

K-NN

28.

(Adevărat sau Fals?)

◦ CMU, 2010 fall, Ziv Bar-Joseph, midterm, pr. 1.be

Care dintre următoarele afirmații sunt adevărate pentru clasificatorii k -NN?
(Justificați pe scurt răspunsul, în dreptul fiecărui punct.)

- F a. Acuratețea la antrenare crește pe măsură ce crește valoarea lui k .
- F b. Granița de decizie este mai netedă (engl., smoother) pe măsură ce valoarea lui k scade.
- A c. k -NN nu necesită o procedură explicită de antrenare.
- F d. Granița de decizie este liniară.
- F e. Este posibil ca un clasificator binar 1-NN să clasifice întotdeauna orice instanță de test ca fiind pozitivă, chiar dacă în setul de date de antrenament există instanțe negative.

multe
călătorii
II
menținere

⁴⁷⁶Adică, atunci când $n \rightarrow \infty$, unde n este numărul de instanțe de antrenament.

469

- a) *Uneori, dec. la antrenare scade pe măsură ce crește k -ul*
 ◦ pt $k=1$, modelul men. complex setul de date =>
 are o acc. de 100% la antrenare.
- b) *Granița decizie mai netedă cu măsură cu $k \rightarrow$ crește*
- c) *Mai susținute → are o formă complexă*
- d) *Dacă avem cel puțin o inst. negativă și avem o inst. de test apropiată 1-NN de un nod negativ din datsetul de antrenare, atunci acesta va fi clasificat negativ.*
- e) *Dacă inst. / date sunt consistenti → 1 ab*
 ↳ modelăză adic. pt. set inconsistent

+	+
-	
+	+

29.

- (Întrebări calitative despre design-ul unor experimente din Învățarea Automată: OK ori ...problematic?)
- CMU, 2009, Geoff Gordon, midterm exam, pr. 3

Fiecare din punctele de mai jos prezintă pe scurt design-ul unui experiment practic de învățare automată. Analizați fiecare din aceste cazuri, indicând apoi dacă respectivul experiment este *ok* ori *problematic* (incercuți varianta pe care o alegeti). Dacă este *problematic*, identificați TOATE defectele [de concepție ale] design-ului respectiv.

a. O echipă de proiectare raportează o eroare mică la antrenare și susține că metoda folosită este bună.

Ok
Problematic → Modelul referitor la *overfitting*.

b. O echipă de proiectare susține că este un mare succes faptul că a obținut 98% acuratețe la antrenare pentru un task de clasificare binară care are următorul specific: unul din cele două cazuri se întâlnește foarte rar comparativ cu celălalt caz. (O astfel de problemă o constituie, de exemplu, identificarea tranzacțiilor bancare frauduloase.) Datele lor au constat din 50 de exemple pozitive și 4950 de exemple negative.

Ok
Problematic → Dacă modelul prezice în beneficiul majorității, cu acestă acuratețe foarte bună, deci ar trebui învățămōn să rezolvă corect probleme.

c. O echipă de proiectare și-a împărțit datele de care dispune în date de antrenament și date de test. Folosind datele de antrenament, ei au construit un *model* de clasificare caracterizat de anumiți *parametri*. Apoi, făcând *cross-validation*, au ales cea mai bună setare a parametrilor. La final, au raportat eroarea obținută pe *datele de test*.

Ok
Problematic

d. O echipă de proiectare a efectuat o procedură de *selecție a atributelor* (engl., features) pe toate datele și apoi a redus setul mare de atribute la un set mai mic. După aceea, membrii echipei au împărțit datele în date de test și date de antrenament. Au construit *modelul* de clasificare pe datele de antrenament folosind mai multe setări ale parametrilor modelului, și au raportat cea mai bună eroare la testare pe care au obținut-o.

Ok
Problematic

selecția atributelor se face strict pe datele de antrenament, nu pe cele de validare.
⇒ trb. să răspundem că datele de testare sunt datele de validare.

Pregresie logistică



32.

(Funcția sigmoidală / logistică: definiție și proprietăți de bază)

• Liviu Ciortuz, 2024

Considerăm *funcția sigmoidală* (sau *logistică*)

$$\sigma : \mathbb{R} \rightarrow (0, 1), \text{ definită prin } \sigma(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z}, \text{ pentru orice } z \in \mathbb{R}.$$

a. Elaborați graficul acestei funcții. Dați toate justificările necesare. (*Indicație:* O parte dintre ele sunt enunțate la punctele b.i și b.ii de mai jos.)

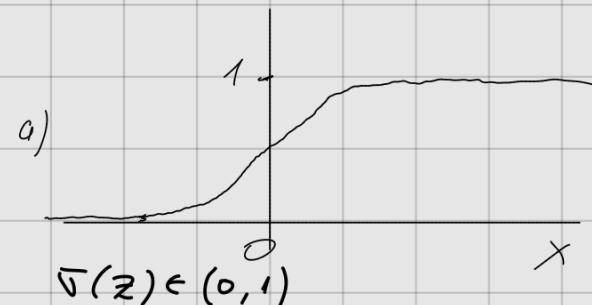
b. Demonstrați următoarele trei proprietăți:

- i. $\sigma(-z) = 1 - \sigma(z)$.
- ii. $\sigma'(z) = \sigma(z)[1 - \sigma(z)]$.

iii. Funcția σ este inversabilă (veți justifica de ce este așa!), așadar există funcția inversă $\sigma^{-1} : (0, 1) \rightarrow \mathbb{R}$, astfel încât

$$\sigma^{-1}(\sigma(z)) = \sigma(\sigma^{-1}(z)) = z \text{ pentru orice } z \in \mathbb{R}. \quad (202)$$

³⁸⁸ Această metodă este descrisă în secțiunea 4.1.3 din cartea *Pattern Recognition and Machine Learning* de Ch. Bishop (Springer, 2006) și în secțiunea 4.2 din cartea *The Elements of Statistical Learning* de T. Hastie, R. Tibshirani and J. Friedman (Springer, 2009).



