edik pont x legközelebbi k szomszédja között van.

k-NN algoritmus



Konzisztencia

- Jel: $L_n = L(g_n)$.
- Konzisztencia:
  - Gyenge: $\mathbb{E}[L_n] \to L^*$
  - Erős: $L_n \to L^*$ w.p.1
- k-NN-re:
  - $k = k_n \to \infty$, $k_n/n \to 0$: $L_n \to L^*$ w.p.1. – feltéve, hogy X eloszlásának van sűrűségfüggvénye
  - $L_{NN} = 2\mathbb{E}[\eta(X)(1 - \eta(X))]$; k rögz., $\mathbb{E}[L_n] \to L_{NN}$.
  - $L^* \le L_{NN} \le 2L^*(1 - L^*) \le 2L^*$

„Plug-in” szabályok

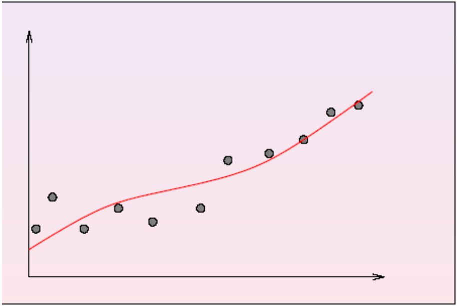
- $\eta \to \tilde{\eta}$, $g^* \to g_{\tilde{\eta}}$
- $L(g_{\tilde{\eta}}) \le L^* + 2\mathbb{E}[|\eta(X) - \tilde{\eta}(X)|]$
- $\eta_n \to \eta$, regressziós probléma!
  - $L(g_n) \le L^* + 2\mathbb{E}[|\eta_n(X) - \eta(X)|\,|D_n]$
  - $L(g_n) \le L^* + 2\mathbb{E}[|\eta_n(X) - \eta(X)|^2\,|D_n]$
- Dimenzió redukció: hagyja invariánsan L^*-ot..

Regressziós problémák

- Adatok: $(X_1, Y_1), \ldots, (X_n, Y_n)$, független, azonos eloszlásúak
- $X_i \in \mathbb{R}^d$
- $Y_i \in \mathbb{R}^m$

Cél

Adott új X példára $Y = ?$.



Példák: idősorok jóslása

Változat: $Z_1, \ldots, Z_n, \ldots$ idősor;

$$X_n = (Z_n, Z_{n-1}, \ldots, Z_{n-k}),$$

$$Y_n = Z_{n+1}$$

- Példák:
  - Holnapi árfolyamok a tőzsdén
  - Mozgó objektum dinamikája

Regressziós függvény

- Regressziós függvény: $f : \mathbb{R}^d \to \mathbb{R}^m$
- Költség (veszteség) függvény:

$$\ell(y, y') = \|y - y'\|^2$$

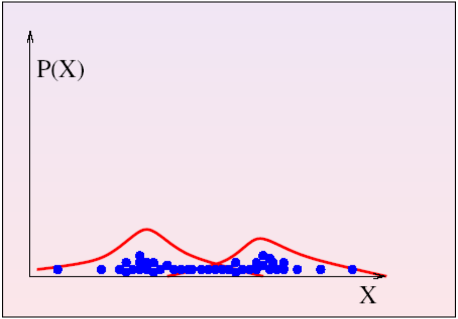
- Hiba: $L(f) = \mathbb{E}(\ell(f(X), Y))$
- „A" regressziós függvény:
  - Feltételes várható érték: $\eta(x) = \mathbb{E}[Y|X = x]$
  - Hasonlóan a Bayes szabályhoz, ez is optimális.

Sűrűségfüggvény becslés

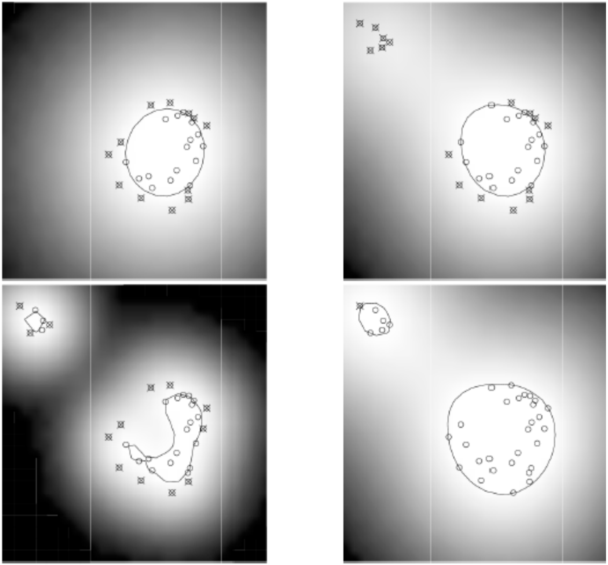
- Adatok: $(X_1, \ldots, X_n)$, független, azonos eloszlásúak
- $X_i \in \mathbb{R}^d$

Cél

Adott új X példára $P(X) =?$ (ill. $P(X) > p_0$?)



Újdonság detektálás

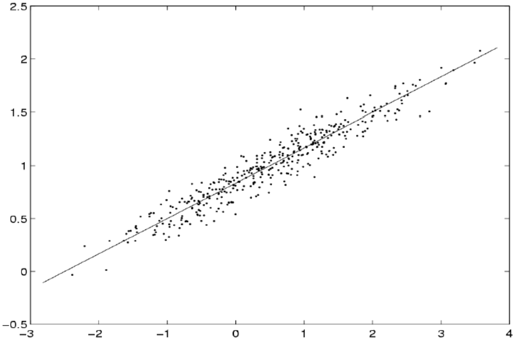


Altér keresés

Az adatok valamely alacsonyabb dimenziós sokaságon „élnek” – eltekintve a zajtól. Találjuk meg ezeket!

