6. Logistička regresija

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, vježbe, v2.2

1 Zadatci za učenje

1. [***Svrha: Znati definirati model logističke regresije. Razumjeti izvod funkcije pogreške unakrsne entropije i pripadne funkcije gubitka. Shvatiti zašto je ta funkcija gubitka unakrsne entropije prikladna za klasifikaciju, dok funkcija kvadratnog gubitka to nije.***]

   (a) Definirajte poopćeni linearni model. Koja je svrha aktivacijske funkcije?

   (b) Definirajte model logističke regresije. Zašto je sigmoidna (logistička) funkcija prikladan odabir za aktivacijsku funkciju?

   (c) Izvedite pogrešku unakrsne entropije $E(\mathbf{w}|\mathcal{D})$ kao negativnu vjerojatnost oznaka na skupu označenih primjera.

   (d) Napišite funkciju gubitka unakrsne entropije i nacrtajte njezin graf. Koliki je najveći a koliki najmanji mogući gubitak?

   (e*) Pretpostavimo da su izlazne oznake $y \in \{-1, +1\}$ umjesto $y = \{0, 1\}$. Reformulirajte funkciju gubitka unakrsne entropije $L(y, h(\mathbf{x}))$ tako da koristi takve oznake te da vrijedi $L(y, 0) = 1$ (kako bi funkcija bila kompatibilna s ostalim funkcijama gubitka koje smo radili).

   (f) Nacrtajte graf funkcije gubitka $L(y, h(\mathbf{x}))$ u ovisnosti o udjelu pogrešne klasifikacije $y\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$, i to za: gubitak 0-1, kvadratni gubitak i logistički gubitak iz (e). Na temelju skice, odgovorite: (i) zašto je logistički gubitak dobar za klasifikaciju, a kvadratni gubitak to nije?; (ii) nanose li ispravno klasificirani primjeri ikakav gubitak?; (iii) možemo li reći da je logistički gubitak *konveksni surogat* gubitka 0-1, i što to znači?

2. [***Svrha: Prisjetiti se definicije konveksnosti funkcije. Razumjeti da konveksnost i unimodalnost nisu jedno te isto.***]

   (a*) Formalno definirajte kada je funkcija $f : \mathbb{R}^n \to \mathbb{R}$ konveksna.

   (b*) Funkcija f je *kvazikonveksna* (ili *unimodalna*) akko je njezina domena $\mathbf{dom}\, f$ konveksna te ako za svaki $\mathbf{x}, \mathbf{y} \in \mathbf{dom}\, f$ i $0 \leqslant \alpha \leqslant 1$ vrijedi

   $$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leqslant \max\{f(\mathbf{x}), f(\mathbf{y})\}.$$

   Kvazikonveksnost je poopćenje konveksnosti: svaka je konveksna funkcija unimodalna, ali obrat ne vrijedi. Pokažite primjerom da obrat ne vrijedi.

   (c) Zašto u strojnom učenju preferiramo konkveksne funkcije pogreške? Koja je veza između konveksnosti funkcije pogreške i konveksnosti funkcije gubitka?

3. [***Svrha: Razumjeti gradijentni spust i potrebu za linijskim pretraživanjem. Znati izvesti gradijentni spust za logističku regresiju. Demonstrirati upoznatost s prednostima i nedostatcima optimizacije drugog reda.***]

   (a) Objasnite ideju gradijentnog spusta i potrebu za linijskim pretraživanjem.

   (b) Objasnite razliku između grupnog (*batch*) i stohastičkog gradijentnog spusta. Koja je prednost ovog drugog?

   (c) Izrazite gradijent funkcije pogreške unakrsne entropije $\nabla E(\mathbf{w}|\mathcal{D})$ i napišite pseudokôd algoritma gradijentnog spusta (grupna i stohastička izvedba).

4. [***Svrha: Razumjeti kako regularizacija i linearna (ne)odvojivost utječu na gradijenti spust i na izgled funkcije pogreške u prostoru parametara.***] Koristimo model L2-regularizirane logističke regresije učene algoritmom gradijentnog spusta. Iskušavamo dvije vrijednosti regularizacijskog faktora $\lambda = 0$ i $\lambda = 100$. Razmatramo posebno linearno odvojiv i linearno neodvojiv problem.

   (a) Skicirajte pogreške učenja i ispitivanja $E(\mathbf{w}|\mathcal{D})$ u ovisnosti o broju iteracija za $\lambda = 0$ i $\lambda = 100$ te za slučaj (i) linearno odvojivih i (ii) linearno neodvojivih primjera (četiri grafikona sa po dvije krivulje).

   (b) Načinite skice izokontura funkcije neregularizirane pogreške $E(\mathbf{w}|\mathcal{D})$ i L2-regularizacijskog izraza u ravnini w_1-w_2. Napravite dvije odvojene skice: za linearno odvojive i linearno neodvojive primjere.

   (c) Na grafikone iz prethodnoga zadatka docrtajte izokonture L2-regulariziranih funkcija pogreške za $\lambda = 100$ i naznačite gdje se nalazi točka minimuma (w_1^*, w_2^*). Gdje bi se nalazila točka minimuma za $\lambda = 0$?