$\sigma(z) \in (0, 1)$
pentru $\sigma(0.5) = 0.5$
- Pd. continuu, derivabilă.
- Pd. liniară,

Arătați mai întâi că

$$\sigma^{-1}(z) = \ln \frac{z}{1-z} \text{ pentru orice } z \in (0, 1).$$

Aceasta este așa-numita funcție *logit*.

Verificați apoi că dubla egalitate (202) este satisfăcută.

La final, faceți graficul funcției logit.

b) i.

$$\bar{\sigma}(-z) = 1 - \bar{\sigma}(z)$$

$$\bar{\sigma}(z) = \frac{1}{1+e^{-z}}$$

calc:

$$\bar{\sigma}(-z) = \frac{1}{1+e^{-(-z)}} = \frac{1}{1+\frac{1}{e^z}} = \frac{1}{\frac{e^z+1}{e^z}} = \frac{e^z}{e^z+1}$$

7

$$= 1 - \frac{1}{1+e^{-z}} = 1 - \bar{\sigma}(z) \geq 0, \forall z.$$

ii.

$$\bar{\sigma}'(z) = \bar{\sigma}(z)(1-\bar{\sigma}(z))$$

$$\left(\frac{1}{x}\right)' = -\frac{1}{x^2} \cdot (x')$$

calc:

$$\bar{\sigma}'(z) = \left(\frac{1}{1+e^{-z}}\right)' = -\frac{1}{(1+e^{-z})^2} \cdot (1+e^{-z})' = -\frac{1 \cdot e^{-z} \cdot (-1)}{(1+e^{-z})^2} =$$

$$= \frac{e^{-z}}{(1+e^{-z})} \cdot \frac{1}{(1+e^{-z})} = \bar{\sigma}(z) \cdot (1-\bar{\sigma}(z))$$

dini. || 1/sim $\bar{\sigma}(z)$

iii.

$$\bar{\sigma}^{-1}: (0, 1) \rightarrow \mathbb{R}, \quad \bar{\sigma}^{-1}(\bar{\sigma}(z)) = \bar{\sigma}(\bar{\sigma}^{-1}(z)) = z, \forall z \in \mathbb{R}$$

- Fct. sigmoid este - monoton crescătoare \Rightarrow este bijecție \Rightarrow deci inversabilă
 - continuă și derivabilă
 (compr. din pct. elementare)
 - are codomeniul bine definit, $\forall x \in \mathbb{R}, \exists$ un
 $y \in (0, 1)$

$$\text{Auszamm c: } \bar{\nu}^{-1}(z) = \ln \frac{z}{1-z}, \quad \forall z \in (0, 1)$$

\hookrightarrow def. Logist.

$$\bar{\nu}(x) = \frac{1}{1+e^{-x}}, \quad \text{Für } z = \bar{\nu}(x)$$

$$\Rightarrow z = \frac{1}{1+e^{-x}} \Rightarrow z(1+e^{-x}) = 1 \Rightarrow z + z \cdot e^{-x} = 1 \Rightarrow$$

$$\Rightarrow z \cdot e^{-x} = 1-z \Rightarrow e^{-x} = \frac{1-z}{z} \quad | \ln \Rightarrow -x = \ln \frac{1-z}{z} \quad | f()$$

$$\Rightarrow x = \ln \frac{z}{1-z}$$

$$z = \bar{\nu}(x) \quad | \bar{\nu}^{-1} \Rightarrow \bar{\nu}^{-1}(z) = \bar{\nu}^{-1}(\bar{\nu}(x)) = x$$

\hookrightarrow contr. Invertierbarkeit $\Rightarrow \bar{\nu}^{-1}(\bar{\nu}(x)) = x$

$$\text{Verifikation: } \bar{\nu}^{-1}(\bar{\nu}(z)) = \bar{\nu}(\bar{\nu}^{-1}(z)) = z$$

$$\text{Dem: } \bar{\nu}^{-1}(\bar{\nu}(z)) = z$$

$$\bar{\nu}^{-1}(\bar{\nu}(z)) = \bar{\nu}^{-1}\left(\frac{1}{1+e^{-z}}\right) = \ln \frac{\frac{1}{1+e^{-z}}}{1 - \frac{1}{1+e^{-z}}} = \ln \frac{\frac{1}{1+e^{-z}}}{\frac{e^{-z}}{1+e^{-z}}} =$$

$$= \ln \frac{\frac{1}{1+e^{-z}}}{\frac{1}{1+e^{-z}}} \cdot \frac{1+e^{-z}}{e^{-z}} = \ln \frac{1}{e^{-z}} = \ln \frac{1}{\frac{1}{e^z}} = \ln e^z = z$$

\nearrow Adversar!

$$\text{Dem: } \bar{\nu}(\bar{\nu}^{-1}(z)) = z$$

$$\log_x \log_x y = y$$

$$\bar{\nu}(\bar{\nu}^{-1}(z)) = \bar{\nu}\left(\ln \frac{z}{1-z}\right) = \frac{1}{1+e^{\ln \frac{z}{1-z}-1}} = \frac{1}{1+\left(\frac{z}{1-z}\right)^{-1}}$$

$$= \frac{1}{\frac{z}{1+z}} = \frac{1}{\frac{z+1-z}{z}} = \frac{1}{\frac{1}{z}} = z \quad \text{Adeseorat.}$$

33.
Liu

(Regresia logistică: particularizare în \mathbb{R}^2 ;
deducerea regulilor de actualizare pentru metoda gradientului;
regularizare L_1)

prelucrare de Liviu Ciortuz, după
CMU, 2018 spring, Nina Balcan, HW3, pr. 1, 4

A. În prima parte a acestei probleme, veți elabora algoritmul de regresie logistică bazat pe metoda gradientului, în cazul bidimensional. Considerăm setul de date de antrenament $\{(x^i, y^i), i = 1, \dots, n\}$ în care fiecare $x^i \in \mathbb{R}^2$ este un vector de trăsături, iar $y^i \in \{0, 1\}$ este o etichetă binară. Presupunând că folosim un model parametrizat de forma

$$p(y=1|x; w) = \frac{1}{1 + \exp(-w_0 - w_1 x_1 - w_2 x_2)} = \frac{\exp(w_0 + w_1 x_1 + w_2 x_2)}{1 + \exp(w_0 + w_1 x_1 + w_2 x_2)},$$

obiectivul nostru este să găsim valorile w_i ale parametrilor w_i care maximizează verosimilitatea condițională (M(C)LE) a setului de date de antrenament.

