Katedra za Signale i sisteme
Elektrotehnički fakultet
Univerzitet u Beogradu

Mašinsko učenje

Drugi domaći zadatak

Logistička regresija i GNB

Student:
Aleksandar Đorđević 2023/3049

Mentor:
dr. Predrag Tadić

1 Logistička regresija

1.1 Postavka problema

Potrebno je projektovati klasifikator na bazi logističke regresije za problem klasifikacije primera u 3 različite klase. Kako logistička regresija klasifikuje primer u jednu od dve moguče klase potrebno je primeniti pristup jedan protiv svih. Poređenjem vrednosti hipoteza sva 3 klasifikatora donosi se odluka kojoj klasi primer pripada.

1.2 Rešenje

Skup podataka je najpre podeljen na trening i test podskupove. Zatim su ulazni podaci oba podskupa standardizovani tako da imaju nultu srednju vrednost i jediničnu varijansu. Oba podskupa su standardizovana na osnovu primera iz trening skupa. Hipoteza logističke regresije se dobija kada se na linearnu kombinaciju prediktora primeni logistička funkcija.

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

Sada u zavisnosti od problema postavljamo granicu sa kojom upoređujemo vrednost logističke funkcije i na osnovu toga određujemo kojoj klasi primer pripada.

Verodostojnost je data sledećim izrazom:

$$p(\mathbf{y}|\mathbf{x};\theta) = \prod_{i=1}^{m} h_\theta(x^{(i)})^{y^{(i)}}(1 - h_\theta(x^{(i)}))^{(1-y(i))} \tag{2}$$

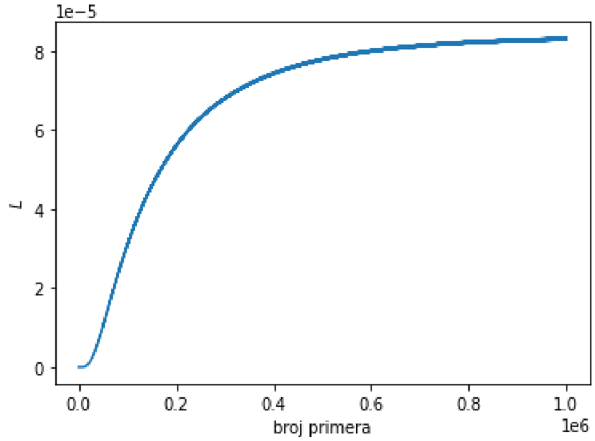
Zbog veoma niskih vrednosti za veliki broj primera postoji rizik od *underflow*-a te se u praksi obično maksimizuje log-verodostojnost.

$$l(\theta) = \sum_{i=1}^{m}(y^{(i)}log(h_\theta(x^{(i)})) + (1 - y^{(i)})log(1 - h_\theta(x^{(i)}))) \tag{3}$$

Kako ne postoji analitičko rešenje parametri se određuju pomoću numeričkih metoda. U ovom zadataku korišćen je mini šaržni (*mini-batch*) gradijentni spust (odnosno uspon kod maksimizacije). Izvod u matričnom obliku je dat sledećom relacijom:

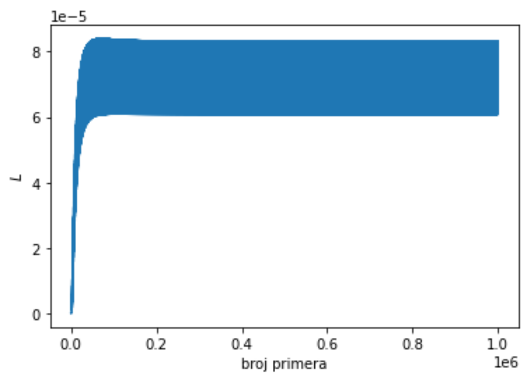
$$\frac{\partial l(\theta)}{\partial \theta} = X^T(Y - h_\theta(X\theta)) \tag{4}$$

I korišćena je formula za gradijentni uspon $\theta^{(t+1)} = \theta^{(t)} + \eta\frac{\partial l(\theta)}{\partial \theta}$.
Možemo videti da je potrebno podesiti 2 hiper-parametra - veličinu *batch*-a m_{bs} i konstantu učenaj η. Oni će biti podešeni manuelno posmatranjem rezultata - tačnosti i verodostojnosti kroz iteracije.
Kako logistička regresija diskriminiše između dve klase, a u našem problemu postoje tri klase primenjen je pristup jedan protiv svih (*one vs all*) odnosno projektovana su 3 klasifikatora na bazi logističke regresije koji govori da li je dati primer iz date klase ili iz neke od ostalih. Poređenjem vrednosti hipoteza primer se klasifikuje u odgovarajuću klasu.
Prilikom treniranja zadat je maksimalan broj obrađenih primera (1 milion u našem slučaju) kako bi se ravnopravno tretirale različite vrednosti veličine *batch*-a.
Za hiper-parametre uzete su vrednosti: $m^*_{bs} = 16$ i $\eta^* = 0.1$. Za ove parametre prikazan je grafik zavisnosti verodostojnosti od broja obrađenih primera za klasifikator za klasu $y = 2$.