Cél

Lehető legkisebb torzítás: $\mathbb{E}[\ell(P^T X, X)] \to \min$



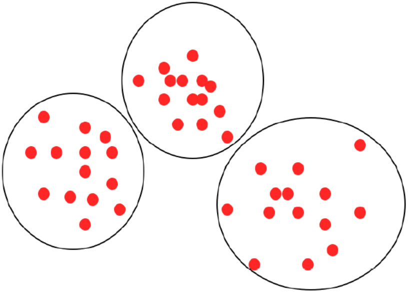
Klaszterezés

Találjuk meg a csoportokat az adatokban!

- Mitől csoport a csoport? („Távolság"/„Közelség")
- Mitől jó a csoportosítás?
- Hány csoport?

Cél

Lehető legkisebb torzítás: $\mathbb{E}[\ell(c(X), X)] \to \min$



Adatok

Tanításkor rendelkezésre álló adatok:

- $D_n = (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$
  - független, azonos eloszlású (iid)
- egy ismeretlen $p(Z)$ eloszlásból dobtuk

Alesetek

- Klasszifikáció:
$$Z = (X, Y) \in \mathbb{R}^d \times \{0, 1\}$$
- Regresszió:
$$Z = (X, Y) \in \mathbb{R}^d \times \mathbb{R}$$
- Sűrűségfüggvény becslés:
$$Z \in \mathbb{R}^d$$

Tanulás: találjunk meg egy jó függvényt az $\mathcal{F}$ függvénytérben.

Példák; $f(\cdot;\theta) \in \mathcal{F}$

- Klasszifikáció:

$$y = \mathrm{sgn}(ax^2 + bx + c); \quad \theta = (a, b, c)$$

- Regresszió:

$$y = ax^2 + bx + c; \quad \theta = (a, b, c)$$

- Sűrűségfüggvény becslés:

$$p(z) = p(z; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(z-\mu)^T \Sigma^{-1}(z-\mu)\right)$$

Költség/veszteség/hiba függvény

Tanulás: találjunk meg egy jó függvényt az $\mathcal{F}$ függvénytérben.

Hiba költsége: $L : \mathcal{Z} \times \mathcal{F} \to \mathbb{R}$;

- Klasszifikáció:

$$L((x,y),f) = \ell(y, f(x)) = \begin{cases} 0, & \text{ha } f(x) = y \\ 1, & \text{ha } f(x) \neq y. \end{cases}$$

- Regresszió:

$$L((x,y),f) = \ell(y, f(x)) = \frac{1}{2}(y - f(x))^2$$

- Sűrűségfüggvény becslés:

$$L(z,f) = -\log(f(z))$$

Várható hiba, empirikus hiba

Tanulás: találjunk meg egy jó függvényt az $\mathcal{F}$ függvénytérben.

- Kritérium: f használatából eredő várható hiba minimalizálása:

$$R(f) = \mathbb{E}[L(Z, f(Z))]$$

- Indukciós elv:
  - $f^* = \mathrm{argmin}_{f \in \mathcal{F}} R(f)$
  - Probléma: Z eloszlása nem ismert és különben is egy integrált minimalizálni nem könnyű!
- Empirikus hiba:

$$\hat{R}(f; D_n) = \frac{1}{n} \sum_{i=1}^{n} L(z_i, f).$$

Várható hiba, empirikus hiba

- Az empirikus hiba a várható hiba *torzítatlan* becslése:

$$\mathbb{E}[\hat{R}(f;D)] = R(f)$$

- Empirikus hiba-minimalizációs elv:

$$f^*(D) = \mathrm{argmin}_{f\in\mathcal{F}} \hat{R}(f;D)$$

- Hiba a tanító mintákon:

$$\hat{R}(f^*(D), D) \quad (= \min_{f\in\mathcal{F}} \hat{R}(f;D))$$

A tanítási hiba torzítatlan, vagy torzított becslése az empirikus hibának?

Tanítási hiba

A tanítási hiba torzítatlan, vagy torzított becslése az empirikus hibának?
Torzított.

$$\begin{aligned}\mathbb{E}[R(f^*(D)) - \hat{R}(f^*(D), D)] &= \mathbb{E}[R(f^*(D)) - \min_{f\in\mathcal{F}} \hat{R}(f, D)] \\ &\geq \max_{f\in\mathcal{F}} \mathbb{E}[R(f^*(D)) - \hat{R}(f, D)] \geq 0\end{aligned}$$

- A D segítségével talált megoldás jobb teljesítményt ad D-n, mint egy másik véletlenül sorsolt D' adathalmazon.
- Tudjuk-e felülről korlátozni a tanítási és a várható hiba különbségét?

$$|R(f^*(D)) - \hat{R}(f^*(D), D)| \leq ?$$

- Válasz: ha $\mathcal{F}$ kellően megszorított, akkor igen
- A korlát az $\mathcal{F}$ kapacitásától (méretétől) függ.

Összefoglalás

- Tanulás az adatokból
- Milyen feladatok?
  - Felügyelt: klasszifikáció, regresszió, ..
  - Felügyelet nélküli: sűrűségfüggvény becslés, altér becslés, klaszterezés, ..
- Empirikus hiba, várható hiba, empirikus hiba-minimalizáció
- Legközelebb:
  - Kapacitás, bias-variance dilemma, regularizáció, modell szelekció
  - Egy kis módszertan: Tanuló algoritmusok összehasonlítása
  - Paraméterbecslés: MLE, MAP, Bayes
  - Klasszifikációs módszerek (k-NN, naive Bayes, döntési fák, SVM)