1. Lineárna regresia

Tomáš Vinař

2-INF-150: Strojové učenie, Leto 2009

Upozornenie: Tieto poznámky obsahujú len hrubý obsah materiálu z prednášky a nie sú náhražkou vlastných poznámok!

1 Motivačný príklad

Snažíme sa predpovedať ceny bytov z ich podlahovej plochy.

Označenie:

$x_i \in \mathbb{R}^n$	n vstupných atribútov (features)	napr. podlahová plocha (1 atribút)
$y_i \in \mathbb{R}$	cieľová premenná (target variable)	napr. predajná cena bytu
$(x_1, y_1) \dots (x_t, y_t)$	trénovacia množina (t príkladov)	napr. ceny a podlahové plochy bytov predaných za posl. mesiac

Hľadáme funkciu $h : \mathbb{R}^n \to \mathbb{R}$ (*hypotézu*), kde $h(x)$ "dobre predpovedá" y.

Úloha: Daná je *trénovacia množina* $(x_1, y_1) \dots (x_t, y_t)$ a *priestor hypotéz* $H = \{h : \mathbb{R}^n \to \mathbb{R}\}$, nájdite hypotézu $h \in H$, pre ktorú $\sum_{i=1}^{t} \mathrm{err}(h(x_i), y_i)$ je najmenší možný.

Príklad: 1 atribút (napr. podlahová plocha), uvažujeme hypotézy tvaru: $h_\theta(x) = \theta_0 + \theta_1 x$, chybovú funkciu $\mathrm{err}(\hat{y}, y) = (\hat{y} - y)^2$.
Hľadáme parametre $\theta = (\theta_0, \theta_1)$ minimalizujúce:

$$J(\theta) = \sum_{i=1}^{t} \mathrm{err}(h_\theta(x_i), y_i) = \sum_{i=1}^{t} (h_\theta(x_i) - y_i)^2 = \sum_{i=1}^{t} (\theta_0 + \theta_1 x_i - y_i)^2 \tag{1}$$

V minime musia byť všetky *parciálne derivácie* nulové:

$$\begin{aligned} \frac{\partial J(\theta)}{\partial \theta_0} = \sum_{i=1}^{t} 2(\theta_0 + \theta_1 x_i - y_i) &= 0 \\ t\theta_0 + \theta_1 \sum x_i - \sum y_i &= 0 \qquad (2) \\ \frac{\partial J(\theta)}{\partial \theta_0} = \sum_{i=1}^{t} 2(\theta_0 + \theta_1 x_i - y_i) x_i &= 0 \\ \theta_0 \sum x_i + \theta_1 \sum x_i^2 - \sum x_i y_i &= 0 \qquad (3) \end{aligned}$$

Riešením tejto sústavy dvoch lineárnych rovníc o dvoch neznámych dostávame:

$$\theta_0 = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{t \sum x_i^2 - (\sum x_i)^2} \tag{4}$$

$$\theta_1 = \frac{t \sum x_i y_i - \sum x_i \sum y_i}{t \sum x_i^2 - (\sum x_i)^2} \tag{5}$$

Zhrnutie:

- **Trénovanie:** Použitím trénovacej množiny $(x_1, y_1), \ldots, (x_t, y_t)$ natrénujeme parametre $\theta = (\theta_0, \theta_1)$ klasifikačnej funkcie $h_\theta(x) = \theta_0 + \theta_1 x$, ktoré minimalizujú trénovaciu chybu $J(\theta)$; riešenie v uzavretom tvare: viď. rovnice (4)-(5).
- **Predpoveď:** Na novom príklade charakterizovanom vstupným atribútom x predpovieme hodnotu cieľovej premennej ako $\hat{y} = h_\theta(x)$, pričom θ označuje parametre natrénované v predchádzajúcom kroku.

2 Viacero atribútov a normálne rovnice

Ak každý príklad x je charakterizovaný n atribútmi $(x_1, \ldots, x_n)$, môžeme trénovaciu množinu t príkladov zapísať pomocou matice a vektora:

$$X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ & \ldots & & \\ x_{t1} & x_{t2} & \ldots & x_{tn} \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_t \end{bmatrix},$$

alebo tiež zjednodušene:

$$X|y = \left[\begin{array}{cccc|c} x_{11} & x_{12} & \ldots & x_{1n} & y_1 \\ x_{21} & x_{22} & \ldots & x_{2n} & y_2 \\ & \ldots & & & \\ x_{t1} & x_{t2} & \ldots & x_{tn} & y_t \end{array}\right].$$

Pre jednoduchosť predpokladáme, že prvý atribút má **vždy hodnotu 1** (t.j., $x_{i1} = 1$ pre všetky i).

