Cursul 21

Regresia liniară şi regresia logistică a două variabile

21.1 Dreapta celor mai mici pătrate şi dreapta de regresie

Termenul de regresie se pare că a fost folosit pentru prima dată în genetică (Francis Galton, 1877) în studiul transmiterii (sau nu) a inteligenţei, de la o generaţie la alta. Termenul a fost apoi folosit în statistică şi în ultimul timp şi în *machine learning*.

Atât în statistică, inginerie, cât şi în *machine learning* se abordează problema predicţiei valorilor unei variabile răspuns sau variabilă dependentă, Y, pe baza valorilor înregistrate de o variabilă intrare sau variabilă predictor, X. Predicţia se efectuează alegând un model ipotetic al dependenţei funcţionale dintre X şi Y, şi construind modelul pe baza unui set de observaţii sau măsurători (x_i, y_i), $i = \overline{1, n}$, asupra perechii de variabile (X, Y).

Pentru a justifica modalitatea de abordare a acestei probleme, reamintim din teoria probabilităţilor că dacă $f_{X,Y}$ este densitatea de probabilitate a vectorului aleator (X, Y), f_X este densitatea marginală a lui X, atunci densitatea variabilei condiţionate $(Y|X = x)$ este (vezi cursul 11, pag. 11):

$$h(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

Astfel valoarea medie a variabilei Y condiţionată de $(X = x)$ este:

$$M(Y|X = x) = \int_{-\infty}^{\infty} yh(y|x)\, dy = \int_{-\infty}^{\infty} y\frac{f_{X,Y}(x, y)}{f_X(x)}\, dy$$

Cum ultima integrală se calculează în raport cu y valoarea integralei este o funcţie de x pe care o notăm $\varphi(x) := M(Y|X = x)$. Această funcţie se numeşte regresia lui Y în raport cu X şi joacă un rol cheie în efectuarea predicţiilor privind valoarea variabilei Y ştiind că $X = x$.

Dacă variabila X ia ca valori lunile anului $1, 2, \ldots, 12$, iar Y este temperatura în acea lună, atunci $M(Y|X = x)$ este temperatura medie în luna x.

Se poate demonstra printr-un calcul simplu că funcţia φ este într-un anume sens cea mai apropiată funcţie de variabila $(Y|X)$, şi anume:

$$\varphi(x) = M(Y|X = x) = argmin_{r(x)} M((Y|X = x) - r(x))^2$$

Adică $\varphi(x)$ este funcţia care minimizează media pătratului erorii (distantei) dintre valoarea variabilei $(Y|X = x)$ şi $r(x)$, unde r este orice funcţie continuă pe $\mathbb{R}$ sau pe mulţimea în care variabila X ia valori.

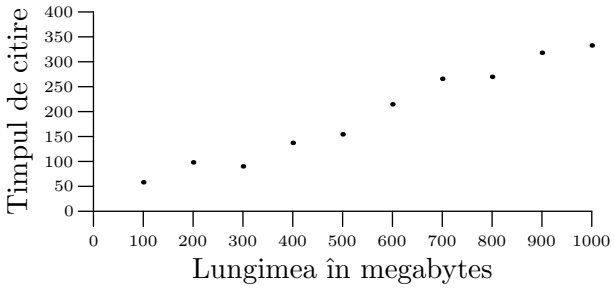

Fig.21.1: Perechile de măsurători (lungime date, timp de citire).

Exemplul 1. Fie Y variabila ce măsoară timpul de citire a unor date stocate pe un anumit tip de DVD şi X lungimea în megabytes a datelor.

Dacă pentru un eşantion de $n = 10$ DVD-uri se înregistrează (în ordine crescătoare) lungimea x, în megabytes, a datelor:

$$100, 200, 300, 400, 500, 600, 700, 800, 900, 1000$$

şi apoi pentru fiecare DVD conţinând informaţie de x_i megabytes se înregistrează timpul de citire, y_i, al acestora, $i = \overline{1, 10}$:

$$58.19, 98.11, 90.14, 137.31, 154.41, 214.69, 265.89, 269.79, 318.05, 332.74$$

se pune problema de a prezice pe baza informaţiilor culese care este timpul de citire a informaţiei de lungime x arbitrară, de pe un DVD din tipul investigat.

