

Úloha binárnej klasifikácie - logistická regresia

Formulácia úlohy

Daných je m dátových bodov a n charakteristík. Hodnoty charakteristík sú uložené vo vstupnom vektore $(u_1, \dots, u_n)^T \in \mathbb{R}^n$. Každý z m dátových bodov zároveň patrí do jednej z dvoch dátových tried, čo je vyjadrené výstupnou hodnotou (label) $v \in \{0, 1\}$ (preto “binárna klasifikácia”). Množina dátových bodov má štruktúru

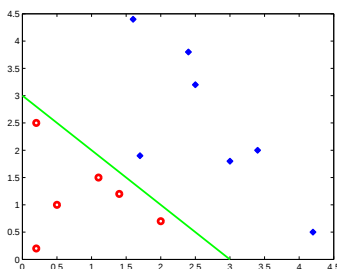
$$M = \left\{ (u_1^{(i)}, \dots, u_n^{(i)}, v^{(i)}) \in \mathbb{R}^n \times \{0, 1\} \mid i = 1, \dots, m \right\}.$$

Hovoríme, že dáta sú lineárne separovateľné, ak existuje nadrovina daná vektorom $(x_1, \dots, x_n)^T \in \mathbb{R}^n$ a konštantou x_0 , ktorá oddeľuje (separuje) body s labelom 0 od bodov s labelom 1. Ak dodefinujeme pre každý dátový bod hodnotu $u_0 = 1$, t.j. vstupné dáta budú $n + 1$ rozmerné vektory $(1, u_1, \dots, u_n)^T$, a označíme

$$u = (1, u_1, \dots, u_n)^T, \quad x = (x_0, x_1, \dots, x_n)^T,$$

matematicky môžeme separovateľnosť bodov vyjadriť ako

$$x^T u^{(i)} < 0 \text{ ak } v^{(i)} = 0, \quad x^T u^{(i)} > 0 \text{ ak } v^{(i)} = 1.$$



Príklad.

V domácej úlohe je $m = 241$, daná je len jedna charakteristika (výška), t. j. $n = 1$. Preto je úlohou nájsť pre danú množinu dát deliacu priamku danú parametrami (x_0, x_1) .

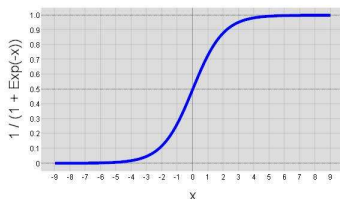
Príklad. V diagnostike nádorov môžu vstupné dáta predstavovať napr. u_1 = polomer, u_2 = plocha, u_3 = textúra, u_4 = symetria, výstupná hodnota je $v = 0$ ak je nádor zhubný, $v = 1$ ak je nádor nezhubný.

Modelovanie problému

Na modelovanie danej situácie je rozumné použiť funkciu

$$h(x; u) = g(x^T u), \quad \text{kde } g(z) = \frac{1}{1 + e^{-z}}.$$

Funkcia g sa nazýva logistická funkcia (alebo aj sigmoid).



Všimnime si, že ak x sú parametre deliacej nadroviny, tak funkcia $h(x; u)$ má nasledujúce vlastnosti:

- $0 \leq h(x; u) \leq 1$;
- $h(x, u^{(i)}) > 0,5$ ak $v^{(i)} = 1$;
- $h(x, u^{(i)}) < 0,5$ ak $v^{(i)} = 0$.

Funkcia $h(x; u)$ sa často interpretuje ako pravdepodobnosť, $h(x; u) = P(v = 1 | u; x)$. Teda pre fixnú hodnotu x , ak $h(x; u) = 0.7$, tak dátový bod u je so 70% pravdepodobnosťou v triede s labelom 1 (a s 30% pravdepodobnosťou v triede s labelom 0).