a. Mai jos, vom face *noi* calculul pentru log-verosimilitatea condițională a datelor de antrenament. Relativ la acest calcul, *dumneavoastră* veți elabora câte o scurtă justificare pentru fiecare linie din demonstrație.

$$\ell(w) = \ln \prod_{j=1}^n p(y^j|x^j, w) - \text{P.d. log verosimilitate. Aplic la } \ell(w) \text{ (203)}$$

$$= \sum_{j=1}^n \ln p(y^j|x^j, w) - \text{transform produs în sumă, conf. prop. log.} \quad (204)$$

$$= \sum_{j=1}^n \ln(p(y^j = 1|x^j, w)^{y^j} p(y^j = 0|x^j, w)^{1-y^j}) \text{ conf. dist. Bernoulli, iau} \quad (205)$$

$$= \sum_{j=1}^n [y^j \ln p(y^j = 1|x^j, w) + (1 - y^j) \ln p(y^j = 0|x^j, w)] \text{ ceea ce} \quad (206)$$

$$= \sum_{j=1}^n [y^j \ln \frac{\exp(w_0 + w_1 x_1^j + w_2 x_2^j)}{1 + \exp(w_0 + w_1 x_1^j + w_2 x_2^j)} + \text{analog form. } P(Y=1|X, w) \text{ de mai sus}] \quad (207)$$

$$= \sum_{j=1}^n [y^j \ln (\exp(w_0 + w_1 x_1^j + w_2 x_2^j)) + \ln \frac{1}{1 + \exp(w_0 + w_1 x_1^j + w_2 x_2^j)}] \quad (208) \text{ sumă cu garantată din faza lini, și fac calculele.}$$

$$= \sum_{j=1}^n [y^j (w_0 + w_1 x_1^j + w_2 x_2^j) - \ln(1 + \exp(w_0 + w_1 x_1^j + w_2 x_2^j))]. \quad (209)$$

\hookrightarrow calc. lini (exp(ceva)) = ceva și lini inversor fractiei din al 2-lea ln.

(Regresia liniară și regresia logistică:
definiții „revizitate“,
o interesantă proprietate comună)

■ □ ● ○ CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW2, pr. 4

Dată fiind o instanță (sau, un input) $X \in \mathbb{R}^d$ împreună cu „răspunsul“ / outputul corespunzător $Y \in \mathbb{R}$, regresia liniară [cu „zgomot“ gaussian] construiește un model de forma

$$Y|X \sim \text{Normal}(\mu(X), \sigma^2),$$

unde media $\mu(X)$ este o funcție liniară de componente / atributelor inputului: $\mu(X) = \theta^\top X = \theta_0 + \theta_1 X_1 + \dots + \theta_d X_d$.

Dacă $Y \in \{0, 1\}$, regresia logistică — care, spre deosebire de regresia liniară servește pentru clasificare — modelează outputul Y astfel:

$$Y|X \sim \text{Bernoulli}(h_\theta(X)),$$

unde $h_\theta(X)$, parametrul acestei distribuții Bernoulli, este obținut din $\theta^\top X$ aplicând funcția logistică / sigmoidală:

$$h_\theta(X) = g(\theta^\top X),$$

unde prin g am notat funcția logistică

$$g(z) \stackrel{\text{def.}}{=} \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

sau, echivalent, folosind funcția logit (care este inversa funcției logistică / sigmoidale):³⁹⁴

$$\text{logit}(h_\theta(X)) \stackrel{\text{def.}}{=} \ln \frac{h_\theta(X)}{1 - h_\theta(X)} = \theta^\top X$$

Comentariu: Definițiile date mai sus pun în evidență un anumit „paralelism“ pentru cele două modele de regresie. În această problemă,

— mai întâi, veți putea vedea încă o similaritate, și anume între vectorii gradient corespunzători funcțiilor de log-verosimilitate condițională pentru cele două metode de regresie. Ne referim aici la expresia $\nabla_{\theta} \ell(\theta) = \sum_{i=1}^n (y_i - h_\theta(x_i))x_i$ (care corespunde relației (181) de la problema 13) pentru gradientul regresiei logistic și la expresia $\nabla_{\theta} \ell(\theta) = \sum_{i=1}^n (y_i - \theta^\top x_i)x_i$ pentru gradientul regresiei liniare, conform termenului principal din relația (164), care a fost obținută la problema 6;

— iar acum veți demonstra o proprietate de tip probabilist, care se scrie în mod identic(!) pentru cele două modele de regresie și care folosește media condițională a outputului Y în raport cu inputul X și cu $\hat{\theta}$, estimarea în sens MLE pentru parametrul θ .

Veți considera setul de date de antrenament $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$.

Demonstrați că pentru fiecare dintre cele două modele de regresie de mai sus, estimarea de verosimilitate maximă a parametrului θ (notație: $\hat{\theta}$) satisfac următoarea proprietate:

$$\sum_{i=1}^n y_i x_i = \sum_{i=1}^n E[Y|X = x_i, \theta = \hat{\theta}] x_i.$$

³⁹⁴Intr-adevăr, $\sigma(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z} = y \in (0, 1) \Rightarrow e^z(1 - y) = y \Rightarrow z = \ln \frac{y}{1 - y}$.