În cazul volumului redus de date, reprezentarea punctelor într-un sistem de axe ortogonale ilustrează distribuţia (împrăştierea) punctelor (x_i, y_i), $i = \overline{1, n}$, în plan şi sugerează dacă datele sunt uşor dispersate în jurul unei drepte, parabole sau alta curbă comună. Alegerea modelului funcţional $y = G(x)$ pentru date nu presupune determinarea unei funcţii continue, care interpolează datele, adică $G(x_i) = y_i$, $i = \overline{1, n}$, ci alegerea unei funcţii G_θ dintr-o anumită clasă (de exemplu, funcţii afine, polinomiale, exponenţiale, etc) care depinde câţiva parametri, $\theta_1, \theta_2, \ldots, \theta_d$, ce se estimează din date. Estimarea parametrilor se face în aşa fel încât pierderea sau eroarea cauzată de alegerea modelului G_θ pentru a aproxima valorile y_i prin $G_\theta(x_i)$, pentru intrările x_i, să fie minimă, $i = \overline{1, n}$. Pentru a înţelege această problematică discutăm mai întâi cazul în care modelul ales este liniar, adică $y = G(x) = \beta_0 + \beta_1 x$.

Modelul liniar se alege când datele din eşantion sunt dispuse în jurul unei drepte aşa cum este exemplul observaţiilor asupra lungimii datelor şi timpul de citire (Fig.21.1).

21.1.1 Dreapta celor mai mici pătrate

Problema construirii modelului din date are o abordare deterministă (folosind noţiuni şi rezultate de algebră liniară şi analiză), iar validarea modelului una probabilistă.

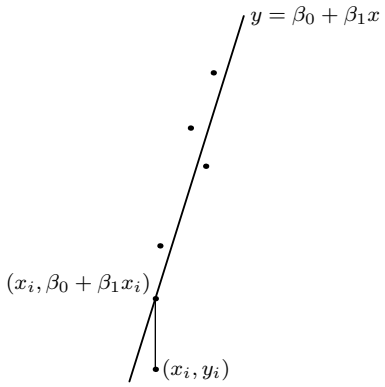

Fig.21.2:

În context determinist, ne punem problema să determinăm dreapta $y = \beta_0 + \beta_1 x$ care "se potriveşte cel mai bine cu datele". Şi anume, determinăm dreapta (mai precis parametrii $\theta_0 = \beta_0$ şi $\theta_1 = \beta_1$ care o definesc) cu proprietatea că suma pătratelor distantelor, dintre punctele (x_i, y_i) şi dreaptă, este minimă, $i = \overline{1, n}$.

Distanţa dintre un punct (x_i, y_i) şi dreaptă se măsoară pe verticală (Fig.21.2), adică distanţa dintre (x_i, y_i) şi punctul de pe dreaptă ce are aceeaşi abscisă $(x_i, \beta_0 + \beta_1 x_i)$. Pătratul acestei distanţe este: $d^2 = (y_i - (\beta_0 + \beta_1 x_i))^2$. Suma acestor distanţe la pătrat depinde doar de β_0 şi β_1, deoarece coordonatele punctelor sunt cunoscute. În concluzie, estimăm dreapta care se potriveşte cel mai bine datelor $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, prin metoda celor mai mici pătrate, adică determinăm dreapta $y = \beta_0 + \beta_1 x$ cu proprietatea că suma pătratelor:

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{21.1}$$

diferenţelor dintre valorile observate y_i şi valorile estimate $\hat{y}_i = \beta_0 + \beta_1 x_i$ este minimă. Funcţia Q se numeşte funcţia eroare asociată datelor, deoarece $(y_i - (\beta_0 + \beta_1 x_i))^2$ este pătratul erorii ce se comite alegând dreapta $y = \beta_0 + \beta_1 x$ ca model predictiv.

Să arătăm că funcţia Q are un singur punct de extrem, $(\hat{\beta}_0, \hat{\beta}_1)$ şi acesta este un minim.

Pentru a determina punctele de extrem ale lui Q calculăm punctele critice, adică punctele ce anulează simultan derivatele parţiale ale funcţiei Q. Dacă datele (x_i, y_i), $i = \overline{1, n}$ nu au aceeaşi abscisă, sistemul:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \tag{21.2}$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \tag{21.3}$$

are soluţia unică:

$$\hat{\beta}_1 = \frac{ss_{xy}}{ss_x}, \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}, \tag{21.4}$$

unde

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{21.5}$$

$$\overline{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} \tag{21.6}$$

$$ss_x = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \tag{21.7}$$

$$ss_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} \tag{21.8}$$

Observăm că soluţia nu este definită dacă şi numai dacă $ss_x = 0$, adică $\sum_{i=1}^{n}(x_i - \overline{x})^2 = 0$, sau echivalent $x_i - \overline{x} = 0$, oricare ar fi $i = \overline{1, n}$:

$$\begin{aligned} x_1 &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ x_2 &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &\vdots \\ x_n &= \frac{x_1 + x_2 + \cdots + x_n}{n} \end{aligned}$$

Dar aceste relaţii au loc dacă şi numai dacă $x_1 = x_2 = \cdots = x_n$.

