Mašinsko učenje – Domaći 2
Logistička regresija i GNB

Nemanja Saveski 2023/3163

12. novembar 2023.

1 Logisitčka regresija

Potrebno je prvo projektovati klasifikator logističkom regresijom. U datasetu, u prvih 5 kolona se nalaze prediktori, a šesta kolona predstavlja klasu. Pre bilo kakvog konstruisanja klasifikatora, standarizovaćemo prediktore i podatke podeliti na trening i test skup. Pošto konstruišemo 3 klasifikatora jedan protiv svih (eng. *one-against-all*), uradićemo *one-hot encoding* za svaku klasu i tako predstaviti kojoj klasi dati odbirak pripada. Za svaki klasifikator na početku su inicijalizovani θ vektori kao nula vektori. Logistička funkcija je oblika:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}},$$

i daje na izlazu 0 ili 1 u zavisnosti od parametara θ i podataka x. Negativna log-verodostojnost se računa:

$$l(\theta) = -\frac{1}{m_{mb}}(y^T \log(h_\theta) + (1 - y)^T \log(1 - h_\theta))$$

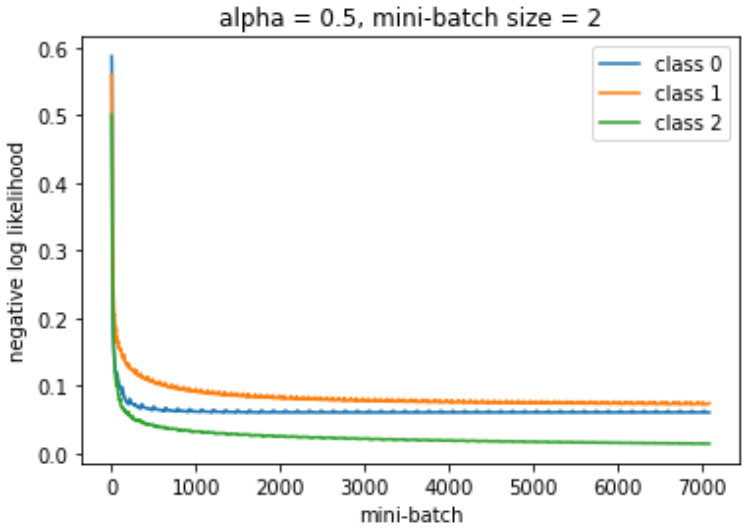
Dalje, na osnovu $h_\theta(x)$, možemo izračunati gradijent u svakoj iteraciji (osim promenljive m_{mb}, sve su vektori, m_{mb} je veličina mini-šarže):

$$\nabla l(\theta) = \frac{1}{m_{mb}} x^T (h_\theta(x) - y),$$

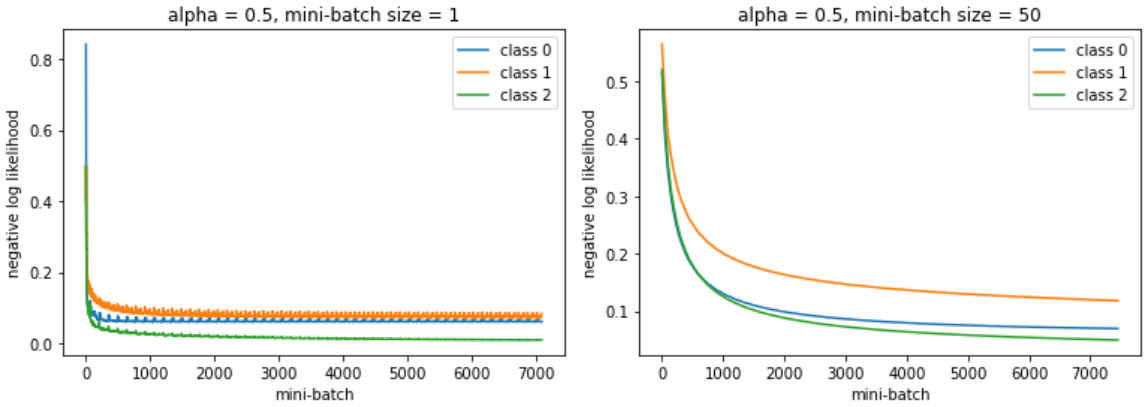
i na osnovu toga dobiti promenu θ u svakoj iteraciji:

$$\theta \leftarrow \theta + \alpha \nabla l(\theta)$$

Dobijaju se optimalne vrednosti za koeficijent obučavanja i veličinu mini šarže $\alpha^* = 0.5$, $m^*_{mb} = 2$, a obučavanje u tom slučaju izgleda ovako:

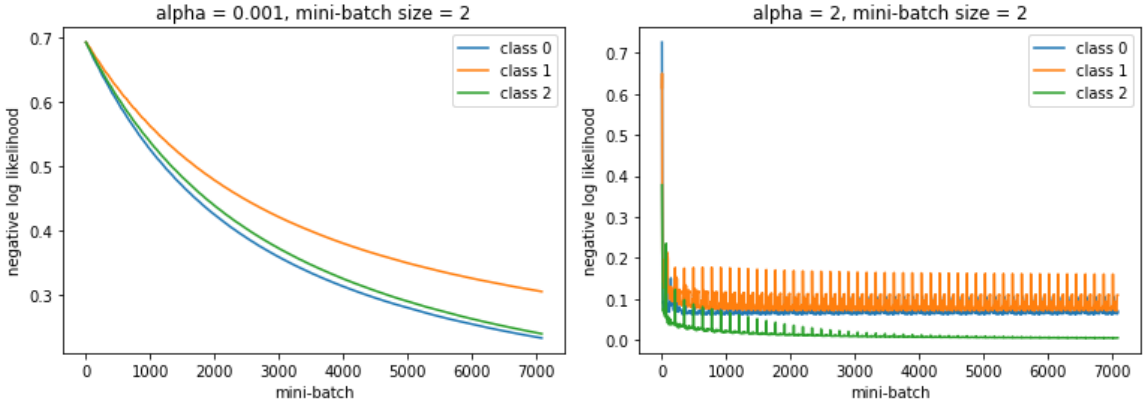


Za optimalno $\alpha^* = 0.5$, u ekstremnim slučajevima $m_{mb} = 1$ i $m_{mb} = 50$, dobijamo sledeća obučavanja:



Deluje da je u redu uzeti i $m_{mb} = 1$, ali dobija se malo veća tačnost sa $m_{mb} = 2$.

Za optimalno $m_{mb} = 2$, u ekstremnim slučajevima $\alpha = 0.001$ i $\alpha = 2$, dobijamo sledeća obučavanja:



2 GNB

U ovom delu domaćeg, ćemo prvo standardizovati podatke, kao u prošloj tački, i podeliti celokupni dataset na train i test skup. Projektovanje GNB klasifikatora podrazumeva da prvo naučimo raspodele svake klase iz podataka. Dakle pronaći ćemo (μ_0, σ_0), (μ_1, σ_1), (μ_2, σ_2) iz trening skupa. Sve ove raspodele su Gausovske pa odatle i ime GNB (eng. *Gaussian Naive Bayes*). Zatim, za svaki novi odbirak, odredićemo verodostojnost da pripada svakoj klasi na sledeći način:

$$p(x_1, x_2, x_3, x_4, x_5|y_i) = p(x_1|y_i)p(x_2|y_i)p(x_3|y_i)p(x_4|y_i)p(x_5|y_i), \quad i = 0, 1, 2,$$

a vrednosti funkcija p se dobijaju na sledeći način iz normalne raspodele:

$$p(x_i|y_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}}$$

Predikcije dobijamo iz sledeće funkcije, u stvari maksimizujemo verodostojnost:

$$\hat{y} = \underset{i}{\operatorname{argmax}}\ p(x_1, x_2, x_3, x_4, x_5|y_i)$$

Dobijamo konačnu tačnost u ovom delu zadatka : **96.48%**, dok je najbolja tačnost (za deo zadatka logistička regresija, sa optimalnim α i m_{mb}) **97.89%**.