Projekt z Metód voľnej optimalizácie: Logistická regresia pomocou kvázinewtonovských metód (Predikcia solventnosti klientov)

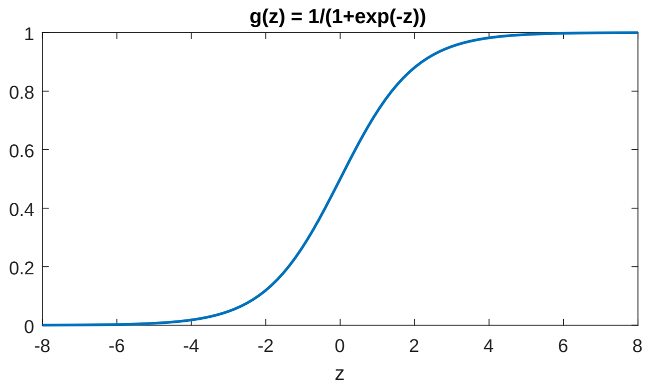
V tomto projekte sa bude odhadovať pravdepodobnosť toho, či klient bude schopný splácať úver, ktorý mu banka poskytne.

K dispozícii máte súbor `credit_risk_train.csv` s údajmi o $m = 699$ klientoch, ktorí žiadali o úver nemenovanú nemeckú banku. Každý riadok prislúcha práve jednému klientovi. Prvý stĺpec obsahuje binárnu premennú v, ktorá nadobúda hodnotu 1, ak klient, ktorý dostal úver, ho bol schopný aj splácať, inak nadobúda hodnotu 0. Ostatné stĺpce obsahujú premenné u_1, u_2, u_3, ktoré udávajú počet mesiacov od otvorenia účtu, pomer úspor a investícii a počet rokov v súčasnom zamestnaní.

Na modelovanie pravdepodobnosti použijeme logistickú regresiu. Logistická funkcia má tvar

$$g(z) = \frac{1}{1+e^{-z}}$$

a vyzerá takto



Zaveďme vektor parametrov logistickej regresie $x = (x_0, x_1, ..., x_5)^T \in \mathbb{R}^6$ a označme $u = (1, u_1, ..., u_5)^T \in \mathbb{R}^6$. Potom sa bude hodnota logistickej funkcie určovať v bodoch

$$x^T u = x_0 + x_1 u_1 + ... + x_n u_n.$$

Hodnota

$$g(x^T u) = \frac{1}{1+e^{-x^T u}} \tag{1}$$

sa potom interpretuje ako pravdepodobnosť toho, že klient s ukazovateľmi u_1, u_2 a u_3 je solventný, teda $g(x^T u) = P(v = 1 \mid u_1, u_2, u_3)$.

Vašou úlohou bude odhadnúť parametre logistickej regresie $x \in \mathbb{R}^4$ na základe údajov v a u. To vedie k optimalizačnej úlohe

$$\text{Min}\left\{ J(x) = -\sum_{i=1}^{m} v^i \ln\left(g\left(x^T u^i\right)\right) + (1-v^i)\ln\left(1 - g\left(x^T u^i\right)\right) \,\middle|\, x \in \mathbb{R}^6 \right\}, \tag{2}$$

kde $u^i = (1, u_1^i, ..., , u_3^i)^T$. Všimnite si, že predpis účelovej funkcie $J(x)$ je zvolený tak, aby sa penalizovala nízka pravdepodobnosť $g\left(x^T u^i\right)$ toho, že klient je solventný, ak naozaj je $(v^i = 1)$ a vysoká pravdepodobnosť $g\left(x^T u^i\right)$ toho, že klient je solventný, ak v skutočnosti nie je $(v^i = 0)$.

a) Ukážte, že predpis funkcie $J(x)$ v úlohe (2) sa dá dosadením funkcie (1) zjednodušiť na tvar

$$\text{Min}\left\{ J(x) = \sum_{i=1}^{m} (1 - v^i)\, x^T u^i + \ln\left(1 + e^{-x^T u^i}\right) \;\middle|\; x \in \mathbb{R}^4 \right\}. \tag{3}$$

b) Vyjadrite prvky gradientu $\nabla J(x)$ účelovej funkcie $J(x)$ v tvare (3).

c) Riešte úlohu (3) pomocou BFGS metódy s približne optimálnou dĺžkou kroku a DFP metódy s približne optimálnou dĺžkou kroku. Na hľadanie približne optimálnej dĺžky kroku využite metódu backtracking (bez potreby odvodenia Hessovej matice). Využite gradient z časti b), štartovací krok zvoľte $x_0 = (0, 0, 0)^T$ a ako kritérium optimality použite $\left\|\nabla J(x^k)\right\| \leq \varepsilon = 10^{-3}$.

d) Riešte úlohu (3) pomocou BFGS metódy s optimálnou dĺžkou kroku a DFP metódy s optimálnou dĺžkou kroku. Na hľadanie optimálnej dĺžky kroku využite metódu bisekcie. Voľte vstupné parametre z časti c), merajte trvanie výpočtu a výsledky porovnajte.

e) Symbolom J^* označme nájdené ε-presné riešenie. Pre rôzne kvázinewtonovské metódy vykreslite do jedného obrázku grafy znázorňujúce vývoj hodnoty $J\left(x^k\right) - J^*$ s rastúcim číslom iterácie k. Na osi y použite logaritmickú mierku.

f) Symbolom J^* označme nájdené ε-presné riešenie. Pre rôzne verzie gradientnej metódy vykreslite do jedného obrázku grafy znázorňujúce vývoj hodnoty $J\left(x^k\right) - J^*$ s rastúcim číslom iterácie k. Na osi y použite logaritmickú mierku.

g) Riešením úlohy (3) sme odhadli model na binárnu klasifikáciu, t.j. ak pravdepodobnosť toho, že klient je solventný, je aspoň 50 %, tak ho klasifikujeme ako solventného, inak ako nesolventného. Na dátach zo súboru `credit_risk_test.csv`, ktoré majú rovnakú štruktúru ako pôvodné dáta `credit_risk_train.csv`, otestujte, ako tento model klasifikuje solventnosť klienta na základe pozorovaných ukazovateľov u_1,u_2,u_3, t.j. zistite, akú časť klientov klasifikuje model správne. Porovnajte kvalitu klasifikácie s vektorom x získaným metódou, ktorú považujete za najlepšiu.

h) *Nadstavba:* Vymyslite vhodnú modifikáciu projektu. Pokúste sa napríklad modelovať binárnu premennú v len pomocou jedného ukazovateľa. Výsledky takejto klasifikácie môžete vykresliť v dvojrozmernom priestore. Môžete porovnať kvázinewtonovské metódy s inými metódami z prednášky, prípadne experimentujte s voľbou štartovacieho kroku (tu však dávajte pozor na maximálny počet povolených iterácií a trvanie metódy).