2 Zadatci s ispita

1. (T) Poopćeni linearni model definirali smo kao $h(\mathbf{x}) = f(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}))$, gdje je f neka (moguće nelinearna) aktivacijska funkcija, a $\boldsymbol{\phi}$ je (moguće nelinearna) funkcija preslikavanja u prostor značajki. **Koji od navedenih uvjeta je dovoljan uvjet da granica između klasa u ulaznom prostoru bude linearna?**

   A f je afina funkcija i $\boldsymbol{\phi}(\mathbf{x}) = (1, \mathbf{x})$ B f je afina funkcija C $f(\mathbf{x}) = \mathbf{x}$ D $\boldsymbol{\phi}(\mathbf{x}) = (1, \mathbf{x})$

2. (T) Kod logističke regresije, pogrešku unakrsne entropije izveli smo modelirajući distribuciju vjerojatnosti oznaka $\mathbf{y}$ u skupu označenih primjera. **Na koji smo način modelirali distribuciju vjerojatnosti pojedinačnog primjera y?**

   A $P(y|\mathbf{x}) = h(\mathbf{x})^y(1 - h(\mathbf{x}))^{1-y}$
   B $P(y|\mathbf{x}) = h(\mathbf{x})(1 - h(\mathbf{x}))$
   C $P(y|\mathbf{x}) = (y - h(\mathbf{x}))\mathbf{x}$
   D $P(y|\mathbf{x}) = y^{h(\mathbf{x})}(1 - y)^{h(\mathbf{x})}$

3. (N) Na skupu označenih primjera $\mathcal{D}$ trenirali smo model logističke regresije. Dobili smo neki vektor težina $\mathbf{w}$ i pomak $w_0 = 0.15$. Tako naučenom modelu neki primjer $\mathbf{x}$, čija je oznaka u skupu primjera $y = 0$, nanosi gubitak unakrsne entropije od $L(0, h(\mathbf{x})) = 0.274$. **Koliki gubitak unakrsne entropije bi nanosio primjer x kada bismo njegove značajke pomnožili sa dva i promijenili mu oznaku**?

   A 4.03 B 2.54 C 7.11 D 1.19

4. (N) Na skupu $\mathcal{D}$ označenih primjera trenirali smo model binarne logističke regresije. Naknadno smo uočili da jedan primjer iz skupa $\mathcal{D}$ modelu nanosi razmjerno velik gubitak. Konkretno, iznos gubitka za dotični primjer je $L(y, h(\mathbf{x})) = 1.20$. Ispostavilo se da je taj primjer pogrešno označen. **Koliko bi iznosio gubitak modela na istom ovom primjeru, ako bismo sada naknadno promijenili njegovu oznaku, ali model ostavili nepromijenjenim?**

   A 0.70 B 0.28 C 0.36 D 0.52

5. (N) Model logističke regresije treniramo stohastičkim gradijentnim spustom. Primjere iz dvodimenzijskog ulaznog prostora preslikali smo u prostor značajki funkcijom

$$\boldsymbol{\phi}(\mathbf{x}) = (1, x_1, x_2, x_1x_2)$$

   U jednoj iteraciji treniranja modela vektor parametara jednak je

$$\mathbf{w} = (0.2, 0.5, -1.1, 2.7)$$

Koliko u toj iteraciji iznosi L_2-norma gradijenta gubitka za primjer $(\mathbf{x}, y) = ((-0.5, 2), 1)$?

A 0.70   B 2.48   C 1.28   D 4.00

6. (P) Na skupu označenih primjera treniramo tri modela: (1) model neregularizirane logističke regresije (NR), (2) model L2-regularizirane logističke regresije (L2R) i (3) perceptron s funkcijom preslikavanja. Sva tri modela koriste iste značajke. Za sva tri algoritma promatramo iznos empirijske pogreške učenja kroz iteracije optimizacijskog postupka. Nakon određenog broja iteracija, algoritam perceptrona uspješno se zaustavio s rješenjem. **Kako se u ovom slučaju ponaša empirijska pogreška učenja kroz iteracije za dva spomenuta modela logističke regresije, NR i L2R?**

   A Pogreška učenja modela NR nakon određenog broja iteracije doseže nulu, dok pogreška učenja modela L2R najprije pada pa raste

   B Pogreške učenja modela NR i modela L2R dosežu nulu, ali modelu L2R za to treba više iteracija

   C Pogreške učenja modela NR i modela L2R obje stagniraju nakon određenog broja iteracija, ali modelu NR za to treba više iteracija

   D Pogreška učenja modela NR asimptotski teži nuli, dok pogreška učenja modela L2R nakon određenog broja iteracija stagnira

7. (P) Razmotrimo sljedeći skup označenih primjera:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((-2,0),-1),((-1,0),+1),((1,0),+1),((2,0),-1))\}$$

   Nad ovim skupom treniramo dva modela: perceptron (P) i neregulariziranu logističku regresiju (LR). Pored toga, razmatramo tri funkcije preslikavanja:

$$\begin{aligned}\boldsymbol{\phi_0}(\mathbf{x}) &= (1, x_1, x_2)\\ \boldsymbol{\phi_1}(\mathbf{x}) &= (1, x_1, x_2, x_1^2, x_2^2)\\ \boldsymbol{\phi_2}(\mathbf{x}) &= (1, x_1, x_2, x_1 x_2)\end{aligned}$$

   Ukupno, dakle, isprobavamo šest kombinacija modela i funkcije preslikavanja. **Za koje će algoritme (model+preslikavanje) empirijska pogreška na skupu za učenje konvergirati?**

   A P+$\boldsymbol{\phi_0}$ B P+$\boldsymbol{\phi_2}$ C LR+$\boldsymbol{\phi_1}$ D LR+$\boldsymbol{\phi_2}$