Fakulta matematiky, fyziky a informatiky Univerzity Komenského, Bratislava

Projekt z metód voľnej optimalizácie

Logistická regresia pomocou kvázinewtonovských metód – predikcia solventnosti klientov

Piati za optimalizáciu
Tomáš Antal, 2DAV, 0.2
Erik Božík, 2DAV, 0.2
Róbert Kendereš, 2DAV, 0.2
Teo Pazera, 2DAV, 0.2
Andrej Špitalský, 2DAV, 0.2

Obsah

0 Predstavenie témy **2**
0.1 Zavedenie značenia 2

1 Odvodenie účelovej funkcie a jej gradientu **3**
1.1 Kompaktnejší tvar účelovej funkcie 3
1.2 Gradient účelovej funkcie 4

2 Riešenie optimalizačnej úlohy **5**
2.1 Kvázinewtonovské metódy 5
2.2 Gradientné metódy 5

3 Vizualizácia konvergencie **6**

4 Binárna klasifikácia solventnosti klientov **7**

5 Nadstavba - všeobecný model pre logistickú regresiu pomocou kvázinewtonovských alebo gradientných metód **8**

6 Záver a diskusia **9**

7 Prehľad kódu **10**

0 Predstavenie témy

0.1 Zavedenie značenia

- $m = 699$ značí počet klientov, o ktorých máme dáta
- $v \in \mathbb{R}^m$, i-ta zložka má hodnotu 1, ak je klient i solventný, inak 0
- $u_j \in \mathbb{R}^m$, $j = 1, 2, 3$, vektory údajov o klientoch
  - u_1 – počet mesiacov od otvorenia účtu
  - u_2 – pomer úspor a investícií
  - u_3 – počet rokov v súčasnom zamestnaní
- v^i, u_j^i označujú i-te položky jednotlivých vektorov pre $i = 1, \dots, m$, $j = 1, 2, 3$

1 Odvodenie účelovej funkcie a jej gradientu

V tejto časti sa budeme venovať odvodzovaniu účelovej funkcie a jej gradientu, ktorú v neskorších častiach budeme minimalizovať, pomocou čoho vytvoríme model na binárnu klasifikáciu.

Do logistickej funkcie $g(z) = \frac{1}{1+e^{-z}}$, ktorá bude odhadovať pravdepodobnosť solventnosti klienta, budeme dosádzať hodnoty $z = x^T u^i$ pre vektor parametrov $x = (x_0, \ldots, x_3)$ a vektor údajov o klientovi $u^i = (1, u_1^i, u_2^i, u_3^i)$, pre $i = 1, \ldots, m$.

Chceme odhadnúť zložky vektora x tak, aby čo najvierohodnejšie predpovedal solventnosť vzhľadom na naše dáta. To vedie k optimalizačnej úlohe:

$$\min J(x) \tag{1}$$
$$x \in \mathbb{R}^4 \tag{2}$$

kde

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

Z predpisu funkcie si môžeme všimnúť, že suma nadobúda záporné hodnoty, čiže $J(x)$ nadobúda kladné hodnoty. Taktiež si môžeme všimnúť, že ak je klient i solventný, čiže $v^i = 1$ a pre nejaký vektor parametrov x je hodnota $g\left(x^T u^i\right)$ blízka nule, má to za následok „výrazné“ zvyšovanie hodnoty účelovej funkcie. Podobnou logikou vidíme zvyšovanie hodnoty účelovej funkcie pre nesolventných klientov, ak pomocou vektora x mu prisúdime veľkú pravdepodobnosť solventnosti hodnotou $g\left(x^T u^i\right)$. Chceme teda nájsť taký vektor x, že $g\left(x^T u^i\right)$ bude blízke 1 pre solvetného klienta a blízke 0 pre nesolventného.

1.1 Kompaktnejší tvar účelovej funkcie

Pre lepšiu manipuláciu a neskoršiu implementáciu si zjednodušíme tvar účelovej funkcie nasledovne:

$$\begin{aligned}
J(x) &= -\sum_{i=1}^{m} v^i \ln\left(g\left(x^T u^i\right)\right) + (1 - v^i) \ln\left(1 - g\left(x^T u^i\right)\right) \\
&= -\sum_{i=1}^{m} v^i \ln\left(\left(1 + e^{-x^T u^i}\right)^{-1}\right) + (1 - v^i) \ln\left(\frac{e^{-x^T u^i}}{1 + e^{-x^T u^i}}\right) \\
&= -\sum_{i=1}^{m} -v^i \ln\left(1 + e^{-x^T u^i}\right) + (1 - v^i)\left(\ln\left(e^{-x^T u^i}\right) - \ln\left(1 + e^{-x^T u^i}\right)\right) \\
&= -\sum_{i=1}^{m} -v^i \ln\left(1 + e^{-x^T u^i}\right) - (1 - v^i) x^T u^i - (1 - v^i) \ln\left(1 + e^{-x^T u^i}\right) \\
&= \sum_{i=1}^{m} (1 - v^i) x^T u^i + \ln\left(1 + e^{-x^T u^i}\right)
\end{aligned}$$

S takýmto vyjadrením funkcie $J(x)$ budeme pracovať v nasledujúcich častiach.