Dacă norul de puncte nu este situat pe o dreaptă verticală $x = x_0$, se poate arăta că punctul $(\hat{\beta}_0, \hat{\beta}_1)$ este un punct de minim pentru Q, adică,

$$\frac{\partial^2 Q}{\partial \beta_0^2}(\hat{\beta}_0, \hat{\beta}_1) > 0 \tag{21.9}$$

şi

$$\begin{vmatrix} \dfrac{\partial^2 Q}{\partial \beta_0^2} & \dfrac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} \\ \dfrac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} & \dfrac{\partial^2 Q}{\partial \beta_1^2} \end{vmatrix} (\hat{\beta}_0, \hat{\beta}_1) > 0 \tag{21.10}$$

Dreapta de ecuaţie $y = \hat{\beta}_0 + \hat{\beta}_1 x$ se numeşte dreapta celor mai mici pătrate, asociată datelor (x_i, y_i), $i = \overline{1, n}$. Teoretic pentru orice intrare
$x \in [\min\{x_i\}, \max\{x_i\}]$, valoarea predicţionată de model pentru variabila răspuns, Y, este $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Faptul că dreapta celor mai mici pătrate există pentru orice set de date (x_i, y_i), $i = \overline{1, n}$, ce nu este dispus pe o dreaptă verticală, $x = x_0$, ne conduce la concluzia că simpla asociere a acesteia nu permite predicţii acceptabile pentru intrări x arbitrare. Cu alte cuvinte nu întotdeauna dreapta celor mai mici pătrate este modelul cel mai potrivit pentru date. Un exemplu este ilustrat în Fig.21.3.

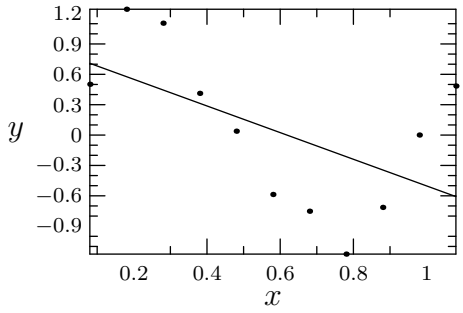

Fig.21.3: Dreapta celor mai mici pătrate asociată unui nor de puncte.

21.1.2 Regresie liniară simplă

Abordarea probabilistă a modelului predictiv pentru valorile variabilei Y ştiind că $X = x$ exploatează faptul că valorile variabilei conditionate $(Y|X = x)$ sunt împrăstiate în jurul mediei sale şi deci putem scrie ca variabila condiţionată, $(Y|X = x) = \underbrace{\varphi(x)}_{M(Y|X=x)} + \epsilon$, este suma dintre media sa şi o variabilă aleatoare ϵ, de medie 0 şi dispersie σ^2. ϵ măsoară abaterea faţa de medie. Modelul funcţional $y = G(x)$ al relaţiei dintre variabilele X şi Y este un model pentru funcţia de regresie $M(Y|X = x)$ a celor două variabile.

În condiţiile în care valorile variabilei X se pot observa (măsura) exact, şi pentru fiecare valoare înregistrată $(X = x)$, abaterea variabilei $(Y|X = x)$ fată de medie este dată de $\epsilon \sim N(0, \sigma^2)$, normal distribuită de medie 0 şi aceeaşi dispersie, atunci şi variabila aleatoare condiţionată $(Y|X = x) = M(Y|X = x) + \epsilon$ are tot distribuţia normală $N(\varphi(x) = G(x), \sigma^2)$, de medie $G(x)$ şi dispersie necunoscută, dar aceeaşi oricare ar fi x (Fig.21.4).

Alegând modelul liniar, $G(x) = \beta_0 + \beta_1 x$, adică admiţând ca regresia lui Y în raport cu X, $M(Y|X = x) = \beta_0 + \beta_1 x$, observaţiile (x_i, y_i), $i = \overline{1, n}$ sunt distribute în jurul dreptei d de ecuaţie $y = \beta_0 + \beta_1 x$.