Formulácia optimalizačnej úlohy

Otázkou stále zostáva, ako nájsť parametre deliacej nadroviny x . Tento problém by sme chceli formulovať ako optimalizačnú úlohu. Keďže chceme aby hodnoty $h(x, u^{(i)})$ boli pre $v^{(i)} = 1$ blízke 1 a pre $v^{(i)} = 0$ blízke 0, ponúka sa možnosť použiť metódu najmenších štvorcov:

$$\text{Min } J(x) = \frac{1}{2m} \sum_{i=1}^m \left(h(x; u^{(i)}) - v^{(i)} \right)^2 = \frac{1}{m} \sum_{i=1}^m C_{quad}(x; u^{(i)}, v^{(i)}), \quad x \in \mathbb{R}^{n+1}.$$

kde

$$C_{quad}(x; u, v) = \frac{1}{2} (h(x; u) - v)^2.$$

Problém však je, že funkcia $J(x)$ nie je konvexná, pretože ani funkcia $C_{quad}(x; u, v)$ nie je konvexná v x . (Overte na malom príklade). Preto, ako vieme, použitie štandardných techník nemusí viesť k nájdeniu globálneho minima.

Ako konvexná alternatíva k funkcii $C_{quad}(x; u, v)$ sa používa funkcia

$$C_{ln}(x; u, v) = \begin{cases} -\ln(h(x; u)) & \text{ak } v = 1, \\ -\ln(1 - h(x; u)) & \text{ak } v = 0, \end{cases}$$

alebo ekvivalentne

$$C_{ln}(x; u, v) = -v \ln(h(x; u)) - (1 - v) \ln(1 - h(x; u)).$$

Takáto funkcia je konvexná v premennej x (overte) a keďže $h(x; u)$ nadobúda hodnoty v intervale $(0, 1)$, funkcia $C_{ln}(x; u, v)$ nadobúda hodnoty v intervale $(0, \infty)$. Zároveň platí

- pre $v = 1$:
 - ak $h(x; u) \rightarrow 1$, tak $C_{ln}(x; u, v) \rightarrow 0$,
 - ak $h(x; u) \rightarrow 0$, tak $C_{ln}(x; u, v) \rightarrow +\infty$,
- pre $v = 0$:
 - ak $h(x; u) \rightarrow 0$, tak $C_{ln}(x; u, v) \rightarrow 0$,
 - ak $h(x; u) \rightarrow 1$, tak $C_{ln}(x; u, v) \rightarrow +\infty$.

Teda pokiaľ je hodnota $h(x; u)$ blízka hodnote v , tak hodnota $C_{ln}(x; u, v)$ sa blíži k minimálnej hodnote, t.j. k nule. Preto je funkcia $C_{ln}(x; u, v)$ vhodná na formuláciu nášho problému, čo vedie k optimalizačnej úlohe

$$\text{Min } J(x) = \sum_{i=1}^m C_{ln}(x; u^{(i)}, v^{(i)}), \quad x \in \mathbb{R}^{n+1}. \quad (1)$$

Dá sa ľahko ukázať, že funkcia $J(x)$ sa dá zjednodušiť do tvaru

$$J(x) = \sum_{i=1}^m \left[(1 - v^{(i)}) x^T u^{(i)} + \ln(1 + e^{-x^T u^{(i)}}) \right],$$

(stačí si uvedomiť, čomu sa rovnajú hodnoty $\ln(g(z))$ a $\ln(1 - g(z))$.)

Vyriešením úlohy nájdeme optimálny vektor $\hat{x} = (\hat{x}_0, \hat{x}_1, \dots, \hat{x}_n)$, ktorý definuje separujúcu nadrovinu, pokiaľ sú dané dátové body separovateľné. Avšak aj pokiaľ nie sú separovateľné, úloha má tiež riešenie, ktoré je dostatočne dobré (definuje nadrovinu, ktorá “takmer” separuje dané triedy bodov).

Pomocou riešenia úlohy môžeme potom zaradiť nové dátové body do jednej z dvoch tried. Teda pre nový vstupný vektor u spočítame hodnotu $h(\hat{x}; u)$. Potom

- ak $h(\hat{x}; u) \geq 0,5$, tak u zaradíme do triedy s labelom 1;
- ak $h(\hat{x}; u) < 0,5$, tak u zaradíme do triedy s labelom 0.