Pentru Regresie liniară

$$\begin{aligned}
 \ell(\theta) &= \ln \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{(y_i - \mu(x_i))^2}{2\sigma^2} \right) \right) \\
 &= \sum_{i=1}^n \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{(y_i - \theta^\top x_i)^2}{2\sigma^2} \right) \right) \\
 &= -n \ln (\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta^\top x_i)^2 \\
 &= -n \ln (\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta^\top x_i)^2 \\
 &= -n \ln (\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta^\top x_i)^T (y_i - \theta^\top x_i) \\
 \nabla_{\theta} \ell(\theta) &= \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta^\top x_i) x_i
 \end{aligned}$$

Pentru regresia logistică

$$\ln \frac{h_\theta(X)}{1-h_\theta(X)} = \theta^T X \Leftrightarrow e^{\theta^T X} = \frac{h_\theta(X)}{1-h_\theta(X)} \Leftrightarrow e^{\theta^T X} = h_\theta(X)(1+e^{\theta^T X})$$

$$h_\theta(X) = \frac{e^{\theta^T X}}{1+e^{\theta^T X}} = \frac{1}{1+e^{-\theta^T X}} \quad \text{și} \quad 1-h_\theta(X) = \frac{1}{1+e^{\theta^T X}}$$

$Y|X \sim \text{Bernoulli}(h_\theta(X))$ înseamnă că $P(Y=1|X)=h_\theta(X)$ și $P(Y=0|X)=1-h_\theta(X)$

care sunt fi scris echivalent ca:

$$P(Y=y|X) = h_\theta(X)^y (1-h_\theta(X))^{1-y} \text{ for all } y \in \{0,1\}$$

Deci, în acest caz funcția conditională log-likelihood este:

$$\begin{aligned} l(\theta) &= \ln \left(\prod_{i=1}^m \{ h_\theta(x_i)^{y_i} (1-h_\theta(x_i))^{1-y_i} \} \right) \\ &= \sum_{i=1}^m \{ y_i \ln h_\theta(x_i) + (1-y_i) \ln (1-h_\theta(x_i)) \} \\ &= \sum_{i=1}^m \{ y_i (\theta^T x_i) + \ln (1-h_\theta(x_i)) + (1-y_i) \ln (1-h_\theta(x_i)) \} \\ &= \sum_{i=1}^m \{ y_i (\theta^T x_i) - \ln (1+e^{\theta^T x_i}) \} \end{aligned}$$

$$\nabla_\theta l(\theta) = \sum_{i=1}^m (y_i x_i - \frac{e^{\theta^T x_i}}{1+e^{\theta^T x_i}} x_i) = \sum_{i=1}^m (y_i - h_\theta(x_i)) x_i$$

Deci: Regresie liniară

$$\nabla_\theta l(\theta) = 0 \Rightarrow \sum_{i=1}^m y_i x_i = \sum_{i=1}^m (\theta^T x_i) x_i$$

Prințind: $Y|X \sim \text{Normal}(\mu(X), \sigma^2)$

$$E[Y|X=x_i, \theta=\hat{\theta}] = \mu(x_i) = \hat{\theta}^T x_i$$

$$\text{So: } \sum_{i=1}^m y_i x_i = \sum_{i=1}^m E[Y|X=x_i, \theta=\hat{\theta}] x_i.$$

Pentru logistic regression

$$\nabla_\theta l(\theta) = 0 \Rightarrow \sum_{i=1}^m y_i x_i = \sum_{i=1}^m h_\theta(x_i) x_i$$

Since $Y|X \sim \text{Bernoulli}(h_\theta(X))$, $\hat{\theta}^T x_i$:

$$E[Y|X=x_i, \theta=\hat{\theta}] = h_\theta(x_i) = \frac{e^{\hat{\theta}^T x_i}}{1+e^{\hat{\theta}^T x_i}}$$

$$\text{So: } \sum_{i=1}^m y_i x_i = \sum_{i=1}^m E[Y|X=x_i, \theta=\hat{\theta}] x_i$$

(Găsirea optimului unei funcții reale de gradul al doilea, folosind metoda analitică, metoda gradientului descendente și metoda lui Newton)

■ University of Utah, 2008 fall, Hal Daumé III, HW4, pr. 1

Să presupunem că dorim să găsim minimul funcției $f(x) = 3x^2 - 2x + 1$.

a. Verificați că această funcție este convexă.

b. Ca urmare a punctului precedent, rezultă că funcția f are un minim global. Găsiți acest minim, folosind cunoștințe de analiza matematică.

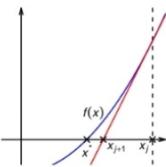
c. La acest punct vom căuta optimul funcției f aplicând algoritmul gradientului descendente.¹⁸¹ Efectuați trei pași ai acestui algoritm pornind de la punctul initial $x_0 = 1$ și folosind rata de învățare $\eta = 0.1$. Cât de aproape a ajuns algoritmul de soluția reală?

d. În sfârșit, căutați din nou optimul funcției f aplicând de această dată metoda lui Newton,¹⁸² pornind tot de la punctul $x_0 = 1$. Ce observați? (Câte iterații sunt necesare pentru ca algoritmul să conveară la soluția optimă?)

Comentariu [Metoda lui Newton]:

În forma sa cea mai simplă, metoda lui Newton — cunoscută în literatură de specialitate și sub numele de *metoda tangentei* — are ca obiectiv identificarea rădăcinii (sau, a rădăcinilor) unei funcții f care este convexă și derivabilă. (Rădăcinile unei funcții sunt acelă valori x^* ale argumentului funcției pentru care se anulează funcția respectivă, adică $f(x^*) = 0$.) În acest scop, adică pentru afilarea rădăcinilor unei funcții, metoda lui Newton, care este o metodă iterative, folosește următoarea relație de „actualizare“:

$$x_{j+1} = x_j - \frac{f(x_j)}{f'(x_j)},$$



cu x_0 ales în mod arbitrar. Facem *observația* că metoda lui Newton poate fi aplicată (modificată ușor) și în scopul găsirii optimului unei funcții. În acest caz, se căută valorile argumentului x pentru care se anulează prima derivată a funcției obiectiv f . Prin urmare, pentru a putea aplica metoda lui Newton se cere ca funcția f să fie dublu derivabilă (adică să existe atât prima cât și cea de-a doua derivată a lui f). Relația de actualizare acum devine:

$$x_{j+1} = x_j - \frac{f'(x_j)}{f''(x_j)}. \quad (106)$$

La problema de față, se folosește metoda lui Newton în varianta aceasta, adică pentru afilarea optimului (mai precis, a minimului) unei funcții.