Prin urmare avem următorul model probabilist pentru orice realizare y a variabilei aleatoare $(Y|X = x)$:

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{21.11}$$

unde ε este *eroarea aleatoare*, adică diferenţa dintre valoarea observată y şi media $M(Y|X = x)$. Se presupune în plus că variabilele condiţionate $(Y|X = x_i)$, $i = \overline{1, n}$ sunt independente, cu alte cuvinte erorile corespunzătoare fiecărui y_i sunt independente, $i = \overline{1, n}$ şi au aceeaşi dispersie σ^2.

Dreapta d, de ecuaţie $y = \beta_0 + \beta_1 x$, se numeşte *dreapta de regresie*. Estimarea parametrilor β_0 şi β_1 se face prin metoda celor mai mici pătrate. Având estimatorii $\hat{\beta}_0$ şi $\hat{\beta}_1$, avem practic şi un estimator pentru media variabilei condiţionate $(Y|X = x)$, $M(Y|X = x)$, şi anume, $\hat{m} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Dispersia comună σ^2 a variabilelor $(Y|X = x_i)$, $\forall\, i$, măsoară variaţia aleatoare a valorilor variabilei Y în jurul dreptei de regresie. Ea este dispersia erorii aleatoare $\varepsilon \sim N(0, \sigma^2)$. Dreapta de regresie(dreapta celor mai mici pătrate) $y = \hat{\beta}_0 + \hat{\beta}_1 x$ se foloseşte

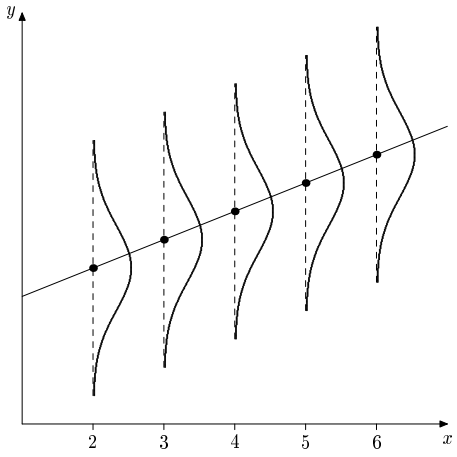

Fig.21.4: Distribuţia normală $N(\beta_0 + \beta_1 x, \sigma^2)$ a variabilelor $(Y|X = x)$.

pentru a prezice valoarea variabilei Y, ştiind că $X = x$. Şi anume valorea predicţionată va fi y-ul punctului de pe dreaptă corespunzător abscisei x, adică $\hat{y} = \beta_0 + \beta_1 x$ ($\hat{y}$ notează valoarea predicţionată). Cu cât dispersia erorii, σ^2, este mai mare, cu atât erorile de predicţie sunt mai mari.

Ţinând seama de importanţa ordinului de mărime al dispersiei, σ^2, este util să avem un estimator al acesteia. Pentru a determina estimatorul de maximă verosimilitate pe baza realizărilor $y_1, y_2, \ldots, y_n$ ale variabilelelor independente $(Y|X = x_i) \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, $i = \overline{1, n}$, considerăm funcţia de verosimilitate:

$$L(\beta_0, \beta_1, \sigma^2; y_1, y_2, \ldots, y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right] \tag{21.12}$$

Fie $\ell = \ln L$. Prin calcul direct obţinem punctul staţionar al lui ℓ (care este un maxim) ca fiind $(\beta_0, \beta_1, \sigma^2)$, cu $\beta_0 = \hat{\beta}_0, \beta_1 = \hat{\beta}_1$ (estimatorii celor mai mici pătrate) şi

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i]^2 \tag{21.13}$$

Observăm că $\dfrac{1}{n} \sum_{i=1}^{n} [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i]^2 = \dfrac{1}{n} Q(\hat{\beta}_0, \hat{\beta}_1)$, adică estimatorul de maximă verosimilitate pentru σ^2 este valoarea minimă a lui Q supra n, volumul eşantionului. Notăm $Q_0 = Q(\hat{\beta}_0, \hat{\beta}_1)$. Pentru implementarea calculului lui Q_0 este mai utilă exprimarea:

$$Q_0 = ss_y - \frac{ss_{xy}^2}{ss_x}, \tag{21.14}$$

unde

$$ss_y = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{2} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}, \tag{21.15}$$

iar ss_x, ss_{xy} au fost precizaţi mai sus.

Estimatorul Q_0/n al dispersiei σ^2 este deplasat, dar ajustându-l la:

$$s^2 = \frac{Q_0}{n-2} \tag{21.16}$$

obţinem un estimator nedeplasat pentru σ^2.