Hľadáme funkciu $h_\theta(x)$ z množiny hypotéz $H = \{h_\theta : x \to x^T\Theta\}$, kde θ je vektor n parametrov $(\theta_1, \theta_2, \ldots, \theta_n)$. [1]

Hypotéza $h_\theta(x)$ má pritom minimalizovať trénovaciu chybu definovanú ako:

$$J(\theta) = \sum_{i=1}^{t} (h_\theta(x_i) - y_i)^2 = (X\theta - y)^T (X\theta - y) \tag{6}$$

Použité definície a vzťahy z viacrozmernej analýzy a maticovej algebry:

Ak A je matica $m \times n$, potom *gradient* funkcie mn premenných $f(A)$ je definovaný ako:

$$\nabla_\theta f(A) =_{\text{def}} \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \frac{\partial f}{\partial A_{12}} & \ldots & \frac{\partial f}{\partial A_{1n}} \\ \frac{\partial f}{\partial A_{21}} & \frac{\partial f}{\partial A_{22}} & \ldots & \frac{\partial f}{\partial A_{2n}} \\ & \ldots & & \\ \frac{\partial f}{\partial A_{m1}} & \frac{\partial f}{\partial A_{m2}} & \ldots & \frac{\partial f}{\partial A_{mn}} \end{bmatrix}. \tag{7}$$

Stopa štvorcovej matice

$$\text{tr}A =_{\text{def}} \sum_i A_{ii} \tag{8}$$

Niektoré užitočné vzťahy maticového počtu:

Pre vektor z:

$$z^T z = \sum_i z_i^2 \tag{9}$$

$$\text{tr}A + B = \text{tr}A + \text{tr}B \tag{10}$$

$$\nabla_A \text{tr} A^T BAC = B^T AC^T + BAC \tag{11}$$

$$\nabla_A \text{tr} BA = B^T \tag{12}$$

[1] V tomto texte sú vektory veľkosti n ekvivalentné matici s $n \times 1$ (stĺpcový variant). To znamená, že $x^T\Theta = \sum_i x_i\theta_i$.

Odvodenie tzv. *normálnych rovníc* pre lineárnu regresiu: Tak ako v predchádzajúcej kapitole, nutnou podmienkou aby daná sada parametrov θ minimalizovala funkciu $J(\theta)$ je, že jej všetky parciálne derivácie musia byť nulové, iným slovom povedané:

$$\nabla_\theta J(\theta) = 0 \tag{13}$$

Jednoduchými algebraickými úpravami a zo vzťahov uvedených vyššie:

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \underbrace{(X\theta - y)^T}_{\theta^T X^T - y^T}(X\theta - y) \\
&= \nabla_\theta \underbrace{\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta + y^T y}_{\text{skalár}} \\
&= \nabla_\theta \mathrm{tr}(\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta + y^t y) \\
&= \nabla_\theta \mathrm{tr}\theta^T X^T X\theta - \nabla_\theta 2\mathrm{tr}y^T X\theta | \nabla_\theta \mathrm{tr}y^T y \\
&= X^T X\theta + X^T X\theta - 2X^T y\theta \\
&= 2(X^T X\theta - X^T y). \qquad (14)
\end{aligned}$$

Dohromady použitím (13) a (14) dostávame:

$$X^T X\theta = X^T y \tag{15}$$

Táto rovnica predstavuje vlastne systém n lineárnych rovníc o n neznámych, pričom ich vyriešením získame parametre $\theta = (\theta_1, \ldots, \theta_n)$ pre regresnú funkciu $h_\theta(x)$. Tento systém rovníc tiež voláme *normálne rovnice lineárnej regresie*.