Facem *observația* că formula (106) se folosește exact în aceeași formă și atunci când — în locul minimului — se căută maximul unei funcții, folosind *metoda lui Newton*. (Pentru a vă convinge, este suficient să înlocuiți f cu $-f$ în

¹⁸¹Pentru o prezentare formală a acestui algoritm, vezi enunțul problemei 164.

¹⁸²Isaac Newton (1642-1727) a fost un matematician, fizician, astronom, alchimist și teolog englez.

a) $\frac{\partial f}{\partial x} = 6x - 2$

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial(6x - 2)}{\partial x} = 6 > 0 \Rightarrow \text{fd. convexă}$$

b) Egalem derivata cu 0?

$$\frac{\partial f}{\partial x} = 0 \Rightarrow 6x - 2 = 0 \Rightarrow x = \frac{1}{3}$$

$$f\left(\frac{1}{3}\right) = 3 \cdot \frac{1}{3} - \frac{2}{3} + 1 = 1 - \frac{1}{3} = \frac{2}{3}$$

fd. de min. local este $\left(\frac{1}{3}, \frac{2}{3}\right)$

$$f(x) = \frac{1}{1+e^{-x}} \quad f'(x) = \frac{e^{-x}}{(1+e^{-x})^2}$$

c) Gradient descend: (met. numerică)

$$x_{(1)} \leftarrow x_{(0)} - \Delta x$$

$$\Delta x = \eta \cdot \frac{\partial f}{\partial x}$$

$$\text{fd: } x_0 = 1$$

$$x_{(1)} \leftarrow x_{(0)} - 0.1 \cdot (6 \cdot x_0 - 2) = 1 - 0.4 = 0.6$$

$$x_{(2)} \leftarrow x_{(1)} - \eta \cdot \frac{\partial f}{\partial x} = 0.6 - 0.1 \cdot (6 \cdot 0.6 - 2) = 0.44$$

$$x_{(3)} \leftarrow x_{(2)} - 0.1 \cdot (6 \cdot x_2 - 2) = 0.44 - 0.1 \cdot (6 \cdot 0.44 - 2) = 0.376$$

Diferența dintre punctul de minim și punctul în care am ajuns este

$$0.376 - 0.333 = 0.043$$

d)

$$x^{(t)} \leftarrow x^{(t-1)} - \frac{f'(x_0)}{f''(x_0)}$$

$$x_0 = 1 \quad x_1 = x_0 - \frac{f'(x_0)}{f''(x_0)} = 1 - \frac{6 \cdot 1 - 2}{6} = 1 - \frac{2}{3} = \frac{1}{3}$$

metoda optimă

164.

(Metoda gradientului: exemple de aplicare,
și un rezultat teoretic: [relativ la] convergență)

□ • CMU, 2019 spring, Nina Blacan, HW3, pr. 1

Fie $f : \mathbb{R}^d \rightarrow \mathbb{R}$ o funcție derivabilă. Vă reamintim că algoritmul *gradientului descendente* — la care ne vom referi aici în forma abreviată, GD — pornește de la un anumit punct $x^{(0)}$ notat $(x_1^{(0)}, \dots, x_d^{(0)}) \in \mathbb{R}^d$, iar după aceea algoritmul „actualizează” în mod iterativ poziția $x^{(k)}$, folosind o *regulă de actualizare*, care se exprimă astfel pentru coordonata i (unde $i \in \{1, \dots, d\}$):

$$x_i^{(k+1)} \leftarrow x_i^{(k)} - \eta \frac{\partial}{\partial x_i} f(x^{(k)}).$$

Aici $\frac{\partial f}{\partial x_i} : \mathbb{R}^d \rightarrow \mathbb{R}$ este derivata parțială a funcției f în raport cu coordonata i , apoi $\frac{\partial}{\partial x_i} f(x^{(k)})$ reprezintă valoarea acestei derivate parțiale în punctul $x^{(k)}$, iar $\eta > 0$ se numește *rata de învățare* (engl., learning rate).

A. Exemple de aplicare a metodei GD...

a. ...în \mathbb{R}

Fie funcția $f(x) = 4x^2 - 2x + 1$ și rata de învățare $\eta = 0.1$. Mai întâi, scrieți expresia derivatei $\frac{\partial}{\partial x} f(x)$. Apoi, pornind de la $x^{(0)} = 1$, pentru fiecare din pașii $k = 0, 1$ și 2 ai algoritmului GD, scrieți vectorul gradient $\frac{\partial}{\partial x} f(x^{(k)})$, noua poziție $x^{(k+1)}$, precum și noua valoare a funcției, $f(x^{(k+1)})$.

²⁷⁹Pentru punctele a și c, puteți vedea graficele din figura de la problema 88.

254

$$\frac{\partial f(x)}{\partial x} = 8x - 2$$

$$x_1 = x_0 - \eta \frac{\partial f(x_0)}{\partial x} = 1 - 0.1 \cdot 6 = 0.4$$

$$x_2 = 0.4 - 0.1(1.2) = 0.4 - 0.12 = 0.28$$

$$x_3 = 0.28 - 0.1(0.28) = 0.28 - 0.028 = 0.256$$

b. ...în \mathbb{R}^2

Fie $f(x_1, x_2) = x_1^2 + \sin(x_1 + x_2) + x_2^2$. Mai întâi, scrieți regula de actualizare pentru gradientul descendente, folosind o rată de învățare oarecare, $\eta > 0$. După aceea, pentru fiecare dintre următoarele *două cazuri* faceți grafice pentru funcția f , care va fi considerată ca fiind definită pe domeniul $[-4, +4] \times [-4, +4]$, folosind [soft pentru] *diagrame de izocontur*,²⁸⁰ precum și săgeți de la o poziție a algoritmului GD către următoarea poziție.²⁸¹

- *Cazul i:* Determinați punctele (x_1, x_2) pentru primii 10 pași făcuți de GD, începând cu $(x_1^{(0)}, x_2^{(0)}) = (3, -3)$ și folosind $\eta = 0.4$,

