

## Pravidlá pre domáce úlohy

Počas semestra budú zadávané domáce úlohy, ktoré sa budú týkať riešenia optimalizačnej úlohy, opísanej v druhej časti tohto dokumentu, pomocou aktuálne preberaných metód.

Dokopy budú mať domáce úlohy hodnotu 40 bodov (40 % hodnotenia) a budete ich robiť v skupinách (1-4 študenti, rovnakí pre všetky úlohy).

Riešenie domácej úlohy bude pozostávať z na $\text{\LaTeX}$ ovaného **pdf súboru** obsahujúceho potrebné výstupy (číselné výstupy, obrázky, popisy, zhodnotenie, ...) a z **matlabovských súborov** s použitými kódmi.

Na domácej úlohe môžete pracovať aj na cvičení (po dokončení obsahu cvičenia) a konzultovať riešenie s cvičiacim.

Termín na odovzdanie domácej úlohy bude do nasledujúceho cvičenia.

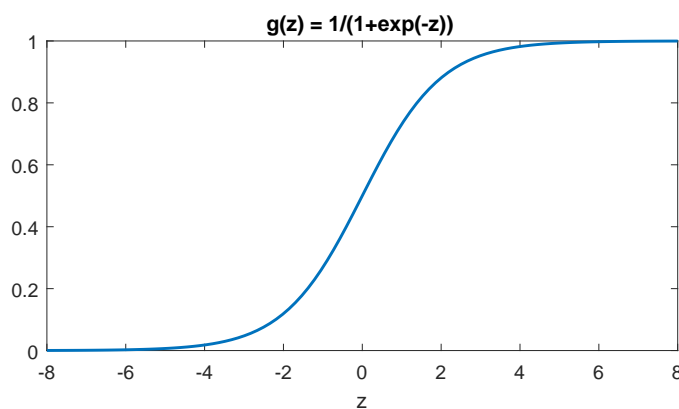
## Logistická regresia

Na základe údajov o výške študenta (alebo študentky) chceme odhadnúť pravdepodobnosť toho, že je to muž. K dispozícii máme súbor *data.csv* s údajmi o  $m = 241$  študentoch, pričom každý riadok zodpovedá jednému študentovi. Prvý stĺpec obsahuje binárnu premennú  $v$  určujúcu, či je to študent ( $v = 1$ ) alebo študentka ( $v = 0$ ). Druhý stĺpec obsahuje premennú  $u_1$ , ktorá udáva výšku uvedenú v metroch.

Na modelovanie pravdepodobnosti použijeme logistickú regresiu. Logistická funkcia má tvar

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

a vyzerá takto



Definujeme  $u = (1, u_1)^T$  a logistickú funkciu budeme vyčísl'ovať v bodoch

$$x^T u = x_0 + x_1 u_1$$

pre nejaké parametre  $x = (x_0, x_1)^T \in \mathbb{R}^2$ . Hodnotu

$$g(x^T u) = g(x_0 + x_1 u_1) = \frac{1}{1 + e^{-x_0 - x_1 u_1}}$$

interpretujeme ako pravdepodobnosť toho, že študent s výškou  $u_1$  je muž, teda  $g(x^T u) = P(v = 1 | u_1)$ .

Vašou úlohou bude na základe údajov zo súboru *data.csv* odhadnúť parametre  $x \in \mathbb{R}^2$  logistickej regresie. To vedie k optimalizačnej úlohe

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

kde  $u^i = (1, u_1^i)^T$ . Tvar účelovej funkcie  $J(x)$  je zvolený tak, aby bola penalizovaná nízka pravdepodobnosť  $g(x^T u^i)$  toho, že študent je muž, ak je to naozaj muž ( $v^i = 1$ ) a, naopak, vysoká pravdepodobnosť  $g(x^T u^i)$  toho, že študent je muž, ak je to naozaj žena ( $v^i = 0$ ).

Funkcia  $J(x)$  sa dá po dosadení zjednodušiť a dostávame úlohu v tvare

$$\text{Min} \left\{ J(x) = \sum_{i=1}^m (1 - v^i) x^T u^i + \ln \left( 1 + e^{-x^T u^i} \right) \mid x \in \mathbb{R}^2 \right\}.$$

Túto úlohu budete počas semestra riešiť rôznymi metódami, ktoré na nej porovnáte.