Observaţia 21.1.1 *În statistică este preferat estimatorul nedeplasat $s^2 = Q_0/(n-2)$, în timp ce în machine learning se lucrează cu estimatorul deplasat Q_0/n, numit şi media pătratelor erorii, şi notat MSE, deoarece:*

$$MSE = \frac{Q_0}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Ce informaţie extragem din estimatorul s^2 al dispersiei σ^2?

$\sigma = \sqrt{\sigma^2}$ măsoară împrăstierea valorilor y în jurul dreptei medie $M(Y|X = x) = \beta_0 + \beta_1 x$. Conform Teoremei lui Cebîşev, aproximativ 75% din valorile y ale lui $(Y|X = x)$ cad în intervalul $(\beta_0 + \beta_1 x - 2\sigma, \beta_0 + \beta_1 x + 2\sigma)$. Deoarece σ nu este cunoscut, $2s$ estimează jumătate din acest interval. Un cunoscător al naturii datelor poate aprecia la ce nivel de mărime un estimator al dispersiei evidenţiază o împrăstiere mare în jurul dreptei celor mai mici pătrate şi poate concluziona când dreapta nu este un model potrivit pentru date.

În cazul Exemplului 1, cu datele vizualizate în Fig.21.1, dreapta de regresie estimată este $y = 0.3245x + 15.4386$ şi estimatorul abaterii standard, $s = 16.93$. Abaterea standard măsoară în acest caz, abaterea timpului de citire de la medie. Ori 16 milisecunde este o abatere redusă şi deci dreapta celor mai mici pătrate este un model bun pentru date.

21.2 Regresia logistică

În problemele de clasificare din machine learning "obiectele" se clasifică în categorii. Considerăm cazul cel mai simplu, în care există doar două categorii. De exemplu pacienţii având un anumit nivel de glicemie se clasifică în bolnav de diabet, sau nu are diabet. Variabilă răspuns, Y, este în acest caz o variabilă aleatoare Bernoulli ce ia doar două valori. Pe baza observaţiilor asupra variabilei X, ce dă nivelul glicemiei pacienţii se clasifică în cele două categorii. Codificând (etichetând) cu 1 clasa "bolnav de diabet" şi cu 0 clasa, "diabetul nu este prezent" ar trebui ca din setul de observaţii (date de antrenament):

$$(x_1, b_1), (x_2, b_2), \ldots, (x_n, b_n),$$

formate din perechi (x_i =nivel glicemie, b_i =eticheta), $b_i \in \{0, 1\}$, să găsim un model funcţional $y = G(x)$ pentru regresia variabilei binare Y în raport cu X, adică pentru funcţia $M(Y|X = x)$.

Variabilele $(Y|X = x)$ au distribuţia Bernoulli, de parametru $p(x)$:

$$(Y|X=x) = \begin{pmatrix} 1 & 0 \\ p(x) & 1-p(x) \end{pmatrix}$$

şi valoarea lor medie este:

$$M(Y|X=x) = 1 \cdot p(x) + 0 \cdot (1-p(x)) = p(x) = P(Y=1|X=x)$$

Prin urmare pentru a putea face predicţii privind clasa unui pacient ce are nivelul glicemiei $(X=x)$, ar trebui să găsim un model funcţional pentru funcţia $M(Y|X=x) = p(x) \in (0,1)$, care este probabilitatea ca pacientul cu nivelul de glicemie x să aibă diabet (de ce?! pentru că am arătat în secţiunea relativ la regresia liniară că modelul funcţional, $y = G(x)$, într-o problemă de regresie, este un model pentru funcţia $M(Y|X=x)$).

Dacă un astfel de model funcţional, $p(x) = G(x)$ este ales şi estimat din datele de antrenament, atunci pentru un pacient cu nivelul glicemiei $x = 138$, de exemplu, se calculează $p(x)$, adică $p(138)$ şi dacă $p(x) \geq 1/2$ atunci pacientul este clasificat ca având diabet, adică valoarea predicţionată pentru variabila $(Y|X=x)$ este 1, iar dacă $p(x) < 1/2$, atunci valoarea predictionată este 0 (pacientul nu are diabet). Aceasta este ideea de bază în clasificarea bazată pe regresia logistică.

Pentru a înţelege modalitatea de alegere a modelului funcţional $y = G(x)$ în cazul regresiei logistice (care în cazul liniar era $y = \beta_0 + \beta_1 x$) fixăm câteva notaţii şi mărimi:

• În pariuri se foloseşte foarte mult, nu probabilitatea, p, ca o echipă să câştige, ci raportul $\dfrac{p}{1-p}$ dintre probabilitatea de succes şi probabilitatea de a pierde jocul. Acest raport se numeşte şansa de câştig. Această şansă se estimează ca raportul dintre numărul cazurilor favorabile şi numărul cazurilor nefavorabile.