- *Cazul ii:* Determinați punctele (x_1, x_2) pentru primii 10 pași făcuți de GD, începând cu $(x_1^{(0)}, x_2^{(0)}) = (3, -3)$ și folosind $\eta = 0.8$,

Ce observați?

$$i) \quad (x_1^0, x_2^0) = (3, -3)$$

$$(x_1^1, x_2^1) = \begin{bmatrix} x_1^0 \\ x_2^0 \end{bmatrix} = \begin{bmatrix} 3 \\ -3 \end{bmatrix}$$

$$(x_1^2, x_2^2) = \begin{bmatrix} 0.2 \\ -5.6 \end{bmatrix} = 0.4 \begin{bmatrix} 0.4 + \cos(-5.6) \\ -11.6 + \cos(-5.6) \end{bmatrix} = \begin{bmatrix} \dots \\ \dots \end{bmatrix}$$

$$\frac{\partial f}{\partial x_1} = 2x_1 + \cos(x_1 + x_2)$$

$$\frac{\partial f}{\partial x_2} = 2x_2 + \cos(x_1 + x_2)$$

$$- 0.4 \cdot \frac{\partial f}{\partial x_1}(x^{(k)})$$

$$- 0.4 \cdot \begin{bmatrix} 7 \\ 7 \end{bmatrix} = \begin{bmatrix} 3 \\ -3 \end{bmatrix} - \begin{bmatrix} 2.8 \\ 2.8 \end{bmatrix} = \begin{bmatrix} 0.2 \\ -5.6 \end{bmatrix}$$

$10X$

ii) learning rate mai mare (probabil recede mai rapid)

B. Analiza algoritmului GD

Metoda gradientului descendente este în sine un *algoritm de căutare locală*: la fiecare pas, el folosește informația din punctul curent (și anume, vectorul gradient) pentru a determina mișcarea pe care trebuie să o facă, în direcția minimizării funcției (care este tocmai direcția descreșterii vectorului gradient). Algoritmul GD nu are cunoștință despre modul în care se comportă funcția pe întreg domeniul de definiție. Totuși, deși GD este un algoritm de căutare locală, dacă f are proprietăți convenabile — mai precis, dacă f este convexă și β -netedă (engl., β -smooth), notiune care este definită mai jos — și are un punct de minim [de abscisă] x^* , atunci există un $\eta > 0$ pentru care $x^{(k)} \rightarrow x^*$ atunci când $k \rightarrow +\infty$. În cele ce urmează vă vom ghida cum să faceți demonstrația acestei proprietăți în cazul unidimensional.

c. [Lema descreșterii]

Presupunem că f este o funcție derivabilă pe tot domeniul de definiție [LC: cu derivata continuă peste tot²⁸²] și β -netedă, unde $\beta > 0$. Prin *definiție*, faptul că f este β -netedă înseamnă că

$$f(y) \leq f(x) + f'(x)(y - x) + \frac{\beta}{2}(y - x)^2 \text{ pentru orice } x \text{ și } y. \quad (146)$$

Stim că $x^{(k+1)} = x^{(k)} - \eta f'(x^{(k)})$ la pasul curent al algoritmului GD. Substituind $x = x^{(k)}$ și $y = x^{(k+1)}$ în relația de mai sus, obținem

$$f(x^{(k+1)}) \leq f(x^{(k)}) + f'(x^{(k)}) \left(x^{(k+1)} - x^{(k)} \right) + \frac{\beta}{2} \left(x^{(k+1)} - x^{(k)} \right)^2.$$

Observație: Conform problemei 78.a, proprietatea 2 (vedeți *Observația* de acolo), faptul că f este convexă (în ipoteza în care f este derivabilă, cu derivata continuă peste tot) este echivalent cu:

²⁸⁰O *diagramă de izocontur* este o modalitate de reprezentare a unor suprafețe, care în mod normal ar trebui să fie reprezentate în spațiu 3D, în spațiu 2D. Ea constă în a desena în plan mai multe curbe, fiecare curbă

$$f(y) \geq f(x) + f'(x)(y - x) \text{ pentru orice } x \text{ și } y. \quad (147)$$

Remarcăți faptul că relația (146) din definiția pentru proprietatea de β -netezire a lui f introduce o margine superioară pentru $f(y)$, pe lângă marginea inferioară din relația (147) dată de proprietatea de convexitate.

Vă cerem să rezolvați următoarele două cerințe:

- i. Folosind regula de actualizare din algoritmul GD, arătați că din relația precedentă rezultă că

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \eta \left(1 - \frac{\eta\beta}{2}\right) (f'(x^{(k)}))^2.$$

- ii. Identificați intervalul în care trebuie să se situeze valorile ratei de învățare η astfel încât să rezulte $f(x^{(k+1)}) < f(x^{(k)})$.²⁸³

NoLo Re

165.

(Metoda gradientului descendente: aplicare / implementare)

• Caltech, 2012, Abu Mostafa, HW5, pr. 4

Considerăm că o anumită funcție de eroare are expresia $E(u, v) = (ue^v - 2ve^{-u})^2$.

a. Calculați derivatiile parțiale ale acestei funcții în raport cu u și respectiv v , adică $\frac{\partial E}{\partial u}(u, v)$ și $\frac{\partial E}{\partial v}(u, v)$.

b. Pentru a identifica minimul funcției de eroare E , veți aplica metoda gradientului descendente, începând cu punctul $(u_0, v_0) = (1, 1)$. Veți folosi *rata de învățare* $\eta = 0.1$.

Câte iterații se efectuează până când valoarea diferenței $E(u_{t+1}, v_{t+1}) - E(u_t, v_t)$ scade pentru prima dată sub pragul de 10^{-14} ? Aveți grijă ca în programul pe care îl veți implementa să folosiți [variabile în] dublă precizie, pentru a putea obține acuratețea cerută.

²⁸³Așadar, vrem să ne asigurăm că valoarea lui η este aleasă astfel încât, la executarea unei iterații a algoritmului GD, valoarea funcției obiectiv să descrească.

