Domáca úloha 3

Support vector machine (SVM) je klasifikačná metóda, ktorá (v našej najjednoduchšej verzii) pomocou vysvetľujúcich premenných $\mathbf{x}_i \in \mathbb{R}^k$ modeluje vysvetľovanú $y_i \in \{-1, 1\}$ $(i = 1, \ldots, n)$. Teda pomocou x-ov sa snaží správne zaradiť ypsilony do kategórie -1 a do kategórie 1.

V datasete `kapusta.txt` máme údaje o n hlávkach kapusty, pričom u každej vieme, akého je typu (stĺpec `Cult`: to je náš vektor $\mathbf{y}$ vyjadrujúci typy 1 a -1), akú má váhu a aký má podiel istej kyseliny (stĺpce `HeadWt` a `VitC`). Vektory $\mathbf{x}_i \in \mathbb{R}^k$ sú v našom prípade riadky matice $\mathbf{X} \in \mathbb{R}^{n\times 3}$:

$$\mathbf{X} = \big[\mathbf{1}_n, \quad \text{stĺpce 2 a 3 z kapusta.txt}\big] = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix},$$

kde $\mathbf{1}_n$ je stĺpec samých jednotiek. Teda máme $k = 3$ a $\mathbf{x}_i = (1, HeadWt_i, VitC_i)^T$.

SVM rieši úlohu

$$\min_{\mathbf{w}\in\mathbb{R}^k} \Big(\sum_{j=2}^{k} w_j^2 + \sum_{i=1}^{n} \max(0, 1 - y_i \cdot (\mathbf{x}_i^T \mathbf{w})) \Big) \tag{1}$$

(všimnite si, že prvá suma nezahŕňa prvok $j = 1$, ktorý reprezentuje intercept, a že sumy sú nasčítavané cez rôzne množiny indexov). [Výraz v druhej sume značí, že pre každé i vypočítame $1 - y_i \mathbf{x}_i^T \mathbf{w}$ a zoberieme maximum z nuly a tohto čísla. Tieto všetky medzivýsledky sa nakoniec sčítajú.] Parametre $\mathbf{w} \in \mathbb{R}^k$ sú teda analogické parametrom β v lineárnej regresii (len v lineárnej regresii sa minimalizuje iná funkcia parametrov β). Aplikujte túto metódu na dáta v súbore `kapusta.txt`, čiže:

1. Načítajte dáta z `kapusta.txt` na stránke (hint: zvoľte správne `sep` a `header`).

2. Pomocou optimalizačných funkcií v Rku pre tieto dáta nájdite $\mathbf{w}^* \in \mathbb{R}^k$, ktoré minimalizuje (1). Hodnoty $\mathbf{x}_i$ a y_i sú popísané vyššie. Na nájdenie $\mathbf{w}^* \in \mathbb{R}^k$ nevyužívajte priamo funkcie riešiace support vector machines.

3. Na overenie, či ste dostali niečo zmysluplné, porovnajte predikcie modelu so skutočnými y_i. Predikcie modelu sú: $\widehat{y}_i = 1$ ak $\mathbf{x}_i^T \mathbf{w}^* \geq 0$ a $\widehat{y}_i = -1$ ak $\mathbf{x}_i^T \mathbf{w}^* < 0$. Čiže $\widehat{y}_i = \operatorname{sgn}(\mathbf{x}_i^T \mathbf{w}^*)$ $(i = 1, \ldots, n)$ (ak ignorujeme možnosť $\mathbf{x}_i^T \mathbf{w}^* = 0$). Môžete napríklad spočítať senzitivitu

$$\frac{\text{počet } i, \text{ že } y_i = 1 \text{ aj } \widehat{y}_i = 1}{\text{počet } i, \text{ že } y_i = 1}$$

a špecificitu

$$\frac{\text{počet } i, \text{ že } y_i = -1 \text{ aj } \widehat{y}_i = -1}{\text{počet } i, \text{ že } y_i = -1}.$$

Zrejme čím bližšie sú tieto čísla k 1 (k 100%), tým lepšie.

Poznámka 1: Výrazy v (1) sa dajú vypočítať elegantne a efektívne (bez použitia cyklov a funkcii ako replicate). Odporúčam tento spôsob, ale nie je to povinné (čiže v kľude

napr. použite cykly, ak vám je to tak príjemnejšie).
Poznámka 2: Funkcia (1) nie je diferencovateľná (môžete si rozmyslieť, prečo), takže klasické optimalizačné metódy pre ňu nie sú stopercentne korektné. Namiesto toho sa minimalizácia (1) typicky rieši pomocou ohraničenej optimalizácie, ktorá už má diferencovateľnú účelovú funkciu aj ohraničenia. Tento problém ale pre potreby domácej úlohy ignorujte.
Poznámka 3: Ak je niečo nejasné, ozvite sa.