Modell választás

Kormányos Andor

Komplex Rendszerek Fizikája Tanszék

2021. március 1.

Mennyire jó a modell?[1]

Mikor jó egy modell?

- nem elég, hogy jól illeszkedik az adatokhoz (a $J(\mathbf{a}_{fit}; x^{(i)}, y^{(i)})$ költségfüggvény értéke kicsi)
- emellett még az **általánosítási hibája** is kicsi
- vagyis új adatok esetén is prediktív

Mit jelent ez?

- detektálni akarjuk, hogy underfittinghez vagy overfittinghez vezet-e a modell
- sokszor a kettő között nem nyilvánvaló a határ
- egyszerű esetben a modell és az adatok ábrázolása segíthet
- többváltozós modell esetén az ábrázolás nem feltétlenül segít

[1] Andrew Ng, Coursera alapján

Modell kiértékelése

Első lépés:

- előfeltétel, hogy ne csak néhány adatpont legyen
- osszuk az adatokat kétfelé:
  - **training halmaz** $\{(\mathbf{x}_{tr}^{(i)}, y_{tr}^{(i)})\}$, pl eredeti adathalmaz 70%
  - **validációs halmaz** $\{(\mathbf{x}_{v}^{(i)}, y_{v}^{(i)})\}$, pl eredeti adathalmaz 30%
- az eredeti adathalmazból egyenletes valószínűséggel válasszunk a training és a validációs halmazba!

Modell kiértékelése

Első lépés:

- előfeltétel, hogy ne csak néhány adatpont legyen
- osszuk az adatokat kétfelé:
  - **training halmaz** $\{(\mathbf{x}_{tr}^{(i)}, y_{tr}^{(i)})\}$, pl eredeti adathalmaz 70%
  - **validációs halmaz** $\{(\mathbf{x}_{v}^{(i)}, y_{v}^{(i)})\}$, pl eredeti adathalmaz 30%
- az eredeti adathalmazból egyenletes valószínűséggel válasszunk a training és a validációs halmazba!

Második lépés:

- a $\{(\mathbf{x}_{tr}^{(i)}, y_{tr}^{(i)})\}$ segítségével kiszámolt $J(\mathbf{a}; \mathbf{x}^{(i)}, y^{(i)})$ minimalizáljuk
- így kapjuk az $\mathbf{a}_{fit}$ paramétereket

Modell kiértékelése

Első lépés:

- előfeltétel, hogy ne csak néhány adatpont legyen
- osszuk az adatokat kétfelé:
  - **training halmaz** $\{(\mathbf{x}_{tr}^{(i)}, y_{tr}^{(i)})\}$, pl eredeti adathalmaz 70%
  - **validációs halmaz** $\{(\mathbf{x}_{v}^{(i)}, y_{v}^{(i)})\}$, pl eredeti adathalmaz 30%
- az eredeti adathalmazból egyenletes valószínűséggel válasszunk a training és a validációs halmazba!

Második lépés:

- a $\{(\mathbf{x}_{tr}^{(i)}, y_{tr}^{(i)})\}$ segítségével kiszámolt $J(\mathbf{a}; \mathbf{x}^{(i)}, y^{(i)})$ minimalizáljuk
- így kapjuk az $\mathbf{a}_{fit}$ paramétereket

Harmadik lépés:

- a $\{(\mathbf{x}_{v}^{(i)}, y_{v}^{(i)})\}$ és az előző lépésben kapott $\mathbf{a}_{fit}$ felhasználásával számoljuk ki a $J(\mathbf{a}_{fit}; \mathbf{x}_{v}^{(i)}, y_{v}^{(i)})$
- ez az ún **validációs hiba**

Hogyan válasszunk modellt?

- olyan modellt szeretnénk, amely prediktív lehet új adatokra is
- vagyis kicsi az általánosítási hibája
- kvantifikáljuk az általánosítási hibát

Egy egyszerű példa: hányad fokú polinomot tartalmazó modellt válasszunk?

$$\begin{aligned}
h_1(\mathbf{a}^{(1)}; x) &= a_{10} + a_{11}x \\
h_2(\mathbf{a}^{(2)}; x) &= a_{20} + a_{21}x + a_{22}x^2 \\
h_3(\mathbf{a}^{(3)}; x) &= a_{30} + a_{31}x + a_{32}x^2 + a_{33}x^3 \\
h_4(\mathbf{a}^{(4)}; x) &= a_{40} + a_{41}x + a_{42}x^2 + a_{43}x^3 + a_{44}x^4
\end{aligned}$$

Hogyan válasszunk modellt?

