Zadatak

Učitajte primjere iz datoteka `X_train.txt` (9982 stupaca i 80 redaka) i `y_train.txt` (80 redaka). U i-tom retku datoteke `X_train.txt` nalazi se i-ti primjer u vektorskom zapisu, a u i-tom retku datoteke `y_train.txt` nalazi se klasifikacija tog primjera (1 = kategorija *Svijet*, 0 = ostali tekstovi).

(a) Napišite (ne u kôdu nego u izvještaju) funkciju pogreške koju optimirate u ovisnosti o primjerima za učenje i parametru λ.

$$E(\tilde{\mathbf{w}}|\mathcal{D}) = -\sum_{i=1}^{N}\left\{y^{(i)}\ln h(\mathbf{x}^{(i)}) + (1-y^{(i)})\ln(1-h(\mathbf{x}^{(i)}))\right\} + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

(b) Implementirajte metodu gradijentnog spusta kojom ćete pronaći mimumum te funkcije. Uz hiperparametar $\lambda = 0$ pronađite parametre logističke regresije koji minimiziraju pogrešku na skupu za učenje (koristeći matrice `X_train` i `y_train`). *Uputa:* pogrešku možete smanjiti ako povećate broj iteracija i smanjite parametar η.

Napišite u izvještaju vrijednost funkcije pogreške na skupu za učenje (uz dobiveni vektor težina). U datoteku `w_a.txt` zapišite vektor težina tako da u retku i piše vrijednost komponente s indeksom $i-1$ (tj. u prvom retku piše vrijednost težine w_0, u drugom težine w_1, itd.). Uz ovako naučen klasifikator broj pogrešno klasificiranih primjera na skupu za učenje trebao bi biti 0.

$E(\tilde{\mathbf{w}}|\mathcal{D}) =$ 0.0017 ← Dobiveno uz eta = 0.007 i 1500 iteracija.

(c) U ovom podzadatku potrebno je pronaći najbolju vrijednost hiperparametra λ korištenjem skupa za učenje i skupa za provjeru.

Odaberite 20-ak vrijednosti parametra λ (obavezno uključite $\lambda = 0$). Raspon parametra λ je od 0 do ∞, pa se preporučuje da odabrani parametri pokrivaju nekoliko redova veličina (kako vrijednosti manje od 1, tako i vrijednosti znatno veće od 1). Za svaki od odabranih vrijednosti parametra λ pronađite najbolji vektor težina korištenjem skupa za učenje. Pomoću svakog od naučenih vektora težina klasificirajte primjere iz skupa za provjeru (datoteke `X_validate.txt` i `y_validate.txt`) i zapišite postotak pogrešno klasificiranih primjera. Broj pogrešno klasificiranih primjera trebao bi biti manji od 20%.

U izvještaju napravite tablicu koja će sadržavati sve vrijednosti λ, postotak pogrešno klasificiranih primjera na skupu za provjeru za svaku vrijednost od λ te vrijednosti koje su korištene u metodi gradijentnog spusta (η i broj iteracija).

λ	postotak krivo klasificiranih	η	broj iteracija
0	18.333 %	0.00005	2500
0.00003	18.333 %	0.00005	2500
0.0001	18.333 %	0.00005	2500
0.0003	18.333 %	0.00005	2500
0.001	18.333 %	0.00005	2500
0.003	18.333 %	0.00005	2500
0.01	18.333 %	0.00005	2500
0.03	18.333 %	0.00005	2500
0.1	18.333 %	0.00005	2500
0.3	18.333 %	0.00005	2500
1	18.333 %	0.00005	2500
3	18.333 %	0.00005	2500
10	15.000 %	0.00005	2500
30	15.000 %	0.00005	2500
100	13.333 %	0.00005	2500
300	13.333 %	0.00005	2500
1000	25.000 %	0.00005	2500
3000	38.333 %	0.00005	2500
10000	51.667 %	0.00005	2500
30000	53.333 %	0.00005	2500

(d) Koristeći parametar λ iz prošlog podzadatka, kojim se minimizira broj pogrešno klasificiranih primjera na skupu za provjeru, potrebno je naučiti klasifikator korištenjem spojenih skupova za učenje i provjeru. Težine tako naučenog klasifikatora zapišite u datoteku `w_d.txt`. Klasificirajte primjere iz skupa za ispitivanje (datoteke `X_test.txt` i `y_test.txt`) i napišite postotak pogrešno klasificiranih primjera. Postotak pogrešno klasificiranih primjera ovako naučenog klasifikatora trebao bi biti manji od klasifikatora naučenog uz $\lambda = 0$.

Uz lambda = 100, eta = 0.00005, iter = 2500, nakon učenja težina na temelju primjera iz spojenih skupova za učenje i provjeru, postotak pogreške na skupu za ispitivanje mi je: 6.667%. Ako ne bih koristio regularizaciju, imao bih postotak pogreške 25%.