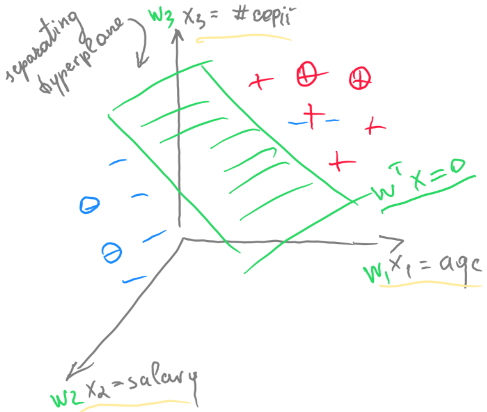
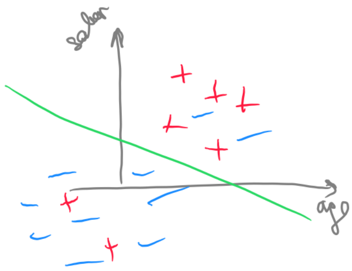
Lectia 6: Logistic Regression → Linear Classification



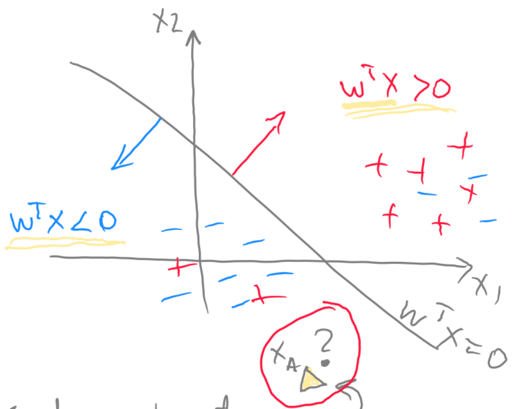
$\theta^T x$

$$w^T x = w_0 x_0 + w_1 x_1 + w_2 x_2 + w_3 x_3, \quad x_0 = 1$$

$$(w_1\ w_2 \ldots) \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix}$$



Intuiția în cadrul logistic regression este că $w^T x = 0$ reprezintă un plan (pentru două variabile - o dreaptă) care împarte spațiul în două zone: zona pozitivă și zona negativă. Exemplu pt 2 variabile:



Având un punct nou cărei clase îl atribuim?

Apropo, geometric $w^T x_A$ are și o semnificație: distanța de la punctul x_A până la planul $w^T x = 0$ este proporțională cu $w^T x_A$. În timp ce distanța nu neapărat ne interesează în această lecție, semnul produsului scalar $w^T x$ ne indică: pozitiv → clasa produsului scalar $w^T x$ ne indică: pozitiv $\to$ clasa $+1$; negativ $\to$ clasa 0.

O proprietate faina a logistic regression este ca inloc sa ne returneze o anumita clasa (0 sau 1), aceasta ne returneaza probabilitatea.

$$p_+ = P(y_i = 1 \mid x_i, w)$$

Avand capacitatea de a returna nu doar un raspuns ("0" sau "1") ci probabilitatea asignarii unei observatii la clasa "1" este o cerinta foarte importanta in multe probleme de business. Spre exemplu, scoringul creditelor este adesea o aplicarea a regresiei ~~liniare~~ logistice - clientii bancii care au aplicat pentru un imprumut sunt evaluati in baza probabilitatilor prezise pentru a intelege daca acestia sunt potentiali platitori "nesatisfacatori" pana la "buni". Iata un exemplu:

client	Probabilitatea prezisă (că nu va returna banii)
Mihai	0.78
Andrei	0.45
Oana	0.13
Natalia	0.06
Cosmin	0.03

Refuz

$p^* = 0.15$

Accept

Intuitia in spatele logistic regression:

$w^T x \in \mathbb{R}$, pentru conveniență $x = (1\ x_1\ x_2 \dots x_d)$

age, salary ... dept

ținem că $w = (w_0\ w_1\ w_2 \dots w_d)$

w_{age} w_{salary} w_{dept}

$w^T x \to [0,1]$ -?

scopul nostru este de a prezice:

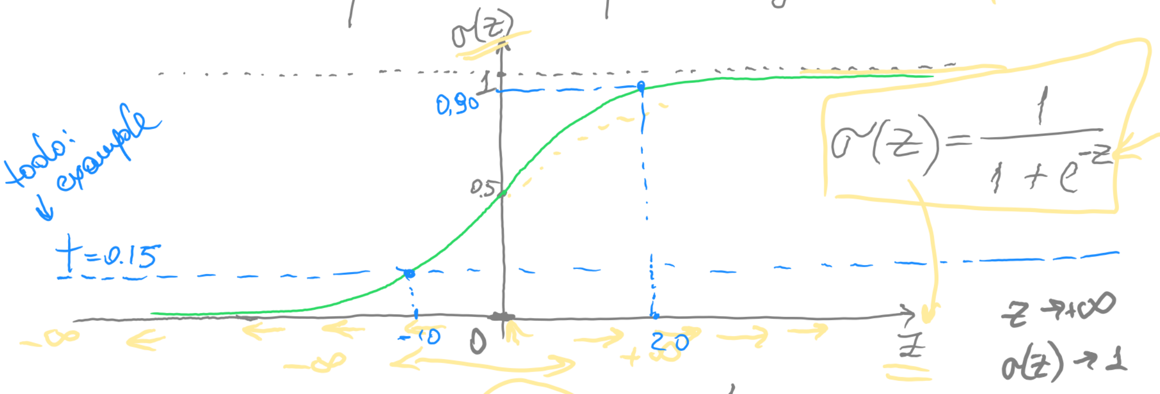
$$p_+ = P\{y = 1 \mid \vec{x}\} \in [0,1]$$