²⁸⁴<http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning>.

256

$$a) \frac{\partial E}{\partial u}(u, v) = 2(u e^v - 2v e^{-u}) \cdot (e^v + 2v \cdot e^{-u})$$

$\frac{\partial E}{\partial v}(u, v) = 2(u e^v - 2v e^{-u}) \cdot (u e^v - 2e^{-u})$

$$\frac{\partial E}{\partial v}(u, v) = 2(u e^v - 2v e^{-u}) \cdot (u e^v - 2e^{-u})$$

$$b) (u_1, v_1) = (1, 1) - 0.1 \cdot \frac{\partial E}{\partial u}(u_0, v_0) = (1, 1) - 0.1 \cdot \left[2\left(e - \frac{2}{e}\right) \cdot \left(e + \frac{2}{e}\right) \right]$$

$$= \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 0.1 \cdot \left[2 \left(\frac{e^2 - 2}{e}\right) \left(\frac{e^2 + 2}{e}\right) \right] = \begin{bmatrix} \dots \\ \dots \end{bmatrix} = \begin{bmatrix} \dots \\ \dots \end{bmatrix}$$

✓

169.

(Metoda lui Newton: o proprietate interesantă
în cazul funcțiilor pătratice)prelucrare de Liviu Ciortuz, după
□ • ◯ Stanford, 2009 fall, Andrew Ng, practice midterm, pr. 6.dFie o funcție $f : \mathbb{R}^d \rightarrow \mathbb{R}$ definită prin expresia $f(x) = \frac{1}{2}x^\top Ax + bx + c$, unde A este o matrice simetrică și pozitiv definită.a. Arătați că f este funcție convexă.

(Sugestie: Puteți folosi fie definiția fie proprietățile formulate la problema 78.)

b. Demonstrați că atunci când se folosește metoda lui Newton pentru a afla minimul funcției f , este suficient să se execute o singură iterație. Veți considera că metoda lui Newton face inițializarea cu vectorul 0 (din \mathbb{R}^d).

d)

$$f(x) = \frac{1}{2} x^\top A x + b^\top x + c$$

$$f(x) = \frac{1}{2} [x_1 \ x_2 \ x_3] \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + [b_1 \ b_2 \ b_3] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + c$$

L''

$$\left[\begin{array}{ccc} x_1 a_{11} + x_2 a_{21} + x_3 a_{31} & x_1 a_{21} + x_2 a_{22} + x_3 a_{32} & x_1 a_{31} + x_2 a_{32} + x_3 a_{33} \end{array} \right], \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

()

C =

$$\frac{1}{2} \left(\begin{array}{c} x_1^2 a_{11} + x_1 x_2 a_{21} + x_1 x_3 a_{31} + \\ + x_1 x_2 a_{21} + x_2^2 a_{22} + x_2 x_3 a_{32} + \\ + x_1 x_3 a_{31} + x_2 x_3 a_{32} + x_3^2 a_{33} \end{array} \right) + b_1 x_1 + b_2 x_2 + b_3 x_3 + c$$

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \end{bmatrix} : \quad \frac{\partial f}{\partial x_1} = \frac{1}{2} \cdot (2x_1 a_{11} + 2x_2 a_{21} + 2x_3 a_{31}) + b_1$$

$$\frac{\partial f}{\partial x_2} = \frac{1}{2} (2x_1 a_{21} + 2x_2 a_{22} + 2x_3 a_{32}) + b_2$$

$$\frac{\partial f}{\partial x_3} = \frac{1}{2} (2x_1 a_{31} + 2x_2 a_{32} + 2x_3 a_{33}) + b_3$$

$$H(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_1 \partial x_3} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_3 \partial x_1} & \ddots & \ddots \end{bmatrix}$$

$$\frac{\partial^2 f}{\partial x_1 \partial x_1} = a_{11}, \quad \frac{\partial^2 f}{\partial x_1 \partial x_2} = a_{21}, \quad \frac{\partial^2 f}{\partial x_1 \partial x_3} = a_{31}$$

$$\vdots \quad a_{21} \quad \ddots \quad - \quad - \quad -$$

$$H(x) = \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & \ddots & \vdots \\ a_{31} & \vdots & \ddots \end{pmatrix} = A \quad \left. \begin{array}{l} \Rightarrow H(x) \succ 0 \Rightarrow f \text{- convex} \\ A \text{ pos def} \end{array} \right\}$$

5)

Met. der Newtons:

$$x^+ \leftarrow x_{t-1} - H^{-1} \cdot \frac{\partial f}{\partial x}(x^{t-1})$$

$$A^T (Ax + b)$$

$$y \leftarrow A^T A x + A^T b$$

$$x^+ \leftarrow x_{t-1} - x_{t-1} - H^T s$$

$$X^T \leftarrow -A^{-1} b$$

$$X^{(2)} = -A^{-1} b$$

Solv.

$\frac{\partial f}{\partial X} = 0 \Rightarrow A X + b = 0$

$X = -A^{-1} b$

Deci, algoritmii să opună la o singură iterare

1.1.2 Regresia logistică

13.

(Regresia logistică, chestiuni introductive: estimare MLE; deducerea regulilor de actualizare a parametrilor, folosind metoda gradientului)

formulare de Liviu Ciortuz, după

■ □ ● ○ Stanford, 2016 spring, Chris Piech,
Introduction to Probability for Computer Scientists course (CS109),
Logistic Regression

În învățarea automată, **algoritmii de clasificare** primesc ca date de intrare n instanțe de antrenament care sunt identice și independent distribuite, $(x^{(1)}, y^{(1)})$, $(x^{(2)}, y^{(2)})$, ..., $(x^{(n)}, y^{(n)})$, fiecare vector $x^{(i)}$ având d atribută. În cazul de față vom presupune că $y^{(i)} \in \{0, 1\}$ pentru $i = 1, \dots, n$.

³⁴⁴Vedeți și problema 25.