Egy egyszerű eljárás, amely

- nem csak lineáris illesztés esetén alkalmazható
- nem függ a költségfüggvény konkrét alakjától (nem csak legkisebb négyzetek esetén)

Lépések:

- a training set $\{(x_{tr}^{(i)}, y_{tr}^{(i)})\}$ segítségével meghatározzuk az illesztett paramétervektorokat minden modellre: $\mathbf{a}_{fit}^{(1)}$, $\mathbf{a}_{fit}^{(2)}$, $\mathbf{a}_{fit}^{(3)}$, $\mathbf{a}_{fit}^{(4)}$
- kiszámoljuk a költségfüggvényt minden modellre a validációs halmaz $\{(x_v^{(i)}, y_v^{(i)})\}$ segítségével: $J(\mathbf{a}_{fit}^{(j)}; x_v^{(i)}, y_v^{(i)})$
- azt a modellt választjuk, amelyre a $J(\mathbf{a}_{fit}^{(j)}; x_v^{(i)}, y_v^{(i)})$ a legkisebb

Hogyan válasszunk modellt?

Ha a különböző modellek az illesztett polinom fokában különböznek:

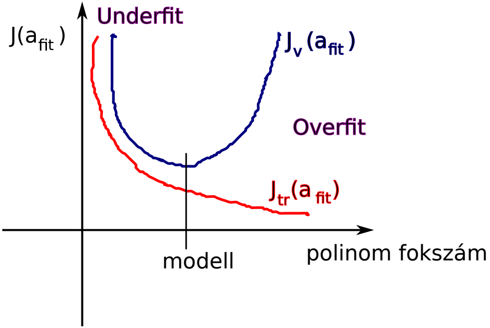

Figure: A költségfüggvény viselkedése különböző modellek esetén. Jelölés: $J_{tr}(\mathbf{a}_{fit}) = J(\mathbf{a}_{fit}^{(j)}; x_{test}^{(i)}, y_{test}^{(i)})$, $J_v(\mathbf{a}_{fit}) = J(\mathbf{a}_{fit}^{(j)}; x_v^{(i)}, y_v^{(i)})$

Mekkora a választott modell általánosítási hibája?

- ha a k-ik modellt választottuk, akkor azt gondolhatjuk, hogy a $J(\mathbf{a}_{fit}^{(k)}; \mathbf{x}_v^{(i)}, y_v^{(i)})$ jellemzi a hibát
- szigorúan véve ez nem egy jó mértéke az általánosítási hibának
- mivel a $\{(x_v^{(i)}, y_v^{(i)})\}$-t használtuk a modell kiválasztására, ezért ezen az adathalmazon számolt $J(\mathbf{a}_{fit}^{(k)}; x_v^{(i)}, y_v^{(i)})$ valószínűleg alulbecsüli az általánosítási hibát

Mekkora a választott modell általánosítási hibája?

- ha a k-ik modellt választottuk, akkor azt gondolhatjuk, hogy a $J(\mathbf{a}_{fit}^{(k)}; \mathbf{x}_v^{(i)}, y_v^{(i)})$ jellemzi a hibát
- szigorúan véve ez nem egy jó mértéke az általánosítási hibának
- mivel a $\{(x_v^{(i)}, y_v^{(i)})\}$-t használtuk a modell kiválasztására, ezért ezen az adathalmazon számolt $J(\mathbf{a}_{fit}^{(k)}; x_v^{(i)}, y_v^{(i)})$ valószínűleg alulbecsüli az általánosítási hibát

Korrekt eljárás:

- három adathalmaz képzése az eredeti adatokból:
  - training adatok $\{(x_{tr}^{(i)}, y_{tr}^{(i)})\}$ (60%)
  - validációs adatok $\{(x_v^{(i)}, y_v^{(i)})\}$ (20%)
  - teszt adatok $\{(x_{test}^{(i)}, y_{test}^{(i)})\}$ (20%)
- $\{(x_{tr}^{(i)}, y_{tr}^{(i)})\}$: $\mathbf{a}_{fit}^{(j)}$ fit meghatározása minden modellre
- $\{(x_v^{(i)}, y_v^{(i)})\}$: modell kiválasztása $J(\mathbf{a}_{fit}^{(j)}; x_v^{(i)}, y_v^{(i)})$ számolásával
- $\{(x_{test}^{(i)}, y_{test}^{(i)})\}$: a kiválasztott modell $\mathbf{a}_{fit}^{(k)}$ paramétereivel kiszámoljuk a $J(\mathbf{a}_{fit}^{(k)}; x_{test}^{(i)}, y_{test}^{(i)})$ **általánosítási hibát**