Zhrnutie:

- **Trénovanie:** Použitím trénovacej množiny o t príkladoch s n atribútmi (pričom prvý atribút je vždy 1), charakterizovanej maticou X a cieľovým vektorom y, zostavíme systém n lineárnych rovníc o n neznámych $\theta = (\theta_1, \ldots, \theta_n)$ podľa vzťahu (15). Riešením tohto systému získame parametre funkcie $h_\theta(x) = x^T\theta$, ktorá minimalizuje trénovaciu chybu $J(\theta)$.
- **Predpoveď:** Na novom príklade charakterizovanom vektorom n vstupných atribútov $x = (x_1 = 1, x_2, \ldots, x_n)$ predpovieme hodnotu cieľovej premennej ako $\hat{y} = h_\theta(x)$, pričom θ označuje parametre natrénované v predchádzajúcom kroku.

Časová zložitosť:

- Zostavenie sústavy rovníc (násobenie matíc): $O(n^2 t)$
- Riešenie sústavy n rovníc o n neznámych: $O(n^3)$
- Spolu trénovanie: $O(n^2 t + n^3)$
- Predpoveď: $O(n)$

3 Iné chybové funkcie

Doteraz sme používali chybovú funkciu $J(\theta)$, ktorá bola sumou štvorcov chýb na jednotlivých trénovacích príkladov (tzv. *metóda najmenších štvorcov*). Táto funkcia sa tiež označuje ako L_2 norma, keďže vlastne optimalizujeme euklidovskú vzdialenosť cieľového vektora všetkých trénovacích príkladov od vektora predpovedaných hodnôt na všetkých trénovacích príkladoch.
Čo ale iné chybové funkcie?

3.1 Norma L_1

Uvažujme nasledujúcu chybovú funkciu (tzv. *Manhattanovská vzdialenosť*):

$$J(\theta) = \sum_i |h_\theta(x_i) - y_i| \tag{16}$$

Trénovanie môžeme riešiť pomocou *lineárneho programovania*:

vypočítaj θ, δ, z
minimalizujúce $\sum_i z_i$
pričom pre všetky i:
$x_i^T\theta + \delta_i = y_i$
$z_i \geq \delta_i$
$z_i \geq -\delta_i$

3.2 Norma L_∞

Uvažujme chybovú funkciu, ktorá minimalizuje najhoršiu chybu:

$$J(\theta) = \max_i |h_\theta(x_i) - y_i| \tag{17}$$

Trénovanie môžeme riešiť opäť pomocou *lineárneho programovania*:

vypočítaj θ, δ
minimalizujúce δ
pričom pre všetky i:
$y_i - \delta \leq x_i^T\theta \leq y_i + \delta$

4 Generalizovaná lineárna regresia

- Množinu n atribútov "preložíme" pomocou m bázových (obvykle nelineárnych) funkcií $\phi = (\phi_1, \phi_2, \ldots, \phi_m)$, kde každá funkcia je $\phi_i : \mathbb{R}^n \to \mathbb{R}$.
- Týmto spôsobom prerobíme trénovaciu množinu:

$$X|y = \left[\begin{array}{cccc|c} x_{11} & x_{12} & \ldots & x_{1n} & y_1 \\ x_{21} & x_{22} & \ldots & x_{2n} & y_2 \\ & \ldots & & & \\ x_{t1} & x_{t2} & \ldots & x_{tn} & y_t \end{array}\right] \quad \Rightarrow_\phi \quad \Phi|y = \left[\begin{array}{cccc|c} \phi_1(x_1) & \phi_2(x_1) & \ldots & \phi_m(x_1) & y_1 \\ \phi_1(x_2) & \phi_2(x_2) & \ldots & \phi_m(x_2) & y_2 \\ & \ldots & & & \\ \phi_1(x_t) & \phi_2(x_t) & \ldots & \phi_m(x_t) & y_t \end{array}\right]$$

- S novou trénovacou množinou $\Phi|y$ narábame pri trénovaní rovnako ako s $X|y$; výsledná hypotéza bude lineárnou kombináciou bázových funkcií pôvodných atribútov.

Napríklad: pomocou bázových funkcií $\phi_1(x) = 1, \phi_2(x) = x, \phi_3(x) = x^2$ sa môžeme naučiť polynómy druhého stupňa

4.1 Lokálne bázové funkcie

4.2 Lokálne váhovaná aproximácia