1.2 Gradient účelovej funkcie

Vyjadríme si najprv parciálnu deriváciu podľa x_0, potom podľa $x_j, j = 1, 2, 3$, keďže tie sa správajú symetricky.

$$
\begin{aligned}
\frac{\partial}{\partial x_0} J(x) &= \frac{\partial}{\partial x_0} \sum_{i=1}^{m} (1 - v^i) x^T u^i + \ln\left(1 + e^{-x^T u^i}\right) \\
&= \sum_{i=1}^{m} \frac{\partial}{\partial x_0} \left((1 - v^i)(x_0 + x_1 u_1^i + x_2 u_2^i + x_3 u_3^i) + \ln\left(1 + e^{-x^T u^i}\right)\right) \\
&= \sum_{i=1}^{m} (1 - v^i) - \frac{e^{-x^T u^i}}{1 + e^{-x^T u^i}} \\
&= \sum_{i=1}^{m} 1 - v^i - \frac{1}{1 + e^{x^T u^i}}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial x_j} J(x) &= \frac{\partial}{\partial x_j} \sum_{i=1}^{m} (1 - v^i) x^T u^i + \ln\left(1 + e^{-x^T u^i}\right) \\
&= \sum_{i=1}^{m} \frac{\partial}{\partial x_j} \left((1 - v^i)(x_0 + x_1 u_1^i + x_2 u_2^i + x_3 u_3^i) + \ln\left(1 + e^{-x^T u^i}\right)\right) \\
&= \sum_{i=1}^{m} (1 - v^i) u_j^i - u_j^i \frac{e^{-x^T u^i}}{1 + e^{-x^T u^i}} \\
&= \sum_{i=1}^{m} \left(1 - v^i - \frac{1}{1 + e^{x^T u^i}}\right) u_j^i \qquad j = 1, 2, 3
\end{aligned}
$$

Toto vieme kompaktne zapísať nasledovne:

$$
\nabla J(x) = \sum_{i=1}^{m} \begin{pmatrix} 1 \\ u_1^i \\ u_2^i \\ u_3^i \end{pmatrix} \left(1 - v^i - \frac{1}{1 + e^{x^T u^i}}\right)
$$

2 Riešenie optimalizačnej úlohy

V tejto časti sa venujeme riešeniu optimalizačnej úlohy 1 rôznymi metódami. Tie boli implementované v Pythone. Konkrétne sme implementovali gradientné metódy (s optimálnou a konštnantou dĺžkou kroku) a kvázinewtonovské metódy BFGS a DFP (s približne optimálnou dĺžkou kroku nájdenou backtracking-om alebo s optimálnou dĺžkou kroku, nájdenou bisekciou).

Ako štartovací bod sme pri každej metóde volili $x_0 = (0, 0, 0, 0)^T$ a ako kritérium optimality bolo použité $||\nabla J(x^k)|| \leq 10^{-3}$. Optimálnym bodom bude teda vektor parametrov x, ktorý budeme používať v logistickej funkcii na odhadovanie solventnosti klienta podľa jeho dát.

Čo sem spísať

1. Analýza bodov, ktoré hodnoty majú najväčší vplyv
2. Odhad solventnosti pre 0,0,0
3. Porovnanie časov výpočtov, či je backtracking rýchlejší

ZLÉ HODNOTY

2.1 Kvázinewtonovské metódy

	BFGS + backtracking	BFGS + bisekcia	DFP + backtracking	DFP + bisekcia
x_0	0.128	0.128	0.208	0.208
x_1	-0.044	-0.044	-0.047	-0.047
x_2	0.304	0.304	0.315	0.315
x_3	0.309	0.309	0.307	0.307

2.2 Gradientné metódy

	optimálny krok	konštantný krok
x_0	0.164	-323.9
x_1	-0.047	-198.3
x_2	0.318	894.4
x_3	0.315	608.2

3 Vizualizácia konvergencie

Čo sem spísať

1. opísať graf, log-škála
2. 2*2 grid pre KNM, 2*1 pre gradientné
3. popísať teoretický typ konvergencie
4. porovnať počet iterácii

4 Binárna klasifikácia solventnosti klientov

Čo sem spísať

1. vypísať výsledky úspešnosti jednotlivých metód
2. porovnať najlepšiu klasifikáciu s najhoršou

5 Nadstavba - všeobecný model pre logistickú regresiu pomocou kvázinewtonovských alebo gradientných metód

Čo sem spísať

1. popísať štruktúru modulu
2. mierne popísať funkčnosť a spúšťanie
3. spomenúť testy

6 Záver a diskusia

7 Prehľad kódu