300

b. Arătați că forma derivatelor parțiale ale funcției de log-verosimilitate condițională în raport cu fiecare componentă w_j a vectorului w este următoarea:

$$\frac{\partial}{\partial w_j} \ell(w) = \sum_{i=1}^n [y^{(i)} - \underbrace{\sigma(w \cdot x^{(i)})}_{P(Y=1|X=x;w)}] x_j^{(i)} \text{ pentru } j = 0, 1, \dots, d. \quad (180)$$

Sugestie: Puteți folosi următoarea proprietate pentru derivata funcției logistică σ în raport cu argumentul ei:

$$\frac{\partial}{\partial z} \sigma(z) = \sigma(z)[1 - \sigma(z)] \text{ pentru } \forall z \in \mathbb{R}.$$

Observație: Din relația (180) urmează că vectorul gradient $\nabla_w \ell(w)$ se poate scrie astfel:

$$\nabla_w \ell(w) = \sum_{i=1}^n [y^{(i)} - \sigma(w \cdot x^{(i)})] x^{(i)}. \quad (181)$$

¹Pentru o ilustrare simplă a modului cum funcționează algoritmul gradientului descendente, vedeți problema 80.c de la capitolul de Fundamente.

²Vedeți pr. 14.

$$\ell(w) = \sum_{i=1}^n (y^i \ln \sigma(w \cdot x^i) + (1-y^i) \ln (1 - \sigma(w \cdot x^i)))$$

$$L(w) = \sum_{i=1}^n \sigma(w \cdot x^i)^{y^i} \cdot (1 - \sigma(w \cdot x^i))^{1-y^i}$$

$$\frac{\partial \ell(w)}{\partial w_j} = \sum_{i=1}^n (\sigma(w \cdot x^i)^{y^i} \cdot (1 - \sigma(w \cdot x^i))^{1-y^i})$$

$$= \sum_{i=1}^n \ln \sigma(w \cdot x^i)^{y^i} + \ln (1 - \sigma(w \cdot x^i))^{1-y^i}$$

$$= -\frac{\partial}{\partial w_i} \left(y^i \ln \sigma(w \cdot x^i) + (1-y^i) \ln (1-\sigma(w \cdot x^i)) \right)$$

$$\sigma'(a \cdot x) = \sigma(a \cdot x)(1-\sigma(a \cdot x)) \cdot (a \cdot x)^t$$

$$= \frac{\partial}{\partial w_i} y^i \ln \sigma(w \cdot x^i) + \frac{\partial}{\partial w_i} (1-y^i) \ln (1-\sigma(w \cdot x^i)) =$$

$$= \frac{y^i}{\sigma(w \cdot x^i)} \sigma'(w \cdot x^i) + \frac{(1-y^i)}{(1-\sigma(w \cdot x^i))} (1-\sigma(w \cdot x^i))^t$$

$$= \frac{y^i}{\sigma(w \cdot x^i)} \cancel{\frac{1}{\sigma(w \cdot x^i)}} \cancel{\sigma(w \cdot x^i)} (1-\sigma(w \cdot x^i)) x^i + \frac{(1-y^i)}{(1-\sigma(w \cdot x^i))} \cancel{(1-\sigma(w \cdot x^i))} \cancel{\sigma(w \cdot x^i)} x^i =$$

$$= y \cdot (1-\sigma(w \cdot x)) x^i - (1-y) \sigma(w \cdot x) \cdot x^i =$$

$$= y x^i - \cancel{y x^i \sigma(w \cdot x)} - x^i \sigma(w \cdot x) + \cancel{y x^i \sigma(w \cdot x)} =$$

$$\sum_{i=1}^n [y - \sigma(w \cdot x)] x^i$$

Nu sun sumo, decarece derivata numei = suno derivatelor

Pentru ca m.c. derivatele pariale pt. cte un $w_j \rightarrow$ suno calc. pt.

a ring, expresie din suno sumo.

Ex 33 / 330

b) $\frac{\partial \ell(w)}{\partial w_i} = \begin{bmatrix} \frac{\partial \ell(w_0)}{\partial w_0} \\ \frac{\partial \ell(w_1)}{\partial w_1} \\ \frac{\partial \ell(w_2)}{\partial w_2} \end{bmatrix}$

(Avizatati)

$$\frac{\partial \ell(w_0)}{\partial w_0} = \sum_{j=1}^m (y^j - p(y^j=1 | X^j; w)) x^j$$

$$\text{ptim: } \ell(w) = \ln \prod_{j=1}^m p(y^j | X^j; w)$$

$$\begin{aligned}
& \frac{\partial}{\partial w_j} \sum_{j=1}^n \ln P(y^i | X^j; w) = \frac{\partial}{\partial w_j} \ln(P(y^i | X^j; w)) = \frac{\partial}{\partial w_j} \ln(p(y=1 | X^j, w) \cdot p(y=0 | X^j, w)) = \\
& = \frac{\partial}{\partial w_j} \ln p(y=1 | X^j; w) + \frac{\partial}{\partial w_j} \ln p(y=0 | X^j; w) = \frac{\partial}{\partial w_j} \ln \frac{\exp(w_0 + w_1 x_1 + w_2 x_2)}{1 + \exp(w_0 + w_1 x_1 + w_2 x_2)} + \\
& + \frac{\partial}{\partial w_j} \ln \left(1 - \frac{\exp(-z)}{1 + \exp(-z)} \right) = \frac{y^i}{\sqrt{z}} \cdot \frac{(1-y^i)}{\sqrt{1-z}} \cdot \frac{\partial}{\partial w_j} (w_0 + w_1 x_1 + w_2 x_2) + \\
& - (1-y^i) \frac{\sqrt{z}(1-\sqrt{z})}{1-\sqrt{z}} \cdot \frac{\partial}{\partial w_j} (w_0 + w_1 x_1 + w_2 x_2) = \\
& = y^i \cdot \frac{1 - \exp(-z)}{1 + \exp(-z)} \cdot X_j^o = (1-y^i) \frac{\exp(-z)}{1 + \exp(-z)} \cdot X_j^o
\end{aligned}$$

My greed Cai-ibine ↑

34

Vo Norma, ~~~