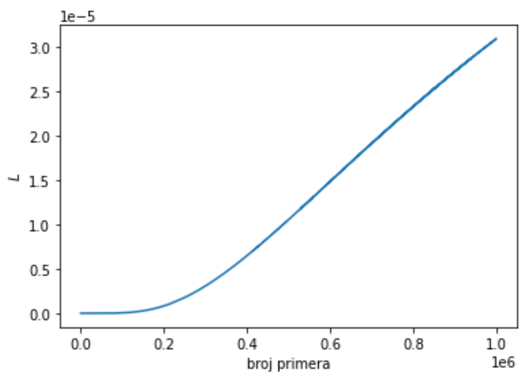


Slika 1: Zavisnost verodostnojnosti od broja obrađenih primera za optimalne parametre

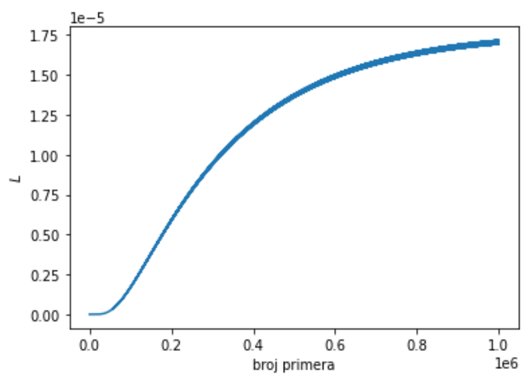
U nastavku su dati grafici za suboptimalne parametre.



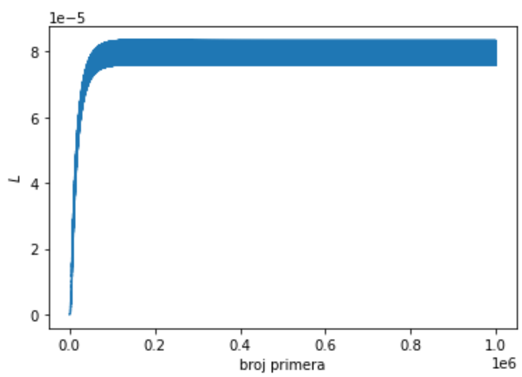
(a) $\eta = 0.1, m_{bs} = 1$



(b) $\eta = 0.01, m_{bs} = 16$



(c) $\eta = 0.1, m_{bs} = 64$



(d) $\eta = 1, m_{bs} = 16$

Slika 2: Zavisnost verodostojnosti od broj obrađenih primera

Možemo primetiti da za mali *batch* verodostojnost veoma brzo konvergira ali sadrži i velike oscilacije, dok je za veliki *batch* konvergencija sporija. Za malu konstantu obučavanja verodostojnost uopšte ne konvergira za dati broj primera dok za veću konvergira veoma brzo uz primetnu oscilatornost.
Na trening skupu postigunuta je tačnost od 97.89%, dok je tačnost na test skupu 94.44%.

2 Gausovski naivni Bajes

2.1 Postavka problema

Projektovati Gausovski naivni Bayesov klasifikator za dati set podataka.

2.2 Rešenje

Podaci su obrađeni na isti način kao u prethodnoj tački.
Naivni Bayes-ov klasifikator polazi od pretpostavke o uslovnoj nezavisnosti prediktora kada je izlazna promenljiva fiksirana odnosno važi:

$$p(x_1, \ldots, x_n|y) = p(x_1|y) \ldots p(x_n|y) \tag{5}$$

Za Gausovski naivni Bayes specijalno je to što pretpostavlja da su raspodele Gausovske te je za procenu gustine verovatnoće svakog prediktora dovoljno je proceniti njihovu srednju vrednost i varijansu.

$$p(x_i|y=j) = \frac{1}{\sqrt{2\pi\sigma_{i,j}^2}} e^{-\frac{(x_i-\mu_{i,j})^2}{2\sigma_{i,j}^2}} \tag{6}$$

Maksimizacijom verodostojnosti se dobija da se srednja vrednost i varijansa procenjuju na sledeći način:

$$\mu_{i,j} = \frac{1}{N_{y=j}} \sum_{k:y^{(k)}=j} x_i^{(k)} \tag{7}$$

$$\sigma_{i,j} = \frac{1}{N_{y=j}} \sum_{k:y^{(k)}=j} (x_i^{(k)} - \mu_{i,j})^2 \tag{8}$$

Takođe, neophodno je proceniti apriorne verovatnoće klasa:

$$p(y=j) = \frac{N_{y=j}}{N} \tag{9}$$

Sada na osnovu Bayes-ove teoreme možemo izračunati aposteriorne verovatnoće:

$$p(y=j|\mathbf{x}) = \frac{p(y=j)\prod_{i=1}^{n} p(x_i|y=j)}{\sum_{k=0}^{L-1} p(y=k)\prod_{i=1}^{n} p(x_i|y=k)} \tag{10}$$

Kako je imenilac uvek isti možemo ga zanemariti i upoređivati samo brojioce. Procena srednje vrednosti i varijanse se vrši isključivo na trening skupu.
Kada dobijemo primer izračunamo aposteriornu verovatnoću za sve 3 klase (to jest samo brojilac) i klasifikujemo ga u onu čija je verovatnoća najveća.
Tačnost klasifikacije na trening skupu iznosi 96.4788%, a na test skupu 91.6666%.