Spre deosebire de p, probabilitatea de succes, care ia valori în $[0,1)$, şansa de câştig,

$$\frac{p}{1-p}$$

ia orice valoare pozitivă. Graficul funcţiei $h(p) = \dfrac{p}{1-p}$, $p \in [0,1)$ este ilustrat în Fig.21.5, stânga.

Luând valori pozitive are sens să considerâm logaritmul natural al raportului $p/(1-p)$,

$$\ln \frac{p}{1-p}$$

$\ln \dfrac{p}{1-p}$ se numeşte *logit*.

Graficul funcţiei logit pe intervalul $(0,1)$ este graficul din dreapta din Fig.21.5.

• Distribuţia de probabilitate a unei variabile aleatoare Bernoulli, B, ce ia valoarea 1 cu probabilitatea p şi valoarea 0 cu probabilitatea $1-p$ se poate exprima concentrat astfel

$$P(B=b) = p^b(1-p)^{1-b} = \begin{cases} p & \text{dacă } b=1 \\ 1-p & \text{dacî } b=0 \end{cases}$$

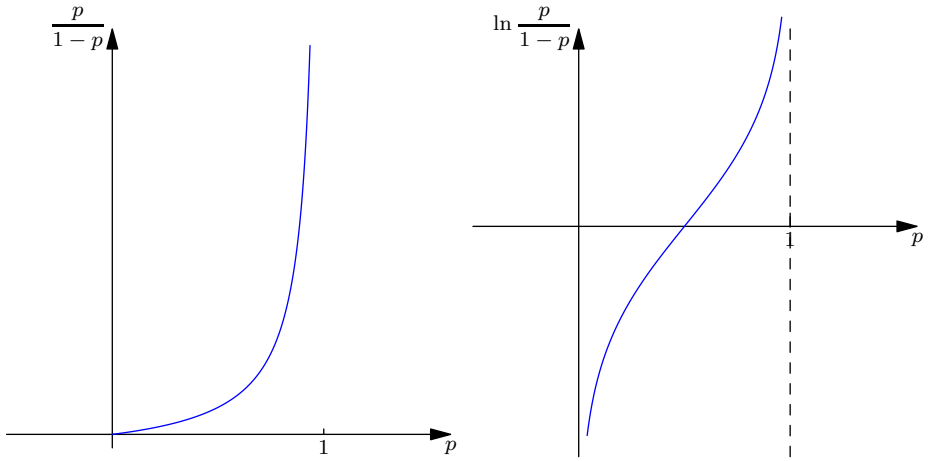

Fig.21.5: Graficul funcţiei $h(p) = p/(1-p)$ (stânga) şi al funcţiei logit (dreapta).

• În exemplul dat mai sus am considerat cazul simplist când clasificarea se face pe baza unei singure caracteristici, nivelul glicemiei. Regresia logistică însă poate lua în considerare mai multe caracteristici ale obiectelor ce se clasifică. În cazul diabetului, pe lângă glicemie se iau în considerare şi alţi parametri.

• În forma sa cea mai generală regresia logistică cu variabila răspuns binară (Bernoulli) consideră construirea unui model funcţional $Y = G(X_1, X_2, \ldots, X_m)$, unde variabilele aleatoare $X_1, X_2, \ldots, X_m$ măsoară nivelul caracteristicilor (trăsăturilor) $1, 2, \ldots, m$, luate în considerare pentru a face clasificarea în două categorii. Notăm cu $X = (X_1, X_2, \ldots, X_m)$, vectorul aleator asociat trăsăturilor de interes.

În acest caz, datele iniţiale, de antrenament, pe baza cărora se construieşte modelul logistic sunt un set de n perechi (caracteristici, clasa):

$$\begin{array}{ll} (x_1, b_1), & x_1 = (x_{11}, x_{12}, \ldots, x_{1m})^T \\ (x_2, b_2), & x_2 = (x_{21}, x_{22}, \ldots, x_{2m})^T \\ \vdots & \\ (x_n, b_m), & x_n = (x_{n1}, x_{n2}, \ldots, x_{nm})^T \end{array}$$

unde pentru fiecare obiect i a cărui clasă (categorie), b_i este cunoscută, avem înregistrate în vectorul x_i, valorile numerice ale trăsăturilor sale, $(x_{i1}, x_{i2}, \ldots, x_{im})^T$.