Q: Cum putem face astfel de preziceri, dat fiind faptul că $w^T x \in \mathbb{R}$ -?

0 1 18 1500 ...

$[0,1]$

A: Folosim pentru asta funcția sigmoid: $[0,1]$



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

$z \to +\infty$, $\sigma(z) \to 1$

$z \to -\infty$, $\sigma(z) \to 0$

$$\Rightarrow P_+ = \sigma(w^T x) = \frac{1}{1 + e^{-w^T x}}$$

w^T x_i

$[0,1]$

Cum verificam prezicerile noastre?

$y = 0, 1$ 0.5 0,3 0.7.

Pentru asta am putea folosi rata de clasificare eronată – missclassification rate:

$\mathcal{D} \to (x, y)$ — datele, label

$$mcr(w, \mathcal{D}) = mcr(\hat{y}(\cdot; w), \mathcal{D})$$

$$= \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} I(y \neq \hat{y}(x; w))$$

adevărat, prezis

$$= \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} I(y \neq I(\underbrace{logistic(w^T x)}_{\sigma(w^T x)} \geq 0.5))$$

0.15, 0.70

$0 \to 1$
$1 \to 0$

Această funcție, însă, nu este potrivită pentru optimizare întrucât nu este continuă.

Funcția proxi care se folosește, în schimb, este dată de formula: log likelihood

$y \in \{0, 1\}$

$$\log L = \sum_{n=1}^{N} y_n \log \hat{y}_n + (1 - y_n) \log (1 - \hat{y}_n)$$

$y_n = 1$; $(1-1) = 0$

$$\frac{e^{w^T x}}{1 + e^{w^T x}} = \frac{1}{1 + e^{-w^T x}}$$

$$= \sum_{n=1}^{N} y_n \log \frac{e^{w^T x_n}}{1 + e^{w^T x_n}} + (1 - y_n) \log \left(1 - \frac{1}{1 + e^{w^T x_n}}\right)$$

$\vdots$

$$= \sum_{n=1}^{N} y_n (w^T x_n - \log(1 + e^{w^T x_n})) + ((1 - y_n)(-\log(1 - e^{w^T x_n})))$$

$$\boxed{= \sum_{n=1}^{N} y_n \cdot w^T x_n - \log(1 + e^{w^T x_n})}$$

w – ? $= \theta$?

$\frac{\partial}{\partial w}$ $y_n \cdot x_n$

$\frac{\partial}{\partial w} \log a = \frac{1}{a} \cdot a'$

Ne reamintim logaritmii:

$$2^4 = 16$$

$$4 = \log_2 16$$

$2^x = 16$

La fel:

$$e^x = y$$

$$x = \log_e y = (\ln y)$$

$e^x = y$

Pana acum am inteles cum facem preziceri cu ajutorul functiei logistice (de unde vine si numele logistic regression). Dar, ca si la regresia liniara, am folosit vectorul de parametri w (lectia trecuta notat cu theta). Cum gasim w?

Ca și în lecția anterioară, pentru a minimiza (în cazul nostru maximiza) o funcție avem nevoie de derivata acestei funcții:

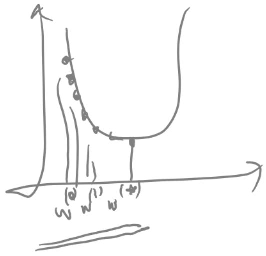
$$\frac{\partial L}{\partial w} = \sum_{n=1}^{N} y_n x_n - \frac{1}{1+e^{w^T x_n}} \cdot e^{w^T x_n} \cdot x_n$$

$$= \sum_{n=1}^{N} x_n (y_n - \hat{y}_n)$$

$$= X^T (y - \hat{y})$$

pentru a minimiza este destul să punem ("–") în față

Având derivata putem scrie formula de update pentru parametrul w



pentru parametrul w:

$$w^{(t)} = w^{(t-1)} - \alpha \cdot \frac{\partial L}{\partial w} \Rightarrow$$

0.1, 0.001 ...

$$w^{(t)} = w^{(t-1)} + \alpha \cdot X^T(y - \hat{y})$$

Mai multe detalii despre pasii de derivare poti gasi in K. Murphy - Machine Learning - A probabilistic approach.

Sa ne uitam la implementarea algoritmului intr-un Jupyter Notebook.