De exemplu un pacient cu două caracteristici, adică $m = 2$, (glicemie=200, tensiune sistolică=180), este clasificat prin $b = 1$, când are diabet.

Notăm cu $p(x_i)$ probabilitatea ca cel de-al i-lea obiect să fie etichetat cu 1, ştiind că el are trăsăturile din vectorul x_i, adică concentrat

$$p(x_i) = P(Y = 1 | X = x_i)$$

În aceste condiţii şi notaţii se caută un model pentru funcţia

$$p(x) = P(Y = 1 | X = x) = M(Y | X = x)$$

adică pentru probabilitatea ca un nou obiect căruia i se observă trăsăturile, $1, 2, ..m$ şi se înregistrează valorile ce sunt coordonate ale vectorului,$x = (c_1, c_2, \ldots, c_m) \in \mathbb{R}^m$ să fie clasificat ca fâcând parte din clasa cu eticheta 1.

În regresia logistică se procedează astfel:

1) se consideră suma ponderată a caracteristicilor, adică suma $\beta_0 + \beta_1 c_1 + \cdots + \beta_m c_m = \beta_0 + \beta^T x$, unde β_i sunt parametri ce se vor estima din date. $\beta = (\beta_1, \beta_2, \ldots, \beta_m)^T$, iar $\beta^T x$ este produsul scalar $< \beta, x >$.

2) Dearece funcţia logit, $\ln \dfrac{p(x)}{1 - p(x)}$, poate lua orice valori reale (vezi graficul), se poate impune ca

$$\ln \frac{p(x)}{1 - p(x)} = \beta_0 + \beta^T x$$

şi din această relaţie rezultă:

$$p(x) = \frac{e^{\beta_0 + \beta^x}}{1 + e^{\beta_0 + \beta^x}} = \frac{e^{\beta_0 + \beta_1 c_1 + \cdots \beta_m c_m}}{1 + e^{\beta_0 + \beta_1 c_1 + \cdots \beta_m c_m}}$$

Expresia pentru funcţia $p(x)$ se numeşte *funcţia logistică*. Evident că ea fiind raportul dintre un număr pozitiv şi numărul plus 1, ia valori în intervalul $(0, 1)$. Graficul ei este ilustrat în Fig.21.6. Funcţia $p(x) = \dfrac{e^x}{1 + e^x}$ are proprietăţile unei funcţii de repartiţie.

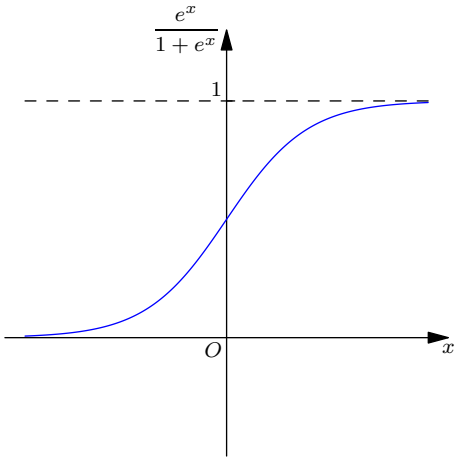

Fig.21.6: Graficul funcţiei logistice.

După ce parametrii β_i, $i = 0, 1, \ldots, m$ se estimează din datele de antrenament atunci, pentru un nou obiect ce are caracteristicile înregistrate într-un vector x se calculează $p(x)$. Dacă $p(x) \geq 1/2$ (probabilitatea de a face parte din clasa 1 este mai mare decât $1/2$, atunci obiectul este clasificat ca fiind din clasa 1, daca $p(x) < 1/2$, atunci el este clasificat în categoria cu eticheta 0.

Estimarea parametrilor β_i din date.

Ca şi în cazul regresiei liniare se consideră că variabilele aleatoare Bernoulli

$$(Y|X = x_1), (Y|X = x_2), \ldots, (Y|X = x_n)$$

sunt independente, adică clasa căruia va fi atribuit un obiect este independentă de clasa altui obiect.

Fiecare variabilă $(Y|X = x_i)$ are distribuţia Bernoulli de probabilitate de succes $p(x_i)$. Estimăm parametrii $\beta_0, \ldots, \beta_m$ ai regresiei logistice prin metoda verosimilităţii maxime. Funcţia de verosimilitate este:

$$L(\beta_0, \beta_1, \ldots, \beta_m; \underbrace{x_1, x_2, \ldots, x_n; b_1, b_2, \ldots, b_n}_{\text{cunoscute;datele de antrenament}}) =$$

$$= P(Y = b_1|X = x_1)P(Y = b_2|X = x_2) \cdots P(Y = b_n|X = x_n) =$$

$$\underbrace{p(x_1)^{b_1}(1 - p(x_1))^{1-b_1}}_{P(Y=b_1|X=x_1)} \underbrace{p(x_2)^{b_2}(1 - p(x_2))^{1-b_2}}_{P(Y=b_2|X=x_2)} \cdots \underbrace{p(x_n)^{b_n}(1 - p(x_n))^{1-b_n}}_{P(Y=b_n|X=x_n)}$$

Prin urmare vom determina parametrii β_i care maximizează probabilitatea înregistrării datelor de antrenament (citiţi cursul relativ la estimatorul verosimilităţii maxime), adică determinăm un punct de maxim $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m)$, al funcţiei $L(\beta_0, \beta_1, \ldots, \beta_m)$.

Ca deobicei e mai simplu să găsim punctul de maxim pentru $\ell = \ln(L)$. Logaritmând avem:

$$\ell(\beta_0, \beta_1, \ldots, \beta_m) = \ln(p(x_1)^{b_1}(1 - p(x_1))^{1-b_1}) + \cdots + \ln(p(x_n)^{b_n}(1 - p(x_n))^{1-b_n}) =$$
$$b_1 \ln(p(x_1)) + (1 - b_1) \ln(1 - p(x_1)) + \cdots + b_n \ln(p(x)n)) + (1 - b_n) \ln(1 - p(x_n))$$

Înlocuind fiecare $p(x_i) = \dfrac{e^{\beta_0+\beta^T x_i}}{1 + e^{\beta_0+\beta^T x_i}}$, $1-p(x_i) = \dfrac{1}{1 + e^{\beta_0+\beta^T x_i}}$, şi calculând derivatele parţiale ale lui ℓ în raport cu $\beta_0, \beta_1, \ldots, \beta_m$ se observă că sistemul din care ar trebui să deducem punctele critice ale funcţiei ℓ:

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_0} &= 0 \\ \frac{\partial \ell}{\partial \beta_1} &= 0 \\ &\vdots \\ \frac{\partial \ell}{\partial \beta_m} &= 0 \end{aligned}$$

nu poate fi rezolvat analitic ca în cazul regresiei liniare, ci doar numeric.

Pachetele software de calcul numeric şi statistic (`MATLAB`, `Python, R`) oferă funcţii de estimare a parametrilor regresiei logistice prin metoda gradientului descendent (în loc să se caute maximul funcţiei $\ell = \ln(L)$, se caută minimul funcţiei $-\ell$).

După ce modelul a fost construit, adică s-au determinat parametrii $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$, spaţiul $\mathbb{R}^m$ al vectorilor trăsăturilor este separat de hiperplanul de ecuaţie $\hat{\beta}_0 + \hat{\beta}_1 c_1 +$

$\cdots, \hat{\beta}_m c_m = 0$ (dacă $m = 1$ atunci avem $\hat{\beta}_0 + \hat{\beta}_1 c_1 = 0$, care reprezintă un punct c_1 în $\mathbb{R}$, pt $m = 2$, $\hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 = 0$, este o dreaptă, pt $m = 3$, $\hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \hat{\beta}_3 c_3 = 0$, este un plan şi de la $m > 3$ avem hiperplan).

Deci dacă caracteristicile x ale unui obiect constituie un punct din hiperplan, atunci $p(x) = \dfrac{e^0}{1 + e^0} = 1/2$. Acest hiperplan este hiperplanul de frontieră între regiunea ce conţine punctele x cu $p(x) > 1/2$ şi regiunea în care $p(x) < 1/2$. El se numeşte hiperplanul de discriminare.

Regresia logistică este folosită de motoarele de căutare, pentru a decide afişarea sau nu a unor ads-uri adecvate pentru cuvintele de căutare ale utilizatorilor. Se iau în calcul mai multe caracteristici ale fiecărui ad şi se cuantifică numeric. Pe baza unor date relativ la metrica numită *click-through rate* (`http://en.wikipedia.org/wiki/Click-through_rate`) înregistrată pentru ads-urile mai vechi se decide afişarea sau nu a unui nou ad pentru un utilizator.

Regresia logistică se aplica foarte mult în medicină pentru a construi predictori pentru diverse afecţiuni. Modelul logistic, aplicat via reţele neuronale, a dat rezultate în predictia unor afecţiuni cardiace sau de altă natură.

Apoi modele bazate pe regresia logistică şi tehnici de inteligenţă artificială pot prezice riscuri financiare, în asigurări, etc.